INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
HEALTH STATUS PREFERENCE MEASUREMENT:
A COMPARISON OF DISCRETE CHOICE CONJOINT
AND CONDITIONAL UTILITY MODELING

Volume I

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of
Philosophy in the Graduate School of The Ohio State University

By

Zafar Mohomed Anwar Hakim, B.Pharm., M.S.

* * * * *

The Ohio State University

1995

Dissertation Committee: Approved by

Dev S. Pathak
Suzan Kucukarslan
Jon C. Schommer

Adviser
College of Pharmacy
Copyright by
Zafar Mohomed Anwar Hakim
1995
To

My Parents
ACKNOWLEDGMENTS

It is a pleasure to acknowledge the contribution Michael Browne, Robert MacCallum, and H. Rao Unnava have made to my academic training at The Ohio State University. I wish to express my sincere gratitude to Suzan Kucukarslan and Jon Schommer, for their patience and advice throughout the dissertation process. I would also like to acknowledge the role my academic adviser Dev Pathak has played in my education, the significance of which extends beyond the purview of my academic training.

In my time as a graduate student, I have enjoyed working with and learning a lot from Bill Arneson, Ann Segraves, Mary Sexton, and Elliott Sogol. Thanks are also due to Iris Brooks, Pamela Dennewitz, Thomas Grubaugh, Gregg Kuck, Chuck and Janet McClary, and Larry Wood for making the experiential part of my education truly memorable and immensely enjoyable.

I am indebted to Fred Ruland, John Downing, and Raymond Steinbacher at the Statistical Consulting Service and Neeraj Arora of the marketing department on campus, Bill Furlong of McMaster University, Jim Mundt of the Vermont Alcohol Research Center, and Richard Johnson of Sawtooth Software for their invaluable help in the design and analysis of my dissertation. It is a pleasure to acknowledge the support of the VA Outpatient Clinic in the data collection phase of my dissertation. I am also thankful to the 140 respondents who provided the data for my dissertation.

I do get by with a little help from my friends and my family. I would like to offer a word of thanks to my colleagues, for tolerating my effective monopolization of “my” computer over the past three months and my “in the zone” behavior at various times in the dissertation process. Thanks are also due to Brian, Conor, Dave, Nabil, Ron, and Saeed for helping me keep things in perspective. I am indebted to Anish and Manju, for their support and understanding over the course of my graduate work. A sincere word of thanks is due to the Kazis and the Pirzadas, for making sure I was never far away from home. Above all, there are my grandparents and my parents, to whom I am especially grateful for their faith ... and their footprints.
VITA

February 8, 1969 .................. Born - Bombay, India

1990 ............................... B.Pharm., University of Bombay, Bombay, India

1990-1991 .......................... University Fellow, The Ohio State University, Columbus, Ohio

1991 ............................... M.S., The Ohio State University, Columbus, Ohio

1991-1993 .......................... Graduate Teaching Associate, College of Pharmacy, The Ohio State University, Columbus, Ohio

1993-1994 .......................... University Fellow, The Ohio State University Columbus, Ohio

1994-1995 .......................... Graduate Teaching Associate, College of Pharmacy, The Ohio State University, Columbus, Ohio
PUBLICATIONS


FIELDS OF STUDY

Major Field: Pharmacy

Studies in: Marketing (Dr. H. Rao Unnava)
             Pharmaceutical Administration (Dr. Dev S. Pathak)
             Pharmacoeconomics (Dr. Dev S. Pathak)
# TABLE OF CONTENTS

**DEDICATION** .......................................................... ii

**ACKNOWLEDGMENTS** ............................................. iii

**VITA** ........................................................................ iv

**LIST OF TABLES** ..................................................... xvi

**LIST OF FIGURES** ................................................... xxii

**CHAPTER**

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. PREAMBLE</td>
<td>1</td>
</tr>
<tr>
<td>Chapter Overview</td>
<td>1</td>
</tr>
<tr>
<td>The Outcomes Movement in Health Care</td>
<td>2</td>
</tr>
<tr>
<td>A Clarification of Terminology</td>
<td>6</td>
</tr>
<tr>
<td>Health</td>
<td>7</td>
</tr>
<tr>
<td>Health Status, Functional Status, and Physiological Status</td>
<td>14</td>
</tr>
<tr>
<td>Quality of Life and Health-Related Quality of Life</td>
<td>17</td>
</tr>
<tr>
<td>Health Status Measurement -- Why the Interest?</td>
<td>21</td>
</tr>
<tr>
<td>Medical Trends</td>
<td>21</td>
</tr>
<tr>
<td>Economic/Regulatory Trends</td>
<td>25</td>
</tr>
<tr>
<td>Social Trends</td>
<td>28</td>
</tr>
<tr>
<td>Health Status Measurement -- A Historical Perspective</td>
<td>40</td>
</tr>
<tr>
<td>The Early Period</td>
<td>41</td>
</tr>
<tr>
<td>The Modern Period</td>
<td>48</td>
</tr>
</tbody>
</table>
III. HEALTH STATUS PREFERENCE MEASUREMENT METHODOLOGIES ................................................................. 239

Chapter Overview ................................................................. 239
Methodologies for Health Status Preference Measurement ................................................. 240
A Framework of Subjective Measurement ........................................................................... 244
Measurement Strategies for Health Status Preference Measurement ................................. 248
Holistic Preference Measurement Strategies ................................................................. 250
Decomposed Preference Measurement Strategies ......................................................... 254
Explicitly Decomposed Preference Measurement Strategies .................................................. 257
The Additive Rating Scale Method ..................................................................................... 257
The Conditional Utility Function-based Procedure ........................................................... 261
Checking Independence Assumptions ................................................................................ 262
Assessing Preference Functions over Single Attributes or Dimensions .................................. 265
Determining Scaling Constants ............................................................................................ 266
Statistically Inferred Decomposed Preference Measurement Strategies .................................. 268
Rank Order Judgment Strategies ....................................................................................... 274
Rating Scale Judgment Strategies ....................................................................................... 277
Functional Measurement .................................................................................................. 278
Social/Clinical Judgment Analysis ....................................................................................... 288
Choice Simulation Methods ............................................................................................... 298
Deterministic Simulation Methods ....................................................................................... 299
Probabilistic Simulation Methods ....................................................................................... 302
Discrete Choice Conjoint Strategies ................................................................................... 307
Introduction ......................................................................................................................... 307
Theory ................................................................................................................................. 309
Design ................................................................................................................................. 319
Analysis ............................................................................................................................... 322
Evaluation of Discrete Choice Conjoint Strategies ............................................................. 328
Hybrid or Mixed Preference Measurement Strategies .......................................................... 339
Hybrid Conjoint Modeling and Huber-hybrid Modeling .......................................................... 339
The Analytic Hierarchy Process ............................................................................................. 343
Scaling Approaches for Health Status Preference

Measurement .................................................. 349
Scaling People Versus Scaling Stimuli ................. 350
Direct Versus Indirect Scaling Approaches .......... 351
Psychophysics .................................................. 353
Decision Theory .................................................. 358
Direct Scaling Approaches ...................................... 363
Rating Scale Methods ....................................... 364
The Magnitude Estimation Method ....................... 372
Gamble Methods .............................................. 377
The Time Tradeoff Method .................................. 389
The Contingent Valuation Method ......................... 391
Indirect Scaling Approaches ...................................... 396
The Paired Comparisons Method ......................... 396
The Pick-one Method ......................................... 399
Evaluation of Existing Research in Health Status
Preference Assessment .................................. 403
The San Diego Group ........................................ 404
The McMaster Group ......................................... 407
The EuroQol Group ........................................... 409

IV. COMPARISONS OF PREFERENCE MEASUREMENT
METHODOLOGIES IN THE PUBLISHED LITERATURE ... 413

Chapter Overview ............................................. 413
Issues in Comparing Preference Measurement
Methodologies .............................................. 414
Issues Relevant to the Comparisons of Interest
in this Study ................................................... 418
Top-down Versus Bottom-up Preference
Measurement Strategies ...................................... 421
Summary ......................................................... 461
Value-based Versus Utility-based Preference Scores .. 467
Summary ......................................................... 531
Judgmental Versus Choice Response Modes .......... 534
Summary ......................................................... 587
V. METHODOLOGY ................................................................. 605

Chapter Overview ......................................................... 605
Conceptualization and Operationalization of
Health Status ................................................................. 606
Rationale Behind the Selection of the EuroQol 606
Conceptual and Methodological Issues in
Health Status Measurement ......................................... 611
Health Status Preference Measurement Strategies
and Scaling Approaches ................................................. 615
Implementation of the Conditional Utility
Function-based Procedures ............................................ 616
Visual Aids used with Rating Scale and Standard
Gamble Scaling Methods ................................................. 625
Implementation of the Rating Scale Scaling Method .......... 631
Implementation of the Standard Gamble Scaling
Method ............................................................................. 636
Implementation of Discrete Choice Conjoint
Methodology ..................................................................... 639
Respondent Evaluations .................................................. 648
University Human Subjects Committee Approval
Procedures ........................................................................ 658
Clinic Research and Development Committee
Approval Procedures ....................................................... 659
Instrumentation Development Procedures .................... 661
Sample Characteristics and Sampling Procedures ........... 669
Data Collection Procedures ............................................. 680
Data Analysis Procedures ............................................... 688
    Data Entry and Coding Procedures ......................... 689
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Questions Pertaining to Modeling the Data</td>
<td>690</td>
</tr>
<tr>
<td>Modeling the Data for the Conditional Utility Function-based Procedure</td>
<td>690</td>
</tr>
<tr>
<td>Modeling the Data for the Discrete Choice Conjoint Methodology</td>
<td>693</td>
</tr>
<tr>
<td>Research Questions Pertaining to Comparing Preference Scores</td>
<td>702</td>
</tr>
<tr>
<td>Preferential Ordering</td>
<td>703</td>
</tr>
<tr>
<td>Magnitude of Preference Scores</td>
<td>712</td>
</tr>
<tr>
<td>Correlations</td>
<td>716</td>
</tr>
<tr>
<td>Concordances</td>
<td>722</td>
</tr>
<tr>
<td>Functional Relationships</td>
<td>729</td>
</tr>
<tr>
<td>Research Questions Pertaining to Comparing Respondent Evaluations</td>
<td>737</td>
</tr>
<tr>
<td>Reliability</td>
<td>737</td>
</tr>
<tr>
<td>Exploratory Factor Analyses</td>
<td>739</td>
</tr>
<tr>
<td>Comparing Respondent Evaluations</td>
<td>756</td>
</tr>
<tr>
<td>VI. RESULTS</td>
<td>761</td>
</tr>
<tr>
<td>Chapter Overview</td>
<td>761</td>
</tr>
<tr>
<td>Response Analysis</td>
<td>762</td>
</tr>
<tr>
<td>Descriptive Characteristics of Sample</td>
<td>765</td>
</tr>
<tr>
<td>Results of the Modeling Procedures</td>
<td>774</td>
</tr>
<tr>
<td>Modeling the CUF-RS Data</td>
<td>774</td>
</tr>
<tr>
<td>Modeling the CUF-SG Data</td>
<td>783</td>
</tr>
<tr>
<td>Modeling the DCCM Data</td>
<td>796</td>
</tr>
<tr>
<td>Results of the Comparison of Preference Scores</td>
<td>819</td>
</tr>
<tr>
<td>Preferential Ordering</td>
<td>820</td>
</tr>
<tr>
<td>Magnitude of Preference Scores</td>
<td>823</td>
</tr>
<tr>
<td>Correlations</td>
<td>828</td>
</tr>
<tr>
<td>Concordances</td>
<td>835</td>
</tr>
<tr>
<td>Functional Relationships</td>
<td>841</td>
</tr>
<tr>
<td>Results of the Comparison of Respondent Evaluations</td>
<td>849</td>
</tr>
<tr>
<td>Reliability</td>
<td>849</td>
</tr>
<tr>
<td>Exploratory Factor Analyses</td>
<td>850</td>
</tr>
<tr>
<td>Comparing Respondent Evaluations</td>
<td>874</td>
</tr>
</tbody>
</table>
Implications of the Results ............................................... 1007
Implications for Health Status Preference
   Measurement .................................................. 1007
Implications for Health Status Measurement 1012
Implications for the Outcomes Movement in
   Health Care .................................................... 1014
Limitations ................................................................. 1018
Avenues for Future Research ............................................. 1020

APPENDICES

A. Correspondence with EuroQol Group and EuroQol
   Instrument Used in the Present Study ......................... 1025
B. Cards Used to Describe Health States ......................... 1030
C. Interviewer Schedule .............................................. 1041
D. Data Collection Form for Rating Scale Scaling Method ..... 1109
E. Data Collection Form for Standard Gamble Scaling Method . 1113
F. Sample Interview for Discrete Choice Conjoint Methodology 1117
G. Respondent Evaluation Forms For Each Individual
   Preference Measurement Methodology ...................... 1134
H. Respondent Evaluation Forms for Overall Comparative
   Evaluation and Other Feedback ............................... 1147
I. Application for Exemption from Full Review Submitted to
   the Behavioral and Social Sciences Review Committee .... 1152
J. Summary Statements and Proposal for Full Review Submitted
   to the Behavioral and Social Sciences Review Committee .. 1154
K. Response From the Behavioral and Social Sciences Review
   Committee ....................................................... 1170

xiv
L. Revised Summary Sheets Submitted to the Behavioral and Social Sciences Review Committee ................. 1173

M. Final Approval Form of the Behavioral and Social Sciences Review Committee ......................... 1177

N. Proposal Submitted to the VAOPC Research and Development Committee ......................... 1179

O. Minutes and Official Response From the VAOPC Research and Development Committee ............ 1190

P. Research Consent Forms ............................................................... 1194

Q. VA Research and Development Project Data Sheet ....................................................... 1199

R. Background Information Questionnaire ......................................................... 1202

S. Summary Information Sheet ................................................................. 1207

T. Materials for Recruiting Patients Over the Telephone ...... 1209

U. Codebook ............................................................................................. 1213

BIBLIOGRAPHY ................................................................. 1249

xv
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Distinctions Between the Preference Measurement</td>
<td>420</td>
</tr>
<tr>
<td>Methodologies Investigated in this Study</td>
<td></td>
</tr>
<tr>
<td>2. Theoretical Relationships Between Value and Utility</td>
<td>479</td>
</tr>
<tr>
<td>Functions</td>
<td></td>
</tr>
<tr>
<td>3. Health Status Classification System of the EuroQol</td>
<td>613</td>
</tr>
<tr>
<td>4. Time Estimates for Elements of the Pre-test</td>
<td>664</td>
</tr>
<tr>
<td>5. Time Estimates for Elements of the Pilot-test</td>
<td>667</td>
</tr>
<tr>
<td>6. Calculations for Sample Size Determination</td>
<td>677</td>
</tr>
<tr>
<td>7. Response Analysis</td>
<td>764</td>
</tr>
<tr>
<td>8. Order of Presentation of Preference Measurement</td>
<td>766</td>
</tr>
<tr>
<td>Methodologies</td>
<td></td>
</tr>
<tr>
<td>9. Demographic Characteristics of Respondents</td>
<td>768</td>
</tr>
<tr>
<td>10. Severity of Hypertension of Respondents</td>
<td>769</td>
</tr>
<tr>
<td>11. Respondent Comorbidities</td>
<td>772</td>
</tr>
<tr>
<td>12. Health Status of Respondents as Operationalized by the</td>
<td>773</td>
</tr>
<tr>
<td>EuroQol</td>
<td></td>
</tr>
<tr>
<td>13. Single Attribute Values Determined using CUF-RS</td>
<td>776</td>
</tr>
<tr>
<td>14. Single Attribute Disvalues Determined using CUF-RS</td>
<td>777</td>
</tr>
</tbody>
</table>
31. Illustration of Conversion of DCCM Scores of Individual Levels of Health Status Attributes to 0.0-1.0 Scale .......................... 811

32. Illustration of Conversion of DCCM Scores of Multiattribute Health States to 0.0-1.0 Scale ................................................. 813

33. Preference Scores for all Individual Attribute Levels on the 0.0 to 1.0 Scale ........................................................................ 815

34. Preference Scores for Multiattribute Health States on the 0.0 to 1.0 Scale ....................................................................... 816

35. Overall Agreement in Rank Ordering of the Three Methodologies .......................................................................................... 821

36. Results of Kendall’s τ Between CUF-RS and CUF-SG .............. 824

37. Results of Kendall’s τ Between CUF-RS and DCCM ............... 825

38. Results of Kendall’s τ Between CUF-SG and DCCM ............... 826

39. Results of Repeated Measures One-way ANOVAs on Preference Scores Obtained Using the Three Methodologies ............. 829

40. Results of Spearman’s ρ Between CUF-RS and CUF-SG .......... 831

41. Results of Spearman’s ρ Between CUF-RS and DCCM .......... 832

42. Results of Spearman’s ρ Between CUF-SG and DCCM .......... 833

43. Results of Pearson’s r Between CUF-RS and CUF-SG ............ 836

44. Results of Pearson’s r Between CUF-RS and DCCM ............ 837

45. Results of Pearson’s r Between CUF-SG and DCCM ............ 838

46. Intraclass Correlation Coefficient Among All Three Methodologies ................................................................. 840
75. Response Analysis of Selected Health Status Preference Measurement Studies ................................................................. 898

76. Comparison of Modeling Studies on the EuroQol ............. 904

77. Estimates of Spearman's ρ Between Rating Scale and Standard Gamble Scaling Methods Reported in the Published Health Status Preference Measurement Literature .......... 938

78. Estimates of Pearson's r Between Rating Scale and Standard Gamble Scaling Methods Reported in the Published Health Status Preference Measurement Literature .......... 942

79. Functional Relationships Between Rating Scale and Standard Gamble Scaling Methods Reported in the Published Health Status Preference Measurement Literature .......... 949

80. Selected Recommended Minimally Acceptable Reliability Coefficients ................................................................. 966

81. Selected Results from Peterson's (1994) Meta Analysis ......... 969

82. Selected Results of Comparisons of Respondent Evaluations of Preference Measurement Methodologies in the Literature 987
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Relationship Between Quality of Life, Health-Related Quality of Life, Health Status, Functional Status, and Physiological Status</td>
<td>20</td>
</tr>
<tr>
<td>2. Charles and DeMaio's Conceptual Framework of Patient Involvement in Health Care</td>
<td>29</td>
</tr>
<tr>
<td>3. A Framework of Subjective Measurement</td>
<td>245</td>
</tr>
<tr>
<td>4. A Classification of Preference Measurement Strategies</td>
<td>249</td>
</tr>
<tr>
<td>5. Graphical Illustration of Attitude Toward Risk</td>
<td>475</td>
</tr>
<tr>
<td>6. Sample Card Describing a Multiattribute Health State</td>
<td>626</td>
</tr>
<tr>
<td>7. Feeling Thermometer Used in Present Study</td>
<td>629</td>
</tr>
<tr>
<td>8. Chance Board Used in Present Study</td>
<td>632</td>
</tr>
</tbody>
</table>
CHAPTER I

PREAMBLE

Chapter Overview

The purpose of this chapter is to set the stage for the study by introducing the topic of health status measurement -- what it is, why it is in vogue today, the current state of the art in health status measurement, what are its potential uses, and where it needs to go in order to fulfill its promise. The chapter begins with a brief discussion of the outcomes movement in health care. This is followed by a clarification of terminology, which is essential to differentiate between health status and other related terms used in the literature. The preamble continues with a discussion of the reasons behind the current interest in health status measurement, followed by an overview of the history of health status measurement. The next sections of this chapter summarize the state of the art in health status measurement and discuss the possible uses for such measures. Finally, some outstanding issues in health status measurement are discussed with the view of leading into the next chapter, which introduces the study in terms of the specific outstanding health status measurement issue which was investigated in this dissertation.
The Outcomes Movement in Health Care

The United States (US) health care system has experienced significant changes over the past half-century. Relman (1988) has described these changes as indicative of three eras in health care. The first of these, the “Era of Expansion,” began in the late 1940s and continued through the 1960s. It was characterized by rapid growth in the number of health-care providers and health-care delivery facilities, expanded insurance coverage of the population through Medicare and Medicaid, and extensive research and development by the private as well as public sectors. These expansive forces were instrumental in increasing per capita health-care spending dramatically, even after correcting for cost of living increases. This led to the incumbent President in 1969 proclaiming that the US faced a “health-care crisis.” The response to this crisis was a federal initiative aimed at restructuring the organization and incentives of the US health care system, which led to the second era in health care, the “Era of Cost Containment” (Relman, 1988).

This second era began in the 1970s and continued through the 1980s. It began with a trend and then a mass movement toward a health-care system influenced by market forces, incentive-based payment arrangements, and aggregation of providers (Relman, 1988). It was characterized by events such as the passage of the Health Maintenance Organization (HMO) Act, the implementation of Diagnosis Related Groups (DRGs), and the establishment of mail-order pharmacies. These cost containment measures led to increased pressure on health-care providers and delivery facilities in terms of not only financing but also decision making. Hospitals, formerly in a position to
collect whatever they charged, found themselves facing tougher buyers who now dictated prices. Some states instituted closely regulated global budgeting for hospitals, while others established tight controls on construction of new health-care delivery facilities and acquisition of equipment for existing facilities via Certificates of Need. Physicians, formerly the sole determiners of patient care decisions, found their decisions subject to second opinions and their practices subject to peer review. Several independent pharmacists, feeling the pressure of intense rivalry from mail-order pharmacies and restricted arrangements of managed care organizations, were forced to go out of business. Even the pharmaceutical industry, which seemed to be immune from these measures (as indicated by the high levels of profitability of its companies), became the subject of Congressional attention on account of its prices and profitability. The combined effect of these cost-containment measures has been described as transforming the traditional producer/provider-driven market to a payor-driven one (Anonymous, 1993).

However, if the intent of these measures was to control health-care spending, they met with limited success, since spending on health care as a percent of Gross Domestic Product (GDP) continued to increase (U.S. Bureau of the Census, 1993). Payor frustrations with the increasing expenditure on health care were compounded by the fact that there was no way to reliably and validly measure the end result or outcome of health-care interventions. Moreover, these measures did little to incorporate the input of the beneficiary of the health-care system, i.e., the patient. As Ellwood (1988) noted,
"Too often, payors, physicians, and health-care executives do not share common insights into the life of the patient. We acknowledge that our common interest is the patient, but we represent that interest from such divergent, even conflicting, viewpoints that everyone loses perspective. As a result, the health-care system has become an organism guided by misguided choices; it is unstable, confused, and desperately in need of a central nervous system that can help it cope with the complexity of modern medicine. The problem is our inability to measure and understand the effect of the choices of patients, payors, and physicians on the patient's aspirations for a better quality of life" (p. 1550).

Ellwood (1988) concluded that the result was concerned patients, frustrated health-care providers, skeptical payors, and besieged health-care executives. Patients expressed concerns about the impersonality of "corporate medicine" and effect of the increased sensitivity of payors, providers, and health-care executives to costs on the quality of care provided to them. Patients also stated that they were too uninformed to make appropriate health-care choices, and that the only power they sometimes had was to choose a health-care provider (Ellwood, 1988). However, they still did know if they made the right choice on such occasions, since they had no way of knowing which choice would yield the best results. Health-care providers shared patients' concerns about the effects of financial constraints on the quality of patient care provided by them. Payors, while expressing some satisfaction with the greater voice that the cost containment measures gave them, were still skeptical about the end result of the health care for which they were paying (Ellwood, 1988). In other words, they were demanding information about the value they got from their mounting expenditures on health care. Finally, health-care executives and corporate health benefit managers faced tough choices in the acquisition of expensive
new technology and drugs and reimbursement of different services. Like payors, they needed information about the impact of such technology and services on the overall health of the patient.

These practical concerns, coupled with conceptual ones deriving in part from the "health accounting" literature in the late 1970s [see, for e.g., Williamson (1978)], led to an increased emphasis on determining and measuring the impact of health-care products and services on the health of the patient. This is the essence of what Ellwood (1988) proposed as the "central nervous system" to help cope with the complexity of modern health care -- Outcomes Management. Ellwood (1988) defined outcomes management as "a permanent national medical data base that uses a common set of definitions for measuring quality of life to enable patients, payors, and providers to make informed health choices (emphasis added)." It is important to emphasize the purpose of measuring quality of life as suggested by Ellwood, i.e., to make decisions about health care. As explained later, this focus on decision-making as the rationale for the measurement of quality of life serves as the guiding light of this dissertation. According to Relman (1988), the move towards outcomes management was part of a larger change that the US health-care system is currently experiencing, aimed at measuring and justifying the value of health care to its stakeholders. Relman described this larger change as signaling a new era in health care, the "Era of Assessment and Accountability."

The Joint Commission on Accreditation of Health-care Organizations (JCAHO) now includes outcome criteria as part of its accreditation procedures for health-care organizations in the US (Joint Commission on
Accreditation of Healthcare Organizations, 1993). Until recently, the JCAHO had relied only on structure and process criteria to evaluate the quality of care in health-care organizations. The assumption behind using structure and process criteria for measuring quality stems from the Donabedian (1966) model of structure-process-outcome, which purports that if the structure and process of delivering health care in an organization are satisfactory, desirable outcomes should follow. However, several studies have shown that process and outcome are poorly correlated [for e.g., see Brook and Appel (1973); Nobrega, Morrow, Smoldt, et al. (1986)]. Given these findings, the emphasis on measuring outcomes seems long overdue -- if one of the objectives of the health-care system is to improve and maintain the health of the people (McDermott, 1981), how can one know whether the objective is being achieved if there is no way of measuring the outcome? Describing the concept of outcomes management in more detail, Ellwood (1988) notes, “The centerpiece and unifying ingredient of outcomes management is the tracking and measurement of function and well-being or quality of life.” The focus of this dissertation is the centerpiece of outcomes management, quality of life, or more specifically, one component of quality of life -- health status.

A Clarification of Terminology

Although most people intuitively understand the connotations of the term quality of life, it has often been used in the health-care literature without explicit definition. Consequently, a wide range of variables have been used as measures of quality of life, from physiological indicators such
as weight loss to standardized psychological measures of emotional distress (see Hollandsworth, 1988; Najman and Levine, 1981). The choice of these measures usually depends on "the objectives of measurement, the political motives of funding sponsors, and the particular concerns of the users, including the patients, clinicians, researchers, and others (Patrick and Bergner, 1990 p.166)." Often, the terms health status, functional status, quality of life, and health-related quality of life have been used interchangeably in the literature, when in fact they are different constructs. Hence it is necessary to distinguish between these terms. A logical starting point in this process is a definition of health, since the definition of health determines its operationalization and subsequent measurement.

Health

The word "health" is derived from the Old English word hoelth, meaning a condition of being safe and sound (Williams, 1934), denoting wholeness (Rogers, 1960). The simplest concepts of health are found in the dictionary. Webster's New World Dictionary (Guralnik, 1984) conceptualizes health as "physical and mental well-being," as "freedom from disease, pain, or defect," and as "normality of physical and mental functions," a "soundness." Ware (1987, 1991) has noted that dictionary conceptualizations of health include words like "completeness," "proper function," and "well-being." In household surveys, Americans have indicated that health is the "ability to do what they need to do" (Milio, 1983), a concept that implies proper functioning of the body and mind. These concepts, while a good starting point, are not specific enough for actual measurement in the sense
that they cannot be operationalized without some division of health into its component parts. In other words, what is needed is a definition of health as opposed to a concept. Definitions, as Hempel (1966) has noted, are "rules of replacement." Rudner (1966) has elaborated, "A definition means that a word or group of words (the definiens) is proposed to be truth-functionally equivalent to the word being defined (the definiendum)." Hunt (1987) has proposed seven criteria for evaluating definitions (constitutive as well as operational) -- inclusivity, exclusivity, differentiability, clarity, communicability, consistency, and parsimony. On the other hand a concept, as Woodruff (1966) has noted, is "a relatively complete and meaningful idea in the mind of a person."

Most scholars agree on the multidimensionality of health and the notion of health encompassing not only the absence of disease and disability, but also the ability to carry out normal tasks and activities and to maintain an overall sense of well-being (Hadley, 1982). This latter concept was advocated by Williams (1934), when he noted, "It is of value to think of health as that condition of the individual that makes possible the highest enjoyment of life, the greatest constructive work, and that shows itself in the best service to the world. .....

Health as freedom from disease is a standard of mediocrity; health as a quality of life is a standard of inspiration and increasing achievement (p. 18)." Larson (1991) has identified five general approaches to defining health: (a) the medical model, (b) the holistic model, (c) the wellness model, (d) the environmental model, and (e) the eclectic model. The medical model is based upon the perspective of illness, disease, and proper functioning. The holistic model adopts a "whole person"
approach, including physical, mental, and social health. The wellness model is concerned with “better than normal” states and progress toward a future state of health. The environmental model describes optimal interaction with the environment, while the eclectic model includes the definitions of health which do not fit into any of the other models (Larson, 1991).

It should be noted that there is no universal agreement on how many models of health exist. Abanobi (1986) has mentioned the first three models noted above -- the medical, holistic, and wellness models. Wylie (see Basch, 1978) offers two types of definitions for health -- the open-ended or “asymptomatic” definition, which is a description of ideal health, and the “elastic” concept in which health is the “ability to withstand stresses of disease” and is “related to interaction with the community and environment.” Hence the elastic concept combines elements of the medical model and the environmental model. Culyer (1983) has also proposed two models -- the medical model as described above and a “characteristic approach,” which “might include functional capacity (ability to perform physical activities), pain, emotional state, or, indeed, anything else deemed relevant for the purposes in mind.” Finally, Hunt, McEwan, and McKenna (1986) have conceptualized health as “absence of illness,” “strength and robustness,” or “high quality of life.” It can be seen that these conceptualizations are similar to the medical, wellness, and holistic models respectively. Therefore, the five models outlined by Larson (1991) appear to be quite inclusive of the different approaches to defining health that have been forwarded in the published literature. Hence, the following discussion will follow Larson’s (1991) classification of the five models of health.
The medical model has traditionally been the most widely used model in health care. Initially, it defined health as simply the absence of disease and used mortality rates to measure health. More recently, the medical model has incorporated morbidity and measures of physical functioning. This model allows the classification of diseases, either on an organic (i.e., infectious and communicable disease patterns) or a functional (i.e., acute and chronic) basis. This classification permits the conceptualization of health along a continuum from death to a state of optimal health, which is useful in program evaluation, i.e., prioritizing between competing programs on the basis of their effects on health. However, this model ignores the social and mental aspects of health which are salient but may not be manifest in measures of functioning. Critics of the model object to the overly negative orientation of the model and note that the absence of disease and infirmity might be necessary, but not sufficient, conditions for individuals to be healthy (see Marvin and Crown, 1976). Moreover, this model has been criticized for overlooking the impact of preventive medicine in its efforts to prevent disease and disabling conditions and improve the health of relatively well individuals (see, for e.g., Larson, 1991; Patrick and Bergner, 1990).

The holistic model includes the physical, mental, and social aspects of health and focuses on the positive aspects of health and well-being, as opposed to the negative aspects emphasized in the medical model. Hence, this model addresses some of the drawbacks of the medical model. According to this model, health in the positive sense, is “well-being, efficiency, and willingness to work (Culyer, 1983 p.33).” Although this
model seems similar to the World Health Organization (WHO) definition of health (World Health Organization, 1947) -- which will be discussed later -- it should be noted that the holistic model is more inclusive in the sense that it can incorporate elements that are external to the health-care system on account of their effect on health, e.g., spirituality or religious activity. Therefore its strict application in measuring health as an outcome of only the health-care system may be difficult to justify.

The **wellness model** focuses on excellence in health and progress toward a future state of health. The focus is on the individual as a system that is improving and moving toward excellence. Health is conceptualized as a state of feeling, "the layman’s intuitive notion of health as physical well-being -- comfort, energy, and ability to perform (Greer, 1986, p.11)." The notion of progressing toward a future state of health is important to note since it implies a dynamic aspect to health. Such a dynamic aspect is desirable since it provides cues for intervention by health-care professionals to affect the health of people. Further, the concept of a future health state is inherent in Pathak and MacKeigan’s (1992) definition of health-related quality of life (which is the adopted definition of health-related quality of life for this dissertation and will be discussed later). Critics of this model contend that (like the holistic model) it expands the meaning of health to include happiness, quality of life, and other global matters which may be beyond the purview of the health-care system (Bice, 1976).

The **environmental model** is influenced by systems theory. Systems theory analyzes the behavior of individuals or systems within the context of a larger environment (Larson, 1991). In the context of health, systems theory
relates the individual to an environment. The state of health of an individual is related to adaptability to the environment, and can therefore vary with environments. In defining health, an environmental model is used rather than a systems model, because the focus is on the relationship between the individual and the environment, not between the components within the system (Larson, 1991). The environmental model is represented by several definitions of health. For instance, Romano (1950) has defined health as “the capacity of the organism to maintain a balance in which it may be reasonably free from undue pain, discomfort, disability, or limitation of action including social capacity,” while Sigerist (1941) has defined health as “not simply the absence of disease ... something positive, a joyful attitude toward life, and a cheerful acceptance of the responsibilities that life puts on the individual.” Finally, Blum (1976) has defined health as “the degree of harmonious interaction among various system levels in the *homo sapiens* hierarchy.” This sampling of definitions of health illustrates the major drawback of the environmental model -- it is extremely hard to operationalize (e.g., how does one operationalize “joyous attitudes” or “harmonious interaction”?) and, consequently, may be of little use in measuring health as an outcome.

The **eclectic model** is a catch-all category for definitions that do not fit into any of the other categories. For instance, Dreitel (see Navarro, 1977) has defined health as the “capacity to help produce the very surplus the owners of the means of production appropriate.” Like the definitions of health in the environmental model, this is also very difficult to operationalize and use as a basis for measurement.
Given the shortcomings of each of the models outlined above, a more
desirable definition of health would be a hybrid one, which draws on the
strengths of each of the models (especially the medical, holistic, and wellness
models) while avoiding their pitfalls. One such definition might be the
WHO definition of health. The WHO's definition of health as “physical,
mental, and social well-being, and not merely the absence of disease or
infirmity” (World Health Organization, 1947) has both positive and negative
elements and accentuates the especially positive elements of health.
Moreover, it is amenable to operationalization -- the positive elements in the
definition can be measured by questions of well-being as perceived by
respondents on each of the dimensions (physical, mental, and social)
identified in the definition, while the negative elements indicate the
traditional measures of morbidity and disability. Further, by extending
beyond disease only, the definition allows for the existence of states of
health along a continuum from death (or worse) to perfectly healthy. As
mentioned, this conceptualization of health -- which was first forwarded by
Reed in 1949 (see Sullivan, 1966) -- is useful for the purpose of program
evaluation. In this sense, the WHO definition would be a hybrid between the
medical and holistic models. Moreover, it can incorporate the dynamism
implied in the wellness model if, as Pathak and MacKeigan (1992) note,
comparisons between present and reference health states (each defined
according to the elements in the definition) are made at different points in
time. In the 1950s, the WHO definition was criticized as being idealistic and
unmeasurable, but it is now widely accepted by health-care researchers and
practitioners alike. Indeed, it forms the definitional basis for most of the
instruments developed in the literature for the measurement of health. Therefore, the WHO definition of health will be adopted for this dissertation.

Health Status, Functional Status, and Physiological Status

The concept of health status follows from the definition of health, since health status is nothing but health as measured at a given point in time. Therefore, using the WHO definition of health, health status is represented by positive as well as negative elements. Positive elements include physical, mental, and social well-being, while negative elements include signs and symptoms of disease, illness, or discomfort. The positive elements are generally considered to be indicative of functional status, while the negative elements are considered indicative of physiological status. In this sense, functional status represents the holistic model of health, while physiological status represents the medical model of health. Hence, it can be seen that functional status and health status are not the same construct -- functional status is one basic dimension of health status.

Functional status comes closest in spirit to the lay concept of health -- "the ability to do what I need to do" (Milio, 1983) -- introduced earlier. Indeed, lay perceptions of quality of life usually include elements that are considered part of functional status as conceptualized above. For instance, DiCicco and Apple's (1958) survey of the elderly concluded that health was perceived in terms of activity (to be active was to be healthy) and poor health was when the state of health interfered with daily activity. Therefore, functional status measurement is essential in any patient-oriented approach
to measuring the outcomes of health care. It would seem that the American College of Physicians agrees with this observation, for in a recent position statement it noted, “The maintenance of the patient’s functional well-being is a fundamental goal of medical practice” (Health and Public Policy Committee, American College of Physicians, 1988). The other basic dimension of health status, physiological status, is essentially the traditional provider-driven view of health status. It is necessary to include physiological status because exclusive reliance on functional status may result in measures which are insensitive to the presence of many chronic conditions. This is because people may sometimes suffer from a disease or illness but still function normally. The point to note is that both these basic dimensions need to be included in any definition of health status. This is in keeping with Pathak and MacKeigan’s (1992) recommendation that in health status measurement, “at a minimum, physical functional status and physiological status should be measured” (p. 37).

In addition to these two basic dimensions, Ware (1987) also calls for including general health perceptions as part of health status because personal evaluations of health experiences are not completely captured by the other components. Personal evaluations of health experiences are important because they offer an overall assessment of how the patient feels. These perceptions may be able to tap into the “prognosis” aspect of health status (which is discussed later). Some studies of lay perceptions of health lend support to this notion. For instance, Baumann’s (1961) study of patients yielded three major dimensions of health – performance of activities of daily living (i.e., functional status), freedom from symptoms (i.e., physiological
status), and feeling-state (i.e., general health perceptions). In addition to these dimensions, as Pathak and MacKeigan (1992) have noted, if the intent of a study is to measure the impact of a particular intervention, additional relevant dimensions may be measured relating to the intervention in question.

It should be noted that there is no agreement in the literature on the specific dimensions of health status that should be included in health status measurement instruments. For instance, Kazis (1991) has recommended that in addition to physical, mental, and social well-being, measures of functional status should include general health perceptions, symptom perceptions, and sexual functioning. Feeny, Furlong, Barr, et al. (1992) have included seven dimensions in their conceptualization of health status -- sensation, mobility, emotion, cognition, self-care, pain, and fertility. On the other hand, the Quality of Well Being or QWB Scale (see Kaplan and Anderson, 1990) has only four dimensions -- mobility, physical activity, social activity, and a symptoms/problems complex, while the Sickness Impact Profile or SIP has 12 dimensions -- sleep and rest, eating, work, home management, recreation and pastimes, ambulation, mobility, body care and movement, social interaction, alertness behavior, emotional behavior, and communication (see Bergner, 1988). Finally, the revised version of the EuroQol instrument has five dimensions -- mobility, self-care, usual activities, pain/discomfort, and anxiety/depression (see Rosser and Sintonen, 1993).

This diversity is largely due to the purpose for which these instruments were developed. Indeed, as Pathak and MacKeigan (1992) have noted, some health-care researchers are interested in only the measurement
of health status *per se*, while others are more interested in explaining changes in an individual's health status (or its dimensions) due to health-care interventions. Hence, as Kirshner and Guyatt (1985) have noted, the content validity of health status instruments should be evaluated according to the specific purpose for which they are intended to be used. The particular dimensions of health status investigated in this study (along with other issues in the conceptualization and measurement of health status) will be discussed in Chapter V.

**Quality of Life and Health-Related Quality of Life**

The final issue that needs to be clarified is the distinction between quality of life (QOL) and health-related quality of life (HRQOL). QOL is a very broad construct. Campbell (1981) has identified 12 dimensions of life -- community, education, family life, friendships, health, housing, marriage, nation, neighborhood, self, standard of living, and work. Thus, in addition to health, there are many dimensions of life that can be affected by disease and treatment. Further, as Ware (1991) has noted, it is unclear as to what extent the health-care system can or should be held accountable for producing these QOL changes, which are the product of many forces in society which are beyond the purview of the health-care system. If QOL has to be used as a measure of outcome in health care, it is necessary to focus on those aspects of QOL which can be attributed directly to disease, illness, and any health-care intervention (preventive or curative). Hence, the outcome of interest to the health-care system is HRQOL.
Although the focus of this study is health status and not HRQOL, it is instructive to illustrate how the two constructs are related. Several different definitions of HRQOL have been proposed in the health-care literature (although they have not always observed the distinction between QOL and HRQOL). For instance, Patrick and Erickson (1988) have defined HRQOL as "the value assigned to the duration of life as modified by the social opportunities, perceptions, functional states, and impairments that are influenced by disease, injuries, treatments, or policy." This definition acknowledges the distinction between QOL and HRQOL and between HRQOL and health status, where the valuation procedure is the link between HRQOL and health status. However, it does conceptualize HRQOL in a rather static sense and is therefore not compatible with the wellness model of health described earlier.

Schipper, Clinch, and Powell (1990) have defined QOL as representing "the functional effects of an illness and its consequent therapy upon a patient as perceived by the patient." This definition concentrates mainly on functional status at the cost of ignoring physiological status. As explained previously, both need to be considered in any conceptualization of health status. Moreover, this definition seems to equate health status and HRQOL. It also suffers from the static conceptualization of HRQOL like the Patrick and Erickson definition. Shumaker, Anderson, and Czajkowski (1990) have defined QOL as "an individual's overall satisfaction with life and general personal well-being." This definition confounds affective responses like satisfaction with the cognitive ones needed to evaluate HRQOL. Like the other definitions, it also views HRQOL in a static light. On
the other hand, Calman (1984) has proposed a dynamic definition of QOL. According to Calman (1984), QOL "measures the difference, at a particular moment in time, between the hopes and expectations of the individual and that individual's present experiences." This definition, however, confounds the terms hopes and expectations. It implies that these terms are interchangeable when, in fact, they are not (see Cribb, 1985). However, it should be noted that this definition explicitly incorporates the cognitive aspect of evaluation of HRQOL evaluation.

The definition of HRQOL adopted for this dissertation is the one proposed by Pathak and MacKeigan (1992): "Health-related quality of life can be defined as a comparative judgment based on a point-in-time assessment of an individual's present health state relative to that individual's reference health state(s)" (p.37). This definition integrates the strengths of some of the other definitions while avoiding some of their pitfalls. By choosing to not define health status, it leaves the choice of dimensions to be included in health status assessment up to individual researchers, which is desirable since different patient populations may have different conceptualizations of the specific dimensions of health status. This allows elements of both, the medical as well as the holistic models of health to be included in the assessment of health status. Moreover, by explicitly separating health status and HRQOL, it incorporates the cognitive aspect of HRQOL evaluation, which most available instruments purporting to measure HRQOL ignore. Finally, it is consistent with the wellness model of health, since comparisons at different points in time could be made, thus providing a dynamic picture of HRQOL.
NOTE: Health status can include other dimensions as deemed relevant

Figure 1: Relationship Between Quality of Life, Health-Related Quality of Life, Health Status, Functional Status, and Physiological Status
The relationship between the constructs discussed in this section is shown in Figure 1. As noted, the emphasis in this study was on health status.

**Health Status Measurement -- Why the Interest?**

The current interest in measuring health status reflects the coalescence of several related trends. The trends may be categorized as medical, economic/regulatory, and social trends.

**Medical Trends**

An overriding factor which has contributed to the increased interest in measuring health status is the gradual change in the focus of primary medical care from decreasing mortality to limiting morbidity and the patient-reported impact of that morbidity (Coons and Kaplan, 1992). This change has been brought about mainly due to a change in the nature of the medical conditions observed in society, which have shifted from a predominance of acute and infectious diseases to a predominance of chronic conditions. Acute and infectious diseases, for the most part, can be treated. Hence the focus of primary medical care for acute and infectious diseases is treatment or cure. Since these diseases are often fatal if not cured, the outcome of health care for such diseases is usually measured in terms of mortality.

Although many acute and infectious diseases remain that can shorten life expectancy (most notably cancer and Acquired Immuno Deficiency Syndrome or AIDS), the majority of medical conditions faced by society
today are of a chronic nature. In chronic illness, the focus of primary medical care is more often a reduction in morbidity than cure or prolongation of life. In a seminal article, McNeil, Weichselbaum, and Pauker (1978) demonstrated the fallacy of using prolongation of life as the criterion for choosing between therapies in chronic illness. Using lung cancer as an example, these investigators interviewed patients with operable lung cancer to determine their attitudes toward two alternative therapies -- surgical extirpation and radiation. Surgical extirpation offered increased five-year survival and prolonged life expectancies compared to radiation, but exposed patients to the immediate risk of thoracotomy. Using only prolongation of life as the decision criterion, the therapy of choice would be surgical extirpation. However, when the expected survival figures for each alternative were adjusted for patient attitudes toward the risk of thoracotomy, it was found that radiotherapy would be the therapy of choice for many patients (McNeil, Weichselbaum, and Pauker, 1978). These results illustrate the importance of looking beyond objective measures of survival in chronic illness.

Even for those conditions where the primary objective is prolonging life, morbidity is an important outcome because patients may suffer from adverse effects of the therapy (MacKeigan and Pathak, 1992). Traditionally, morbidity assessment has relied on physiological measures obtained during laboratory tests or medical examinations (i.e., measures of physiological status). These measures, although objective, do not necessarily reflect changes in the way the patient feels and/or functions (MacKeigan and Pathak, 1992). For example, plasmapheresis has been shown to alter
serologic indicators of rheumatoid arthritis such as erythrocyte sedimentation rate without improving pain, grip strength, or walking ability of patients (Dwosh, Giles, Ford, et al., 1983). There is acceptance now that physiological responses to treatment for heart failure do not always translate into functional improvement (Schipper, 1983). Sole reliance on such measures, then, might miss clinically relevant changes in patients. In such cases, a measure of functional status would be more appropriate. Recognizing these factors, the Oncologic Drugs Advisory Committee of the Food and Drug Administration (FDA) has recommended that health-related quality of life -- actually health status according to the clarification provided in the previous section -- be included as a key parameter (along with survival) to be considered in the approval of new anticancer drugs (see Johnson and Temple, 1985; Shoemaker, Burke, Dorr, et al., 1990).

Further, there are behavioral consequences that come into play in such cases. Since chronic illnesses usually require lifelong therapy, the role of the patient in following the therapy becomes crucial in determining the effectiveness of therapy. If patients cannot see or feel a noticeable improvement in their functioning -- which, as mentioned previously, is their frame of reference for their concept of health (see Milio, 1983) -- they may be inclined to discontinue or be noncompliant with therapy, which could have deleterious consequences for the health of the patient. In other words, patients need a measure of health which is centered around how they feel and function since they may be unlikely to accept any therapy -- whatever its scientific merit -- if they see nothing in it for themselves.
The merit of these arguments has not been lost on health-care providers, who are becoming increasingly convinced that one of the primary objectives of medical care is enhancement of health status and health-related quality of life (see Frieds, 1980; Revicki, 1990). In the profession of pharmacy, the focus on patient outcomes is central to the recently proposed mission of pharmacy practice called pharmaceutical care, "The responsible provision of drug therapy for the purpose of achieving definite outcomes that improve a patient's quality of life" (Hepler and Strand, 1990). The recent initiative by the Agency for Health Care Policy and Research or AHCPR (1992) should increase the focus on these outcomes of health care. Further, the role of the so-called "me-too" pharmaceuticals in bringing about this acceptance cannot be discounted. Although these me-too pharmaceuticals may be equally effective when compared according to the traditional clinical and physiological parameters, they may have different effects on health status. In this sense, the selection of drug therapy may depend on the documented differential impact of alternative drugs on health status.

The pharmaceutical industry has been quick to note this and is sponsoring and conducting several studies examining the impact of its products on the health status of patients. A 1988 Institute of Medicine Survey found that 66% of the pharmaceutical companies surveyed stated they were currently using or plan to incorporate such measures in clinical trials of new pharmaceuticals (Luce, Weschler, and Underwood, 1989). Apart from traditional clinical trials, these measures are also being incorporated into post-marketing studies comparing different marketed drugs in a variety of practice settings. The pharmaceutical industry is also actively engaged in
establishing entire departments to conduct and manage such research (Revicki, Rothman, and Luce, 1992). Health-care providers can therefore expect to see increased use of health status and health-related quality of life claims by pharmaceutical marketers. It is possible, then, that health-care providers may be receptive to such claims and use these measures to a greater extent in medical decision making. In this way, the medical trends discussed above have contributed to the increased interest in health status measurement.

Economic/Regulatory Trends

The economic/regulatory trends that prompted interest in health status measures have been mentioned earlier, in the discussion of the “Era of Cost Containment.” To reiterate, payors and health-care executives were increasingly demanding proof of the value they were getting from their increasing health-care expenditures. In the US, the high prices of prescription drugs and high profitability of pharmaceutical companies in the 1980s was (and is) the subject of intense Congressional attention. In order to demonstrate the value of their products, pharmaceutical manufacturers and marketers have sponsored and conducted pharmacoeconomic analyses of their products. Pharmacoeconomics has been defined as “the description and analysis of the costs of drug therapy to health-care systems and society” (Townsend, 1987). Pharmacoeconomic research usually consists of the description, analysis, and evaluation of costs and outcomes associated with the use of pharmaceutical products, programs, or services (Boyer and Pathak, 1994). This research involves the investigation of the efficiency of
resource allocation among competing alternative medicines, programs, or services for delivering pharmaceutical care (Boyer and Pathak, 1994).

Although the term "pharmacoeconomics" is of recent origin, the principles of economics have been applied to health care for many decades. The initial interest was in program evaluation, wherein decisions were made about prioritization among alternative programs. Incorporating health status measures into economic analyses was initially proposed in the early 1970s (see, for e.g., Bush, Fanshel, and Chen, 1972; Fanshel and Bush, 1970, Patrick, Bush, and Chen, 1973). This was based on the notion that health status was one component of an index of health, the other being duration of life. By combining these into one index, it was argued that different health-care interventions could be compared in terms of their overall effect on the quality and quantity of life. This idea led to the concept of quality-adjusted life years (QALYs), which form the denominator in the unit of comparison (i.e., cost/QALY) used in cost-utility analyses. Cost-utility analyses are one of the two types of pharmacoeconomic evaluations in which some effort is made to incorporate health status measures into the unit of comparison between alternatives. The other type is cost-benefit analysis, wherein health status can be valued in currency amounts (e.g., US dollars) and used as the denominator in the unit of comparison between alternatives.

However, most of the pharmacoeconomic evaluations that have been conducted have used other measures as the denominator in the unit of comparison between alternatives, usually physiological indicators (in cost-effectiveness analyses). Since these other measures do not capture the outcomes of therapy from the patient's point of view, users of this research
are demanding that patient-centered outcomes, such as health status measures, be incorporated into the analyses. From a financial point of view, pharmaceutical marketers can use health status information to increase sales of their products. For instance, Jack (1991) has estimated that the combined effects of expanding the sales force and changing the labeling conditions of Capoten® (captopril) to incorporate information about health status were instrumental in increasing the sales of this product by approximately US $375 million over a three-year period.

Policy makers are also driving the interest in health status measurement from a regulatory perspective. Recent steps taken by the Ontario Formulary Committee (see Detsky, 1993) and the Australian Pharmaceutical Benefits Advisory Committee (see Henry, 1992) indicate that economic analyses will play an increasingly important role in the reimbursement of pharmaceuticals in global market. Although the US FDA does not currently have any such requirements, the recommendation of the Oncologic Drugs Advisory Committee that health-related quality of life data be considered along with survival data in the approval decision of anticancer drugs (see Johnson and Temple, 1985) may be indicative of future FDA policy. Some manufacturers are already submitting such data along with the regular information of Phase I, II, and III trials to the FDA in apparent belief that the additional information may assist them in the regulatory process by demonstrating additional patient outcomes not accounted for in the traditional safety and efficacy data (Henderson-James and Spilker, 1990).
Social Trends

The notion of patient involvement in health care has been documented for some time (see Hamilton, 1982). Charles and DeMaio (1993) have provided a conceptual framework of patient involvement in health-care decision making that defines involvement in three dimensions -- decision making domains, role perspectives, and level of participation. This framework is presented in Figure 2. Charles and DeMaio have drawn a distinction among three decision making domains -- treatment, service delivery, and broad macro or system level decision making contexts. These three domains reflect an underlying continuum from individual to system levels. Treatment domains are at the individual patient level, service delivery domains are at the community level, while macro domains (as the name implies) are at the general system level. Charles and DeMaio acknowledge that the three sub-domains are not entirely discrete (Charles and DeMaio, 1993).

The second dimension, role perspectives, has two levels -- user and public policy. This distinction is important because each role perspective incorporates different attitudinal assumptions and expectations which individuals bring to a particular decision making context (Charles and DeMaio, 1993). The user perspective reflects a concern with the potential impact on the individual or any person/group with which the individual feels a special affinity. The public policy perspective, on the other hand, reflects a more altruistic concern with some notion of a broader public or community good (Charles and DeMaio, 1993).
NOTE: Role Perspective on the x-axis; Decision Making Domain on the y-axis; Level of Participation on the z-axis

Figure 2: Charles and DeMaio's Conceptual Framework of Patient Involvement in Health Care [adapted from Charles and DeMaio (1993)]
The third dimension, level of participation, refers to the extent to which individuals have control over the decision making process. Drawing on the work of Arnstein (1969) and Feingold (1977), Charles and DeMaio (1993) have suggested three categories of decision making control -- consultation, partnership, and patient control. According to Charles and DeMaio (1993), consultation provides an opportunity for patients to express their view but offers no guarantee that individual views will be taken into account, partnership involves shared planning and decision making responsibilities between patients and other decision makers, while patient control occurs when patients are in full charge of decision making.

This conceptual framework is useful in understanding the role of social trends in influencing the current interest in health status assessment. The traditional role of the patient in the health-care system was at most consultative, both in a user as well as a public policy perspective. System and community level decisions were made by the appropriate regulatory agencies and committees without much input of patients, as were individual treatment decisions by physicians. However, the consumerism movement of the 1960s (see Buskirk and Rothe, 1970; Day and Aaker, 1970; Herrmann, 1970; Reeder, 1972) was instrumental in bringing about changes in the traditional role of the patient.

The essence of the consumerism movement was the conceptualization of the patient as a consumer rather than a client (see Bashshur, Metzer, and Worden, 1967; Campbell, 1971; Hochbaum, 1969; Thursz, 1970). It is important to note the difference between a client and a consumer. As a client, a patient delivers himself or herself into the hands of the professional
who is usually the sole decision maker regarding the nature of the service to be provided (Reeder, 1972). In such a case, the patient's participation is at best in a consultative capacity. On the other hand, as a consumer, a patient is viewed as an active user of services who tends to be guided by *caveat emptor* (Reeder, 1972). Although in principle this means that patients have total mobility and freedom to choose their health-care providers, the reality of the health-care system (especially the role of insurance and asymmetry of information) suggests otherwise. Therefore, in health care, consumerism could potentially move the patient into a partnership level of participation -- not patient control. However, as Brody (1980) has noted, not all patients have the desire and/or capability to participate in individual clinical decision making, nor will all physicians possess the time, interest, sensitivity, or communication skills needed for fostering patient participation.

It should be noted that a rather strict interpretation of the term consumer (i.e., as user) was adopted in this discussion. If one adopts the more liberal interpretation of the term (i.e., user and payor), consumers would include patients as well as payors such as insurance companies, corporate health benefit managers, and the government. Since the effect of payors in increasing interest in health status measures has already been dealt with as an economic/regulatory trend, it was decided to adopt the more conservative interpretation here.

The initial effects of consumerism in health care (in the early 1970s) were mainly felt at the public policy-macro and public policy-service dimensions (cells D and F in Figure 2) of Charles and DeMaio's framework.
Many government supported programs during this period mandated patient involvement in the planning and management of health-care services in the belief that such involvement would increase provider responsiveness to the goals and needs of users of health-care services (Metsch and Veney, 1973). However, it was found that consumer involvement did not meet its intended purpose due to several reasons related to the implementation of the involvement mechanisms as well as the nature of the provider-consumer interaction (see Metsch and Veney, 1976). There are several published studies on the elicitation and incorporation of patient attitudes and values in public policy decision making (see, for e.g., Pauker, Pauker, and McNeil, 1981; Sackett and Torrance, 1978). Since patient attitudes and values in these studies were measured as a result of the researchers' belief that they needed to be measured (and not because patients demanded their attitudes and values be measured), these studies will not be reviewed.

During the 1980s, patients' demands for participation in individual clinical decisions (cell A of Figure 2) increased. Apart from the influence of consumerism in general, there were some other specific reasons for this increased demand for participation. A major reason was the increased incidence of chronic conditions (as discussed previously) and the consequent focus on prevention and behavior modification as opposed to cure. Since patients play a greater role in controlling chronic conditions (as chronic conditions usually require lifestyle adjustments in order to be controlled) and because chronic conditions are usually lifelong, patients have a vested interest in increasing their involvement in decisions about their therapy which they will probably be taking for life and which will have long-lasting
effects on them. As a consequence of the expanded role of patients in controlling chronic conditions, the self-care movement gained in popularity. Rising education levels in the population and increased access to technical information through multiple media sources may have also contributed to the increased desire for participation (Levin, Katz, and Holst, 1976). Moreover, patient cost-sharing mechanisms were introduced and grew in popularity in the 1980s. It is logical for patients to demand more involvement in health care if they are bearing an increased responsibility for financing it. Finally, it has been suggested that the increasingly technological nature of the health-care system may have been responsible for this desire (see MacKeigan and Pathak, 1992), which has been expressed by Naisbitt (1982) as “hi-tech-hi-touch: whenever a new technology is introduced in society, there must be a counterbalancing human response.” This may explain the increasing popularity of home health care and alternative medicine that has been seen in recent times (see DerMarderosain, 1994; Eisenberg, Kessler, Foster, et al., 1993; Furnham and Smith, 1988).

The rationale and importance of formally introducing patient attitudes into medical decision making processes was clearly stated by Arrow (1973): “The economist and the decision analyst address themselves to ... the implementation of values among feasible alternatives. ... The choice may be ... between two forms of treatment, one of which reduces the overall probability of death but increases the probability of an earlier one.” As Pauker and McNeil (1981) have noted, in such cases the preferences of individual patients need to be considered. Goldsmith (1984) has also commented on the changing role of the patient in modern health care when
he noted, "According to physicians, patients are increasingly vigilant and demanding. They are increasingly suspicious of medical direction, particularly women. They are increasingly specific about their needs and how they want them met, and increasingly sophisticated regarding medical technology, and diagnostic and therapeutic options. Consumer surveys suggest that patients (particularly younger ones) are increasingly willing to switch physicians and to shop for physician care on the basis of price (p.11)."

Given this increased demand of patients to participate in individual clinical decision making, it is necessary to know the effect of this increased participation on the outcomes of these decisions. Kleinman, Eisenberg, and Good (1978) have observed that the failure to directly elicit patient attitudes and desires about their care can have deleterious effects on therapeutic success. Starfield, Steinwachs, Morris, et al. (1979) have shown than patients report greater problem improvement when both the practitioner and the patient believe that problems require follow-up, as opposed to the outcomes of problems that only the patient considers worthy of a follow-up visit. Greenfield, Kaplan, and Ware (1985) found that increased patient participation (in terms of reading their medical records, asking physicians questions about their condition and therapy, and negotiating medical decisions with physicians) was responsible for better patient outcomes. More specifically, participants reported fewer limitations in physical and role-related activities and were as satisfied with their care as non-participants (Greenfield, Kaplan, and Ware, 1985). In another study, Greenfield, Kaplan, and Ware (1988) found that when diabetics were given access to their
medical records, it was associated with better regulation of their blood glucose levels. These researchers explained their findings by hypothesizing the existence of an intervening variable between participation and patient outcomes like physical and role-related activities -- "sense of control."

Mahler and Kulik (1990) have suggested two mechanisms through which sense of control could be associated with positive outcomes. First, patients who feel they have control may suffer less emotional disturbance, thereby experiencing better heart functioning -- since psychological upset in general and anxiety in particular have been linked with the development and duration of arrhythmias (see Alpert, 1980) -- and therefore better functional status. Alternatively, perceived control may be associated with following physician advice, thus leading to better outcomes. In their study, Mahler and Kulik (1990) found greater sense of control to be associated with shorter hospital stay following a coronary-bypass operation. Moreover, studies have shown the existence of a link between sense of control and physiological status (see, for e.g., Imboden, 1972; Pennebaker, Burnam, Schaeffer, et al., 1977; Sklar and Anisman, 1979) and elements of functional status such as ability to tolerate pain (see, for e.g., Bowers, 1968; Langer, Janis, and Wolfer, 1975) and effective daily functioning (see, for e.g., Langer and Rodin, 1976). However, greater perceived control may also lead to noncompliant or undercompliant patient behavior as a form of reasoned decision making (Donovan and Blake, 1992).

It is necessary to remember that there are different steps in individual clinical decision making. The question that arises, then, is in which step(s) do patients desire to be involved? Strull, Lo, and Charles (1984) have
conceptualized the individual clinical decision making process as consisting of three major steps -- disclosing information about the disease or condition, discussing therapeutic alternatives, and making the final decision. It should be noted that these steps are distinct from the levels of patient participation visualized by Charles and DeMaio (1993), as each step can occur at all three levels of participation.

Strull, Lo, and Charles (1984) surveyed 210 hypertensive outpatients about their preferences for participation in these different steps and their 50 clinicians about their appreciation of patients' preferences for participation. Regarding the disclosure of information, it was found that although most (52%) patients reported receiving considerable information about their condition (more, in fact, than what clinicians reported giving), about 41% preferred receiving additional information about their condition. On the matter of discussion of therapeutic alternatives, the majority of patients (55%) preferred extensive discussion. Clinicians were found to be likely to underestimate this demand for discussion. However, regarding actual decision making, nearly half (47%) the patients wanted the clinician to make the decision using "all that's known about the medicines," while 31% wanted the clinician to make the decision but "strongly consider the patient's opinion." Less than one-fifth (19%) of the patients wanted to make joint decisions with clinicians. Clinicians tended to overestimate patient preferences for participation in actual clinical decision making. It was also found that patient preferences for participation declined as the severity of their condition increased (Strull, Lo, and Charles, 1984). Therefore, this study found that patients desired information about their condition and
wanted to discuss alternative therapies with their physicians, but did not desire to be very involved in actual clinical decision making. Other studies investigating patient preferences for participation in specific steps of individual clinical decision making have generally supported the results of the Strull et al. study in different patient populations (see, for e.g., Degner and Sloan, 1992; Ende, Kazis, Ash, et al, 1989; Sutherland, Llewellyn-Thomas, Lockwood, et al., 1989). This is quite a departure from the assumptions of policy makers and clinicians, who are reported to have felt that increased demands for patient participation in individual clinical decision making meant that patients wanted to make final therapy decisions on their own.

These results, coupled with Goldsmith’s (1984) observations which were quoted earlier, suggest that although patients may be very vocal about their needs and how they want them met and are eager to discuss alternative therapeutic options with their physicians, they still prefer physicians to make the final therapy decisions on their behalf. In other words, patients seem to prefer a level of decision making control in between the consultation and partnership categories of Charles and DeMaio’s (1993) framework. This middle level is one in which patients want to let the physicians know their preferences and discuss therapeutic alternatives with physicians, but let the physician make the final therapy decision on their behalf after incorporating their preferences and on the basis of the discussion with them. This would suggest a level along the lines of what Szasz and Hollender (1956) have termed a "guidance-cooperation" model of physician-patient behavior, where the physician offers guidance to the patient in terms of what to do on the basis
of a discussion with the patient. The patient, then, is expected to cooperate and follow this guidance. It should be noted that the patient is far from a helpless pawn in this model, for he or she can actively decide whether or not to follow the physician's guidance.

Given this observation, physician understanding of the preferences of patients becomes critical to success of therapy. If physicians understand what patients prefer, they could personalize therapy to suit these preferences and thus be responsive to patient preferences. Patients, on the other hand, are likely to be more compliant with personalized therapy. Assuming that the therapy in question is more effective in compliant patients than noncompliant or undercompliant ones, it would be successful in achieving the desired patient outcomes, i.e., improved health status. This improved health status would then correspond more closely to the patient's reference state [as conceptualized in the Pathak and MacKeigan (1992) definition of health-related quality of life, which was discussed earlier]. This in turn would improve the patient's health-related quality of life.

The importance of incorporating patient preferences in clinical decision making is emphasized in the recently released AHCPR clinical practice guideline for diagnosis and treatment of benign prostatic hyperplasia (BPH), in which it is specifically stated that it should be the preference of the patient with moderate symptoms who has not yet developed complications and not the preference of the physicians that is most important (Agency for Health Care Policy and Research, 1994). To reiterate this point, the BPH Panel Chair noted, "The individual patient's view of his symptoms and the benefits versus risk of treatment should be the
prime determinants of both the need to treat as well as the way to treat (reported in Anonymous, 1995 p.13)."

It should be noted that this entire line of reasoning rests on the physician understanding the preferences of the patient. This begs the question, "Preferences for what?" The patient enters the formal health-care system for a particular reason, usually the maintenance of good health or the need to suppress or cure a particular condition or disease. Hence, the answer to the question, "Preferences for what?" lies in the end result or outcome of the reason for the patient entering the formal health-care system. In other words, the preferences of patients that physicians need to understand are the preferences for the outcomes of the care patients receive when they enter the formal health-care system. Since the end result or outcome of the health-care system is health, physicians then need to understand patients' preferences for health. Assuming that patients prefer to be healthy as opposed to unhealthy (i.e., in a higher state of health along the continuum of health states) and keeping in mind the layperson's definition of health (Milio, 1983), the common language between patients and physicians is health status. However, a necessary prerequisite to understanding patient preferences for health states is the reliable and valid measurement of health status. In this way the social trend of increasing demands for patient involvement in health-care decision making has increased interest in health status measurement.

A salient force driving the above trends was the development of sophisticated measurement techniques and their application to health care, which enabled the interest in measuring health status to be realized. This
was made possible by the efforts of researchers from the fields of economics, psychology, decision analysis, and operations research. To understand the role these techniques have played in health status measurement, it is necessary to review the history of health status measurement.

**Health Status Measurement -- A Historical Perspective**

Although the current interest in measuring health status may reflect the coalescence of the medical, economic/regulatory, and social trends discussed above, the development of health status measures dates back to the 1940s. The initial interest in health status measurement was driven by regulatory forces for use in program evaluation. Program evaluation includes both, resource allocation decisions as well as monitoring progress (or lack thereof). For both purposes, the type of instrument required was one which could give an overall assessment of the effect of the program(s) in question on the health of the general population. Since then, however (mainly due to the medical and social forces discussed earlier), increasing attention has been focused on the development of health status measures for use in clinical decision making. For this purpose, the emphasis is more on the particular dimensions of health status deemed clinically relevant and the demonstration of clinically relevant change over time. These goals have largely been realized due to advances in measurement sophistication in health status research. The development of health status measures can be conceptualized as comprising of two periods -- the early period and the modern period.
The Early Period

The early period began in the late 1940s and extended up to the early 1970s. Goldsmith (1972, 1973) and Hennes (1972) have provided excellent reviews of this period. Traditionally, the “health status” of a community was measured using population statistics regarding death (e.g., infant mortality rates) or disease (e.g., incidence of coronary heart disease). In addition to the basic age-adjusted death rate, several other population-based measures were used, such as proportional mortality and the unnecessary death index or UDI (Hennes, 1972). Proportional mortality is the number of persons who have died from a given disease over a specific age (e.g., 50 years) as a proportion of all deaths from that disease. It is based on the premise that the older one is when one dies, the healthier one may be considered (Swaroop and Uemura, 1957). The UDI is the ratio of unnecessary deaths to all deaths for a given disease and geographic region. Unnecessary deaths are determined by taking the average adjusted rate from the lowest rate regions and applying that average rate to the population under study. The result is the expected number of deaths for the region, and all deaths beyond this number are considered unnecessary (Guralnick and Jackson, 1967).

It can be seen that both these measures are based on faulty premises. Regarding the proportional mortality measure, the premise of age being positively correlated with good health is not a truism, especially in the case of chronic illnesses. Further, age may be negatively correlated with some dimensions of health status such as functional status. Regarding the UDI, the basis of comparison is flawed -- to apply average mortality rates from one region blindly to another implicitly assumes that the two regions are
equivalent in all other respects, which may not necessarily be true. According to the clarification of terminology provided earlier, neither of these indicators measure health status. Moreover, these measures focus unduly on mortality. Moriyama (1968) has noted the fallacy of using mortality-based indicators as measures of health in a society which faced more chronic conditions than acute and infectious diseases: "... by 1950 the mortality from the diseases of infectious origins had reached a level where death rates for the infectious diseases of infectious origins no longer contributed in a major way to the overall mortality rate ... Further reductions in total mortality in the United States are possible, but any substantial decreases must come from the lowering of the death rates from chronic noninfective diseases and for accidents and other violence. ... The nature of past changes in mortality and the past behavior of the death rates have made moot the value of statistics of deaths from all causes as a measure of health in countries like the US" (Moriyama, 1968 p.576).

In apparent agreement with this observation, several proposals were made to broaden vital statistics to include the five "d’s" -- death, disease, disability, discomfort, and dissatisfaction (White, 1967). In the following years, several health indices were constructed by different researchers, including those Chiang (1965), Sullivan (1971), Meyer (1971), and the Q index (1970). However, it has been noted that most of these measures are overly influenced by the medical model of health and end up being measures of sickness (as opposed to health) or sophisticated ways of assessing the absence of disease (Hennes, 1972). As Logan (1964a, 1964b) has noted, "So, while it would be ideal to measure the prevalence of 'well-
being, what is usually done in estimating and comparing the health of populations is to measure the deaths from various causes, the extent to which death is avoided, the absence of a demand for medical care, or, at least, the absence of an obvious need for medical care."

While the health status measures for program evaluation developed in this period may have been overly influenced by the medical model of health, the measures developed for use in clinical settings showed more balance between the negative and positive aspects of health. Measures for use in clinical settings were of two general types -- those measuring physical function and physiological status and those measuring social and mental function. Measures of physical function and physiological status were usually developed by clinicians to collect and record systematically the information they thought to be important for diagnosis and treatment (Ganz, 1990). On the other hand, measures of social and mental function were generally developed by social scientists. This difference is probably indicative of the difference in purpose and research interests of the developers.

Budd, Bleich, Boyd, et al. (1970) have reviewed nearly 20 of the physiological status measures used in the early period. Some of the more commonly used measures included the Cornell Medical Index (Brodman, Erdmann, Lorge, et al., 1949), the Mayo Clinic Questionnaire (Martin, Mayne, Taylor, et al., 1969), and the Kaiser Permanente Patient Inventory Questionnaire (Collen, Cutler, Seigelaub, et al., 1969). These measures were intended to be proxies for a physician and obtain diagnostic symptom information useful for patient care. However, as Hennes (1972) has noted,
they are more illness questionnaires than health questionnaires. In other words, they are products of the medical model of health. This is probably because they were developed by clinicians, whose training emphasized the medical model over other models of health. Moreover, although widely accepted by clinicians, the reliability and validity of these instruments was not the subject of much investigation. The few studies on the reliability and validity of these instruments were not very supportive of these measurement properties of the instruments (Hennes, 1972). Most of them are further limited by the fact that they are rated by clinicians rather than patients. Patient reports are desirable since they are more likely to include symptoms as well as signs, as opposed to clinician reports which usually only include signs. As Jette (1980), has noted, signs are directly observable events which may be seen by others as evidence of illness (e.g., elevated blood pressure) while symptoms are subjective phenomena experienced by individuals which are not directly observable by others, but rather only become known through the report of the individual being assessed (e.g., pain, dizziness). In order to get a complete representation of physiological status, it is therefore to obtain information about signs as well as symptoms. Hence, patient reports are usually preferred.

The early period also witnessed the initial development of measures of physical, mental, and social function. One of the earliest measures of physical function was the Karnofsky Performance Status (Karnofsky and Burchenal, 1949). A number of other instruments measuring physical function were developed in this period, especially for application in the elderly, such as the Activities of Daily Living Scale (Katz, Ford, Moskowitz,
1963), the Shanas Incapacity Index (Shanas, Townsend, Wedderburn, et al., 1968), and others (e.g., Berg, Browning, Hill, et al., 1970; Kelman and Willner, 1962; Pihlblad and Rosencranz, 1967; Simon, 1961; Zeman 1945). These measures are noteworthy because they represent some of the earliest efforts to operationalize health in terms of the elements of models other than the medical model of health. Indeed, some of them have been incorporated into later health status instruments which measured functional status. However, as Hennes (1972) noted, there was little research conducted on the measurement properties of these instruments and on comparing or relating these instruments which purported to measure the same construct (i.e., physical function).

Apart from developing measures of physical function, researchers in this period also made progress toward developing measures of mental and social function. The most commonly used instruments measuring mental function in this period were the Langner Screening Scale (Langner 1962; Langner and Michael, 1963) and the Bradburn Index of Psychological Well-Being (Bradburn 1969). The Bradburn Index is especially noteworthy because it included scales for both positive and negative affect, which were found to be two independent measures. Regarding social function, Roen and Burnes (1968) developed the Community Adaptation Schedule, which attempted to measure an individual's perception of the community, affects toward it, and behaviors in it. Social performance in everyday life was measured in the areas of work, family, the social community, the larger community, and the professional community (Roen and Burnes, 1968). Other measures for measuring mental and social function which were published in
the early period included the Hessler scale of emotional and social health (Hessler, New, Kubish, et al., 1971) as well as the Pihlblad and Rosencranz scale on life satisfaction and social participation (Pihlblad and Rosencranz, 1969).

Measures of mental and social function which were developed during this period tended to be more carefully investigated about their measurement properties as compared to measures of physiological status and physical function (Hennes, 1972). This is probably due to the social science background of the developers of such measures, since the principles of measurement for such instruments were developed in the social sciences. For instance, Fabrega and McBee (1970) compared the Langer Screening Scale with clinical psychiatric evaluation ratings. They found the scale to relate significantly to anxiety, depression, and neuroticism, but not to other dimensions of psychiatric evaluation like disorganization or regression, nor to clinical ratings of social and vocational impairment. Items representing positive affect in Bradburn’s Index of Psychological Well-Being were found to correlate positively with indicators of social involvement and change, while items representing negative affect were found to correlate with mental health measures such as anxiety (Hennes, 1972).

Hennes (1972) concluded his review of the health status measures of this period by noting: “Most instruments have only superficial validation .... For the most part, the measurements have been on illness more than health, and often have been attempts to be proxy measures for a professional evaluation rather than attempts to measure concepts beyond the clinical level (p.1279)." Other researchers, apparently concerned with these and
other drawbacks of the existing "health status" measures, proposed guidelines and evaluation criteria for such measures (see, for e.g., Bush and Fanshel, 1970; Goldsmith, 1972; Moriyama, 1968; Sullivan, 1966).

However, the major drawback with all the measures developed during the early period was the fact that they were stand-alone types of measures -- they focused on the different dimensions of health status and no attempt was made to integrate them so as to provide a complete measure of health status. In other words, none of them had content validity as a measure of health status. One possible reason for this lack of integration could be the lack of appreciation of the merits of social science research by clinicians of the period. This is evident is Sullivan's (1966) criticism of the National Health Survey: "...the validity of disability data based on interview reports will be open to question until extensive use of such measures in a variety of studies has established their relation to clinical measures on the one hand and social variables on the other (p.15). ..... Reliable measurement requires elimination or control of extraneous factors influencing the measurement. ..... many aspects of survey procedure influence the measures obtained (p.14)." In spite of these concerns, social science research methods were embraced by health status researchers. This is because these methods offered guidance in solving not only the content validity problem identified above, but also other problems which faced health status measures of the early period such as assessing the reliability and validity of the measures. These methods formed the basis of the new health status measures which were developed in the modern period.
The Modern Period

The modern period began in the early 1970s and extends up to the present. It spans such a vast range because the instruments that were developed in the early 1970s are still being used and in fact represent the state of the art today. In contrast to the early period, this period was (and is) characterized by the application of sophisticated measurement techniques to multidimensional (as opposed to stand-alone) health status measurement instruments. Much of the initial developments in this period came about due to research funded by the National Center for Health Services Research to develop general measures of health status. Examples of such measures included the Quality of Well-Being Scale (see Bush, Chen and Patrick, 1973; Kaplan and Anderson, 1990), the Sickness Impact Profile (see Bergner 1988; Bergner, Bobbit, Carter, et al., 1981; Gilson, Gilson, Bergner, et al., 1975), and the General Health Rating Index, which was subsequently adapted by the RAND Corporation under Department of Health and Education and Welfare grants for use in the Health Insurance Experiment (see Kisch and Torrens, 1974; Ware, Brook, Davies-Avery, et al., 1980).

The modern period can be said to have begun with the work of Bush and his colleagues -- first at New York University and then at the University of California at San Diego -- who developed the General Health Policy Model for program evaluation. These researchers defined health as a composite of an individual's level of function at a point in time (i.e., function level or Quality of Well-Being) and his or her expected transition to other levels at future times (i.e., prognosis). More specifically, they defined health
as the product (expected value) of the social preferences assigned to levels of function and the probabilities of transition among the levels over the life expectancy of an individual or group (see Patrick, Bush, and Chen, 1973a; Kaplan and Anderson, 1990). Function level was defined as consisting of three dimensions which were intended to be mutually exclusive and collectively exhaustive -- social activity, mobility, and physical activity.

These dimensions were decided on the basis of a review of the literature and existing health surveys and were operationalized using decision logic tables (McDaniel, 1968) to ensure the ordinality of each component (Patrick, Bush and Chen, 1973a). Their instrument had three scales (one for each component) representing the objective disturbances that diseases and disabilities cause in role performance. The social activity scale initially had five steps (or levels), the mobility scale had five, and the physical activity scale had four (Patrick, Bush and Chen, 1973a). Taken together, these defined 100 (i.e., 5x5x4) different function states. However, some function states were impossible or illogical and were omitted, leaving a total of 29 logical function states in the initial instrument (Patrick, Bush and Chen, 1973a). The most recent version of the instrument has five levels for the social activity scale, three levels each for the mobility and physical activity scales (Kaplan and Anderson, 1990). The instrument also contained an independent set of symptom/problem complexes to denote subjective and symptomatic disturbances. Thus, the symptom/problem complexes were a measure of physiological status. Initially, the instrument had 42 different symptom/problem complexes (Patrick, Bush and Chen, 1973a). However, the latest version of the instrument has 27 different
symptom/problem complexes (Kaplan and Anderson, 1990). These changes reflect the result of ongoing studies on the instrument. It is notable that they did not consider mental function to be a basic dimension of health status, but believed "... the impact of mental health on general health status (was) expressed through its impact on life expectancy, functioning, and symptoms" (Kaplan and Anderson, 1990 p.136).

On the basis of these dimensions, each member of a given population can be classified into a particular function level (as some combination of the dimensions). However, to be used as an index of health status, the above classification needs to incorporate measures of preference of being in different function levels. Recognizing this, the relative preference for each function level was judged by equal numbers of graduate students and nurses (both groups had 31 respondents) using category scaling (Patrick, Bush and Chen, 1973a). (The details about different scaling methods will be discussed in Chapter III). Although the nurses' evaluations were statistically significantly higher than the students' (p<0.02), the difference was not considered practically significant (0.02 units on a 0-1 scale). Hence, the data were combined for computing the mean preferences for the function levels.

In later studies, the relative preferences of different function states was measured on different types of respondents using different scaling methods like magnitude estimation and equivalence rating (Patrick, Bush, and Chen, 1973b). No significant differences were detected. It was decided to use the results of category scaling for further calculations on account of the fact that this scaling method was easier for respondents, less sensitive than the other scaling methods studies to modifications in procedure
(Patrick, Bush, and Chen, 1973b), and could provide reliable and valid results if the response continuum was made clear to respondents (Kaplan and Ernst, 1983).

The measurement methodology used was rating scale conjoint methodology (discussed in Chapter III) based on these category ratings of multidimensional profiles of function level. Using these methods, a model of the preference structure was developed that assigned weights to each level (or step) of each dimension and symptom/problem complex, thus providing scores for all possible function levels with a high degree of accuracy -- $R^2 = 0.96$ (Kaplan and Bush, 1982). This model was cross-validated on a totally new set of function level profiles with an $R^2$ of 0.94 (Kaplan, Bush, and Berry, 1978).

Together, the three scales and symptom/problem complexes along with the social preferences for each function level define the Quality of Well-Being (QWB) Scale (previously referred to as the Index of Well-Being or the Function Level Index). On the basis of this scale, each function level can be placed on a scale of 0 to 1, where 0 is death and 1 is asymptomatic optimum function. Scoring is facilitated as follows: each respondent indicates his or her level of social activity, mobility, and physical activity on the scale and picks one of the 27 symptom/problem complexes which was the most undesirable over a given time period. As explained, each level of each dimension and symptom/problem complex has a weight or social preference associated with it. The QWB score is the simple sum of the weights associated with the selected levels. This allows the preferences for each function level to be measured on an interval scale.
The QWB Scale, however, is only one part of the General Health Policy Model (although it is the aspect relevant to this dissertation). The QWB Scale is only the point-in-time component of this model. The other component is prognosis, or the probability of moving between health states in time. Prognosis is operationalized as the expected duration of stay in a given function level over a standard life period as determined by an appropriate statistical (preferably stochastic) model. In this way, the effect of each program can be calculated in terms of well-life expectancy, which is the sum of the product of the QWB scores associated with particular function states and the expected duration of stay in each function state. In this way, the impact of any proposed program can be determined by measuring the difference the program makes in moving the population from one function state to another. Further, well-life expectancy can be used as the denominator in cost-utility analysis, thus providing an estimate of the cost incurred per year of well-life provided by the program(s) under evaluation.

The QWB and the General Health Policy Model represent a significant advance in health status measurement over the previous efforts of the 1960s. Not only did the QWB provide a measure of health status defined as a multidimensional construct, it was also the subject of extensive psychometric tests of validity and reliability (see, for e.g., Kaplan, Bush, and Berry, 1976; Kaplan, Bush, and Berry, 1978; Kaplan and Ernst, 1983; Patrick, Bush, and Chen, 1973a, 1973b). According to the definition of health forwarded by these researchers, the QWB has content validity. Tests of construct validity, including both convergent and divergent validity, were generally supportive of the QWB (Kaplan, Bush, and Berry, 1976). In the absence of a “gold
standard" of health status measurement, criterion validity of the QWB cannot be assessed. Moreover, by incorporating the effect of prognosis, the General Health Policy Model provides a meaningful estimate of the impact of different programs on health, thus fulfilling its intended purpose.

Limitations of the QWB include the length of time taken to administer the instrument and the consequent risk of respondent fatigue. The issue of picking a single most undesirable symptom/problem complex deserves further comment. It is possible that the respondent may have experienced more than one symptom/problem complex in the relevant time period, and that the interactive effect of these multiple symptom/problem complexes would be more undesirable than any one alone. Thus, by constraining the respondent to choose only one symptom/problem complex, it is possible that social preferences for different programs may be underestimated. It should also be noted that the QWB inherently assumes an additive model and does not allow for interactions between its dimensions. The importance of including interactions in models of health status is explained in detail in a subsequent section. However, it is necessary to point out here that the assumption of an additive model may bias the preference weights obtained when the underlying model is actually multiplicative or multilinear.

Further, the exclusion of mental function as a basic dimension of health status might preclude the use of the QWB in the evaluation of certain programs, e.g., programs directed at improving mental health such as psychological counseling programs. For evaluating such programs, it may be advisable to measure the direct effect of the program (i.e., on mental function) as opposed to observe the effect of the program on mental function.
via some other dimension of health status such as physical function. Finally, the QWB may not be responsive to minor but clinically significant changes in health status. This is especially true in cases of diseases or conditions which are asymptomatic or mildly symptomatic, such as hypertension (the condition of interest for this dissertation) and allergic rhinitis. Thus, the QWB may not be useful for individual or group clinical decision making in such diseases or conditions.

In spite of these limitations, the QWB and the General Health Policy Model have enjoyed widespread application. Published applications include total hip replacement, chronic obstructive pulmonary disease, arthritis, hypertension, phenylketonuria screening, pneumococcal vaccines, and even seat belt laws (see Kaplan and Anderson, 1990). A major public policy application of the QWB and the General Health Policy Model was for the Medicaid program for the state of Oregon (Kaplan, 1994; Strandberg, 1991). Using the General Health Policy Model, the state prioritized more than 700 treatments on the basis of their cost-utility ratios. The intent was to allocate resources to the treatments with the more favorable cost-utility ratios until the budget was exhausted. By diverting resources away from programs which had unfavorable cost-utility ratios, it was expected that Medicaid coverage could be extended to more people below the poverty level. This plan has received approval from Congress and the Health Care Financing Administration and is in the process of being implemented.

Since the pioneering research by Bush and his colleagues, several health status instruments have been developed. Some of these are intended
for program evaluation while others are intended for use in clinical decision making. Prominent among the instruments intended for program evaluation are the McMaster Health Utility Indices (Marks I, II, and III) and the EuroQol instruments. Among the health status instruments which are intended for clinical decision making, the most frequently used (at least in published studies) are the Sickness Impact Profile, the Nottingham Health Profile, the Short-Form 36 (commonly referred to as the SF-36), and the Dartmouth COOP Charts. While it is acknowledged that there are other instruments which measure health status, the following discussion shall limit itself to the instruments identified above. This is because these instruments represent significant advances in the development of health status measurement sophistication in the modern period.

The McMaster Health Utility Indices represent the collective effort of a group of researchers -- most notably George Torrance -- at McMaster University in Hamilton, Ontario. Much of the initial work by this group was methodological and conceptual in nature (see, for e.g., Torrance, 1973, 1976a, 1976b, 1976c; Torrance, Sackett, and Thomas, 1973; Torrance, Thomas, and Sackett, 1972). The group made its first attempt at developing a multiattribute health status classification system in the early 1980s, as part of a cost-utility analysis of a neonatal intensive care for very-low-birth-weight infants (see Boyle, Torrance, Sinclair, et al., 1982; Torrance, 1982a; Torrance, Boyle, and Horwood, 1982). The resulting classification, the Mark I Health Utilities Index, consisted of four dimensions of health status -- physical function (including mobility and physical activity), role function (including
self-care and role activity), social-emotional function (including emotional well-being and social activity), and health problems. The physical function dimension had six levels, the role function dimension had five levels, the social-emotional function dimension had four levels, and the health problems dimension had eight levels. Since then, this classification has undergone two major revisions on the basis of field studies. The most recent health status classification system (Mark III) has eight dimensions -- vision (six levels), hearing (six levels), speech (five levels), ambulation (six levels), dexterity (six levels), emotion (five levels), cognition (six levels), and pain (five levels).

The greatest contribution these researchers made to the health status measurement literature was methodological in nature. In 1972, Torrance developed the time tradeoff scaling method of measuring preferences for health states (Torrance, Thomas, and Sackett, 1972). This scaling method measures the value individuals place on a given health state with respect to another health state (details of the time tradeoff scaling method will be provided in Chapter III). The McMaster group of researchers were also instrumental in arguing for explicitly incorporating the notion of risk into the assessment of preferences for health states by using scaling methods such as the standard gamble to measure health state preferences. Depending on respondents' attitudes toward risk (risk-seeking, risk-neutral, or risk-averse), their preferences for different states of health will be different. Thus, preference assessment techniques which do not account for the inherent riskiness and uncertainty of health states may yield biased estimates of preference assessments.
These researchers were also the first to apply the principles of multi-attribute utility theory to the measurement of health state preferences. Multi-attribute utility theory is concerned with expressing the utilities (or measures of worth) of multiple attributes or dimensions as a function of the utilities of each attribute or dimension taken individually. The theory specifies several possible models (additive, multiplicative, and multilinear) and the conditions (independence conditions to be met) under which each would be appropriate (see Keeney and Raiffa, 1976). Although the details of multi-attribute utility theory will be provided in Chapter III, it is relevant to note here that this measurement methodology is appropriate and relatively efficient compared to overall or holistic utility measurement techniques when health states are described by a multidimensional system.

Apart from these measurement differences, the approach of the McMaster group is similar to that of the San Diego group (i.e., Bush, Kaplan, and their colleagues). Once the preferences for each level of each level of each dimension are empirically determined, the preference for a given health state can be determined according to the model (i.e., additive, multiplicative, or multilinear) deemed appropriate. The flexibility in allowing non-additive models is noteworthy, which represents an improvement over the QWB approach where only an additive model is permitted. However, the use of these preference weights is similar to those of the QWB in that they are used to adjust life expectancy, so as to get adjusted life expectancy for a given health state, also called Quality Adjusted Life Years or QALYs. These QALYs can be used as the denominator in cost-utility analyses, thus expressing the impact of different programs in terms of
cost per QALYs. In this sense, QALYs are similar to well-years in the General Health Policy Model approach.

The European Quality of Life Instrument (EuroQol) has been recently developed (in the late 1980s) by an interdisciplinary group of researchers in five different countries in Europe -- the U.K., Sweden, Norway, Finland, and The Netherlands (The EuroQol Group, 1990; Rosser and Sintonen, 1993). The principal aim of this group of researchers was to “test the feasibility of jointly developing a standardized non-disease-specific instrument for describing and valuing health-related quality of life (The EuroQol Group, 1990 p.200).” The instrument was intended to complement other instruments measuring health status or health-related quality of life and facilitate the collection of a common data set for reference purposes by creating a common core of items that measured health status. Of particular interest was the ability to make cross-cultural comparisons of health state valuations. It was recognized from the outset that the pursuit of a common core data set would not be a comprehensive measure of health status, but would need to cover most of the dimensions of health status that were represented in published research and the ranges of severity within each dimension. In this way, it was believed that the common core would offer scope for application to many different populations, from fairly healthy people living in their own homes and performing their usual activities to severely ill patients in hospitals and nursing homes (The EuroQol Group, 1990).

Although the instrument is of recent origin it has already undergone a revision in the number of dimensions of health status it contains. The
original instrument contained six dimensions -- mobility, self-care, pain (three levels each), main activity, social relationships, and mood (two levels each). These dimensions were selected on the basis of a review of the existing health status measurement instruments like the Quality of Well-Being Scale (Kaplan and Anderson, 1990), the Sickness Impact Profile (Bergner, 1988), the Nottingham Health Profile (Hunt, McEwen, and McKenna, 1985), and the Rosser Index (Rosser and Watts, 1972). The revised instrument has only five dimensions, each with three levels -- mobility, self-care, usual activities, pain/discomfort, and anxiety/depression (Rosser and Sintonen, 1993). The most significant aspect about the revision was the omission of the social relationships dimension. The precise reason for this omission is unclear. All that is reported about the revision is that it was based on “experiences thus gained” (Rosser and Sintonen, 1993).

Since it is of such recent origin and because most of the data collected are proprietary, there is not much published research on the EuroQol. Results of three preliminary studies have, however been published (Brooks, Jendteg, Lindgren, et al., 1991; Essink-Bot, Bonsel, and van der Maas, 1990; The EuroQol Group, 1990). These studies -- which were conducted in Lund (Sweden), Bergen op Zoom (in The Netherlands), and Frome (in the UK) -- used the original version of the instrument (i.e., with six dimensions). Respondents were asked to rate their preferences for 16 health states on a finely calibrated visual analog scale ranging from 0 (worst imaginable health state) to 100 (best imaginable health state). The selection of the 16 health states from the universe of 216 (i.e., 3x3x3x2x2x2) health states was purposeful -- states were selected which represented a wide range of
severity and which the researchers believed (from their experience) were likely to occur fairly frequently in practice. Moreover, as a check on the ability of respondents to make consistent valuations of health states, one health state was repeated twice. In a departure from other commonly used preference assessment techniques for health states, data were collected through the mail.

The results of these preliminary studies were remarkably similar across the study sites, in terms of both the absolute values given to the health states as well as their relative positions within the samples. This indicates that there are no significant differences in preferences for health states between respondents of different European countries. Regarding the consistency of respondents' valuations of the same health state, mixed results were found. In the Bergen op Zoom study, the two scores were highly correlated but had small although non-significant differences (Essink-Bot, Bonsel, and van der Maas, 1990). However, in the Lund study, the two scores were significantly (p<0.01) different (Brooks, Jendteg, Lindgren, et al., 1991). Separate results for the Frome study have not been reported. These mixed results suggest the reliability of the instrument may be suspect. However, further analyses of the psychometric properties of the instrument need to be reported before making a conclusion on this matter.

The EuroQol instrument has been the subject of some criticism from other researchers (see, for e.g., Carr-Hill, 1992; Gafni and Birch, 1993). Much of the criticism centers around the need for further analysis on the instrument, which is reported to be in progress (Rosser and Sintonen, 1993). Although the usable response rates were rather low (between 15 and 30
percent), these need to be judged in terms of the novelty of the response task, i.e., valuing health states through the mail. Indeed, such low response rates are commonly observed in pharmaceutical administration research. The important issue is the existence of nonresponse bias, which was empirically investigated and found to be absent (Essink-Bot, Stouthard, and Bonsel, 1993). Other criticisms of the EuroQol are more serious and address the methodological basis of the instrument (Gafni and Birch, 1993). It should be noted that most of these methodological criticisms also apply to other instruments which measure QALYs or well-years. Further, it should be noted that although this instrument claims to measure health-related quality of life, according to the clarification of terminology provided earlier in this chapter, it is actually a measure of health status.

In defense of the EuroQol instrument, it should be noted that it was not intended to stand alone as a measure of health status. Moreover, its inherent simplicity in identifying some core dimensions of health status has intuitive appeal, especially if its purpose is to foster cross-cultural comparisons of health status measurements. Its contribution to the development of health status measures lies in its use of the mail survey as a method of data collection for preference assessment of health states and the identification of core dimensions of health status for which respondents in different countries seem to have comparable valuations. The EuroQol instrument was used to operationalize health status for the purpose of this study, and the rationale for its selection in this study will be discussed in Chapter V.
One of the earliest health status instruments to be developed for the purposes of individual and group clinical decision making was the Sickness Impact Profile (SIP). The SIP was developed in the mid-1970s by researchers at the University of Washington, most notably Marilyn Bergner. The final version of the SIP consists of 136 items grouped into 12 categories -- sleep and rest, eating, work, home management, recreation and pastimes, ambulation, mobility, body care and movement, social interaction, alertness behavior, emotional behavior, and communication. Three of the categories (ambulation, body care and movement, and mobility) can be aggregated into a dimension of physical dysfunction, while another four (social interaction, communication, alertness behavior, and emotional behavior) can be aggregated into a dimension of psychosocial dysfunction. The remaining five categories cannot be aggregated into a coherent and consistent dimension. The choice of the word "dysfunction" as opposed to function requires some clarification. The items in the SIP are phrased so as to indicate changes in behavior associated with carrying out one's daily life activities due to sickness. It is argued that although these changes in behavior may be instrumental in accelerating the return to good health, they are dysfunctional in the sense that they represent impaired or ineffective role performance (Bergner, 1988).

The SIP is based on a conceptual model of sickness behavior which is operationalized in terms of behavioral dysfunction items. These items were generated by means of open-ended statements describing sickness-related changes in behavior from patients, health-care professionals, and informal caregivers (Bergner, Bobbit, Kressel, et al., 1976). The items were scaled
using Thurstone's method of equal-appearing intervals by patients and health-care professionals so as to get a weight for each item on each category. Respondents indicate agreement with each item on a nominal scale. A score for the overall SIP is computed by summing the scale values of all statements checked by a respondent and dividing that sum by the grand total of all statement scale values. This ratio is then multiplied by 100 to convert it to a percentage. Similarly, scores can be computed for each category or the physical and psychosocial dysfunction dimensions. The higher the score on the SIP, the more dysfunctional the respondent.

Like other health status instruments of the modern period, the SIP offers a multidimensional view of health status. The psychometric properties of the SIP have been extensively tested and found to be adequate (see Bergner, Bobbit, Carter, et al., 1981; Bergner, Bobbit, Pollard, et al., 1976; Gilson, Bergner, Bobbit, et al., 1978; Pollard, Bobbit, Bergner, et al., 1976). The SIP has been widely used in monitoring progress of individual patients as well as determining the effectiveness of drug therapy (see Bergner, 1988). The SIP is the subject of cross-cultural research, and has been translated into several languages, including German, French, Dutch, Chicano-Spanish, Norwegian, Danish, and Swedish (de Bruin, de Witte, and Diederiks, 1992). Recently, a shorter version of the SIP -- consisting of 68 items and six dimensions -- has been developed and found to have acceptable psychometric properties (de Bruin, Diedericks, de Witte, et al., 1994). Further, according to the conceptual model of sickness behavior used in the derivation of its items, the SIP may have content validity. However, using the definition of health status provided earlier which will be used for this
dissertation, the SIP focuses overly on the medical model of health and does not include much in the way of positive items about health status. In this sense, it may not adequately represent the construct of health status, i.e., it may not have content validity as a measure of health status.

Another health status instrument developed in the mid-1970s for the purpose of individual and group clinical decision making was the Nottingham Health Profile or NHP (Hunt, McEwen, and McKenna, 1985). Developed at Nottingham University in the UK, the NHP is a two-part instrument designed “to measure perceived health problems and to assess the extent to which such problems affect the activities of everyday life (Hunt, 1984 p.165).” Part I consists of 38 items covering six dimensions of health status -- pain, physical mobility, sleep, emotional reactions, energy, and social isolation. Part II explores the effect of health problems on seven areas of daily life -- paid employment, jobs around the house, personal relationships, social life, sex life, hobbies, and holidays (Hunt, 1984; Hunt and McEwen, 1980). Hence, while the dimensions included in Part I may reflect the physical, mental, and social aspects of health, the areas included in Part II go beyond the scope of health-related quality of life and are more reflective of quality of life in general.

Each item of the six dimensions included in Part I was weighted for severity using the Thurstone method of paired comparisons (which will be explained in Chapter III) with a sample of 215 members of the general public (McKenna, Hunt, and McEwen, 1981). On both parts of the instrument, respondents are required to indicate whether each item applies to them or
not on a nominal scale. In Part I, positive answers are given the appropriate weighting in the computation of scores. A weighted score is obtained for each of the dimensions in Part I of the instrument. The maximum score on any dimension is 100. The higher the score, the greater the extent and severity of the problems. In this way, a respondent score for each dimension of Part I of the instrument can be obtained. However, thus far, there is no arrangement for an overall score on all dimensions. Items in Part II of the instrument are not weighted.

The instrument has been tested for validity in a variety of populations, including elderly persons of varying degrees of fitness (Hunt, McKenna, McEwen, et al., 1980), persons who consult and persons who do not consult with general practitioners (Hunt, McEwen, McKenna, et al., 1981), patients with peripheral vascular disease, (Hunt, McKenna, McEwen, et al., 1982), patients with rheumatoid arthritis and osteoarthritis (Hunt, McKenna, and McEwen, 1981), fracture victims (McKenna, Hunt, McEwen, et al., 1984), and patients undergoing minor surgery (Hunt, McEwen, McKenna, et al., 1984). Reliability has been assessed using the test-retest technique through a mail survey on different populations (Hunt, McKenna, and McEwen, 1981; Hunt, McKenna, McEwen, et al., 1982). The results of these psychometric tests generally support the validity and reliability of the NHP as a measure of health status.

Although not as major an advance in the development of health status measures as some of the other instruments reviewed in this section, the NHP is noteworthy for the fact that it was one of the earliest examples of using mail surveys as a data collection method for health status assessment.
Response rates obtained were encouraging, ranging from 68 to 93 percent depending on the respondents surveyed. This established the feasibility of using mail surveys as a data collection method for health status assessment. However, apart from the EuroQol instrument, data collection through the mail has not been frequently used by health status researchers. The NHP was also the first (and still only) health status measurement instrument to use Thurstone's method of paired comparisons as a scaling method.

Limitations of the NHP are very similar to those of the SIP discussed above. Like the SIP, the NHP is overly influenced by the medical model of health and investigates the negative aspects of health only. Therefore, it cannot be used to assess positive feelings of well-being, since zero scores do not necessarily indicate a total absence of distress. Further, individuals who score zero cannot be shown to improve over time, even though in actuality their state of health may be much better. Further, the items in Part I of the NHP represent rather severe problems. It has been explained that this was a deliberate choice of the developers of the instrument, who wanted to avoid including a large number of false positives i.e., individuals who feel temporarily uncomfortable (Hunt, 1984; Hunt, McEwen, and McKenna, 1985). However, the result of this focus on severe conditions is that some people who are in distress may not show scores on the NHP. This represents a significant drawback of the NHP, since it may be said that it does not adequately represent the entire spectrum of health states and therefore cannot be used as a measure of health status for individuals in more positive states of health.
A significant event in the history of development of health status measures was the RAND Health Insurance Experiment (HIE). The HIE was not primarily designed to measure health status -- the purpose of the HIE was to determine the effect of financing health care through alternative insurance plans on several dependent variables, one of which was health status. Health status measures in the HIE were developed by an interdisciplinary group of researchers at the RAND Corporation, most notably John Ware, Anita Stewart, and Ron Hays. Although the health status measures used in the HIE have been criticized as being crude (Jette, 1980), they are significant in the history of health status measures mainly because of their conceptual development and future applications in the Medical Outcomes Study (Tarlov, Ware, Greenfield, et al., 1989). Hence, they need to be reviewed.

Health status was conceptualized in the HIE as a multidimensional construct, consisting of four dimensions -- physical health, mental health, social health, and general health perceptions. In addition to these dimensions, several other health-related constructs were also measured, such as health habits i.e., smoking, alcohol consumption, exercise, etc. (Ware, Brook, Davies-Avery, 1980). Physical health was visualized as consisting of physical functioning and physiological status. Six categories of physical functioning were identified on the basis of a review of the existing literature -- self-care, mobility, physical activities, role activities, household activities, and leisure activities. Three separate batteries of questions were designed to measure limitations and abilities in each of the six categories of physical functioning: (a) the Functional Limitations battery focused on both chronic
and acute limitations in performance of a variety of normal activities considered basic to daily living; (b) the Physical Abilities battery focused on the current ability (or capacity) to perform both activities basic to daily living and others that a person might elect to do if he or she were in good physical health; and (c) the Disability Days battery measured acute changes in performance, focusing on day-to-day changes in a person's usual activities without further specifying the abilities required to perform those activities (Stewart, Ware, Brook, et al., 1978). Physiological status was defined as the presence or absence and severity of about 20 chronic diseases (Ware, Brook, Davies-Avery, et al., 1980).

Mental health was measured in the HIE using two batteries of questions that focused on psychological states rather than physiological and somatic states since the latter were expected to be represented in the measures of physical health. An effort was made to include both favorable and unfavorable aspects of these psychological states. Both batteries included scales hypothesized to measure six different categories mental health -- anxiety, depression, positive well-being, self-control, vitality, and general health (Ware, Johnston, Davies-Avery, et al., 1979).

Measurement of social health seemed to have received lower priority in the HIE (Donald, Ware, Brook, et al., 1978). As opposed to items measuring physical and mental health, items measuring social health were not included in the first Medical History Questionnaire (at Dayton). Further, only 11 items measuring social health were included in the questionnaire (after Dayton enrollment). In agreement with the prevailing literature of the time (Dohrenwend, Dohrenwend, and Cook, 1973; Myers, Lindenthal,
Pepper, et al., 1972; Myers, Lindenthal, and Pepper, 1975), these items defined social health in terms of participation and interpersonal interaction and focused on the number or frequency of several kinds of participation and interaction.

HIE measures of general health perceptions differed from other measures of health status in that they did not focus on any particular dimension of health status. Instead, these measures asked respondents for an assessment or self-rating of their health in general. In the HIE, these measures were defined with respect to time (i.e., perceptions of prior, current, and future health), resistance to illness, health worry and concern, and sickness orientation or the extent to which people believe illness to be a part of their lives (Davies-Avery and Ware, 1981; Ware, Brook, Davies-Avery, 1980; Ware, Davies-Avery, and Donald, 1978). Like measures of social health, these measures were not included in the first questionnaire but were added subsequently.

Items were scaled using two different types of scaling methods. Likert’s method of Summated Ratings was used to score one battery of physical health measures (Physical Abilities), the mental health scales, and the general health perceptions scales, while Guttman scaling was used to score the Functional Limitations battery and the social health scale. In this way, a score can be generated for each individual battery of questions. Some batteries (e.g., the mental health batteries) can be aggregated into a single index, while others cannot. The reliability and validity of each battery of questions were extensively tested and found to be satisfactory (see Brook, Ware, Davies-Avery, et al., 1978; Davies-Avery and Ware, 1981; Donald,
Ware, Brook, et al., 1978; Rogers, Williams, and Brook, 1979; Stewart, Ware, Brook, et al., 1978; Ware, Brook, Davies-Avery, et al., 1980; Ware Davies-Avery, and Brook, 1980; Ware, Davies-Avery, and Donald, 1978; Ware, Johnston, Davies-Avery, et al., 1979).

The importance of the HIE health status measures in the history of development of health status measurement instruments lies in the process of their development and construction. Regarding their development, they were one of the first to adopt a truly expanded view of health in that they focused on the positive as well as negative aspects of health. Although the QWB also did so, the items in the HIE health status measures had more of a balance in positive and negative aspects of health than those in the QWB. Moreover, the developers of these measures reviewed a wide variety of interdisciplinary research while constructing the scales to measure health status. Finally, efforts were made to correct for potential sources of bias in responses, such as acquiescent response set and social desirability response set. Thus, the HIE health status measures offered conceptual as well as methodological advances over existing health status measures.

The HIE also deserves some credit for the development of disease-specific health status instruments, which measure the health status of individuals suffering from a particular disease or condition. The development of these instruments has been greatly facilitated by the series of booklets produced by the RAND Corporation describing the conceptualization and measurement of physiological status. Each booklet described the problems in conceptualization and measurement of a specific disease or condition as well as the particular aspects of these diseases or
conditions that need to be considered in measuring the health status of individuals suffering from these diseases or conditions. The detailed account provided in these booklets provided a conceptually and methodologically sound basis for developing disease-specific health status measures.

The major limitations of the HIE health status measures are due to the fact that they were designed for the purpose of evaluating the effect of financing health care through alternative insurance plans on health status and not to measure health status per se. On account of this, they were designed to be responsive to only those changes in health status which may be expected due to the effects of financing health care through alternative insurance plans. This is evident in the developers' comments explaining the omission of certain aspects of physical, mental, and social health (see Donald, Ware, Brook, et al., 1978; Stewart, Ware, Brook, et al., 1978; Ware, Johnston, Davies-Avery, et al., 1979). It is because of these omissions that some critics have termed these measures "crude" (Jette, 1980). In spite of these limitations, the advances the HIE health status measures offered were significant. Indeed, some of the measures were subsequently used unchanged or slightly modified in another major study, the Medical Outcomes Study (Tarlov, Ware, Greenfield, et al., 1989).

The Medical Outcomes Study (MOS) was a study conducted in the mid-1980s involving more than 500 physicians and 22,000 patients in three major US cities. The purpose of this study was to assess the impact of the structure and process of delivery of medical care on the outcomes of medical care, one of which was health status (Tarlov, Ware, Greenfield, et al., 1989). The principal researchers who developed the health status measures in the
MOS were the same as those in the RAND HIE. The instrument used to measure health status in the MOS has since been modified, and will therefore not be discussed in detail. The discussion will focus on its successor, the Short-Form 36 (SF-36).

The conceptual development of the SF-36 is similar to that of the HIE health status measures, and will therefore not be repeated. The SF-36 was designed for use in clinical practice and research, health policy evaluations, and general population surveys (Ware and Sherbourne, 1992). The instrument that has come to be known as the SF-36 has undergone three revisions, the third producing the current version of the SF-36. These revisions were made according to field experiences. The first version consisted of 18 items measuring four dimensions of health status -- physical functioning, role functioning, mental health, and general health perceptions. The second version, which was the one used in the MOS, had 20 items measuring six dimensions of health status -- physical functioning, social functioning, role functioning, mental health, pain, and general health perceptions. Finally, the most current version (i.e., the SF-36) has 36 items and measures eight dimensions of health status -- physical functioning, social functioning, role limitations due to physical problems, general mental health (including psychological distress and well-being), role limitations due to emotional problems, vitality (including energy and fatigue), bodily pain, and general health perceptions (Ware and Sherbourne, 1992). Apart from changes in the number of dimensions, the different versions also have more items per dimension of health status than the previous versions.
The SF-36 is one of the most widely used health status measurement instruments today, and has been extensively tested for its psychometric properties. Tests on the validity and reliability of the SF-36 in a variety of patient and non-patient populations have supported the conclusion that it is valid and reliable as a measure of health status (McHorney, Ware, and Raczek, 1993; McHorney, Ware, Lu, et al., 1994). Each dimension is represented by a multi-item summated scale. Hence the scaling method used is Likert's method of Summated Ratings for each dimension. Separate scores are obtained for each dimension. For each individual dimension, raw scores are transformed into a 0-100 scale. At present, aggregating scores across dimensions to obtain a single health status is not recommended by the developers (Ware, Snow, Kosinski, et al., 1993), although it is permissible to obtain two “mega-dimensional” scores of physical health and mental health (Ware, Kosinski, Bayliss, et al., 1995).

This has been a very brief discussion of the SF-36. The reason for this short discussion is that most of the advancements in health status measurement methodology represented by the SF-36 are actually the products of the HIE (whose principal developers were the same as those of the SF-36). The major improvement of the SF-36 over and above these advancements was more practical in nature in that content validity of the health status measure was not sacrificed while yet dramatically reducing the length of the instrument and thus easing respondent burden. This would allow it to be routinely used in clinical practice, which is one of the reasons for its popularity. However, since the scores obtained cannot be aggregated into an overall score, the SF-36 cannot operationalize tradeoffs between the
different dimensions of health status, which would be very useful in individual and group clinical decision making. Further, this may even hinder its use in health policy evaluations, which typically require measuring the impact of a given proposal on the overall health (i.e., an overall score on all dimensions of health status) of a given population. It should, however, be noted that the issue developing an aggregate score for the SF-36 is currently under investigation at the University of Sheffield in the UK (Brazier, 1995).

One final health status instrument that deserves mention in this section is the set of instruments known as the COOP Function Charts, developed by the Dartmouth Primary Care Cooperative Information Project or COOP Project in the mid-1980s (Nelson, Landgraf, Hays, et al., 1990; Nelson, Wasson, Kirk, et al., 1987). Although there are several researchers who are part of this project, the principal investigator in the published articles on the COOP Function Charts is Eugene Nelson. The COOP Function Charts are a series of charts which measure functional status. There are currently nine different charts which comprise the COOP Function Charts, each of which measures a different dimension of health status. The dimensions measured are physical function, role function, social function, pain, emotional status, social support, perceptions of quality of life, perceptions of overall health, and perceptions of health change over the past four weeks. Each chart is designed as a five-point Likert-type scale.

These charts are designed to measure functional status in routine clinical practice. The improvement these charts offer over existing health
status measures for this purpose is their design -- pictures are used as part of each scale to assist patients in quickly grasping the concept to which each chart relates. This is particularly useful for patients whose first language is not English, who are marginally literate, or who rarely think in abstract constructs. However, pictures may introduce unwanted specificity by causing respondents to focus on a particular aspect of the dimension in question i.e., the one depicted (Palmer, 1987). In spite of this concern, the COOP Function Charts have been positively received by both patients and clinicians (Nelson, Landgraf, Hays, et al., 1990; Nelson, Wasson, Kirk, et al., 1987). Further, they have been found to be reliable and valid measures of health status in different patient populations (Nelson, Landgraf, Hays, et al., 1990; Nelson, Wasson, Kirk, et al., 1987).

In summary, the field of health status measurement has shown significant advances over the past few decades, especially since the 1970s. Although the initial health status measures were overly influenced by the medical model of health, the newer ones adopt a more inclusive view of health -- more along the lines of the WHO definition of health (World Health Organization, 1947). Further, the field of health status assessment has shown increasing measurement sophistication over its course of development. Most instruments that are in use today have been extensively tested for their psychometric properties in different populations and have been found to be reliable and valid measures of health status in these diverse populations. It is largely due to these advances in measurement sophistication that the interest in health status measurement has become a reality.
Health Status Measurement -- State of the Art

At present, there are several instruments available that measure health status. It should be noted that some of these measures purport to measure health-related quality of life, but according to the clarification provided earlier and the definitions adopted for this dissertation, these instruments measure health status. It must also be remembered that these different health status measurement instruments are not strictly substitutable, although they purport to measure the same construct.

Read, Quinn, and Hoefer (1987) evaluated the General Health Rating Index or GHRI (one of the RAND HIE health status measures), the Quality of Well-Being Scale or QWB, and the Sickness Impact Profile or SIP in terms of their practicality and validity. Practicality was assessed in terms of interviewer training required, administration time, and respondent burden. In terms of validity, content, convergent, and divergent, validity were assessed. It was found that each instrument, although different in application, performed according to the claims of its developers and could provide valid, useful data on health status. It was concluded that each was specifically suited for a particular purpose -- the GHRI for the collection of data through brief, self-administered forms for routine clinical practice, the QWB for cost-utility analyses, and the SIP for dealing with a wide range of specific dysfunctions (Read, Quinn, and Hoefer, 1987).

Hornberger, Redelmeier, and Peterson (1992) have shown how different results can be obtained in a cost-effectiveness analysis using different measures of health status. Kirshner and Guyatt's (1985) advice on evaluating health status measures for use offers guidance on how to deal
with this issue -- measures should be evaluated on the basis of their intended purpose. Apart from the basis of their intended purpose, MacKeigan and Pathak (1992) have noted that health status instruments can be classified in several different ways, including scope, level of aggregation, and scaling focus.

Scope

In addition to instruments which measure general health status, several instruments have been developed which measure health status as affected by a particular disease or condition. As noted previously, a major influence in the development of such instruments was the series of booklets produced by the RAND Corporation on the conceptualization and measurement of physiological status in different diseases and conditions. These instruments are referred to as disease-specific health status instruments. Such instruments have been developed for a wide variety of diseases and conditions (e.g., hypertension, diabetes, arthritis, cancer, asthma, angina, AIDS, etc.), and are routinely being used in evaluations of clinical programs and new drugs on these diseases and conditions (MacKeigan and Pathak, 1992). It is expected that these instruments will be more capable of detecting subtle yet clinically significant changes in health status resulting from clinical programs or drug therapy for these diseases or conditions, since these instruments only include those elements which are more important to that disease. These instruments, however, are of little use in making comparisons between clinical programs or drug therapies affecting different diseases or conditions.
A more recent innovation is the development of "multiple-disease specific" instruments, which focus on health status measurement in patients who have specific multiple diseases and/or conditions. An example of such an instrument is the one developed by Testa and Simonson (1988) for use in hypertensive patients who are also diabetic. Since this dissertation will focus only on the general dimensions of health status, disease-specific health status instruments will not be reviewed. However, in establishing the applicability of the general health status measure to hypertensive patients, the salient dimensions of most health status measurement instruments for hypertension will be reviewed in Chapter V.

Level of Aggregation

Health status measurement instruments may also be classified according to the level at which they aggregate scores on the different dimensions measured by them. Within the field of health status measurement, there are two different schools of thought on the issue of aggregation of scores on the individual dimensions of health status (The EuroQol Group). Both schools agree that health status is a multidimensional construct, but they differ in the implications of this multidimensionality. The first school of thought holds that this multidimensionality implies that the measurement of health status must necessarily be multidimensional, in which one can only measure health status within a particular dimension. The second school of thought holds that since people have to weigh the diverse dimensions (implicitly or explicitly) and arrive at a decision about which, on balance, is best and how these diverse dimensions fit into the
overall health status evaluation, health status measurement should be conducted in such a way that provides an overall or aggregate score of health status. In this regard, health status measurement instruments can be profiles, batteries, or indices. Profiles and batteries are the products of the first school of thought, while indices are the products of the second school of thought.

A profile is an instrument that provides separate scores for each of the dimensions of health status that it measures and makes no attempt to aggregate these scores into a single, overall score of health status. Examples of health status measurement profiles are the Nottingham Health Profile and the SF-36.

A battery is a collection of independent instruments or individual dimensions from different instruments that have been assembled to obtain a comprehensive assessment of health status. Scores are reported separately for each instrument or dimension. An example of a health status measurement battery is the one used by Croog, Levine, Testa, et al. (1986) in their measurement of the “quality of life” of hypertensive patients treated with captopril, methyldopa, or propranolol. These researchers used a health status battery composed of ten unidimensional instruments measuring physical symptoms, sleep dysfunction, sexual function, work performance and satisfaction, cognitive function (two scales), social participation, emotional status, general well-being, and life satisfaction (Croog, Levine, Testa, et al., 1986).

Profiles and batteries provide detailed information about the effect of clinical programs and drugs on different health status dimensions. However,
their major limitation is that they cannot compute an overall score of health status as a composite of all its dimensions. Therefore, while these instruments may be useful for individual and group clinical decision making they may be of little use in program evaluation, where the emphasis is on the effect of the program(s) under evaluation on the overall health of the population. Health status indices are useful in this latter application.

An index is a single health status instrument that aggregates the scores obtained on its different dimensions into one single score. A critical issue in the development of health status indices is the basis for aggregating the scores of the different dimensions, i.e., the weights assigned to each dimension. Although the issue of determination of weights for different dimensions of health status instruments is the subject of a subsequent section, it is relevant to note here that a simple summing up of the different scores on each dimension is not a satisfactory method since it implicitly assigns equal weights to each dimension. Since one dimension may be more important than another, it is necessary that the score for each dimension be weighted by its relative importance when an overall score is calculated (MacKeigan and Pathak, 1992). The issue of relative importance of different dimensions of health status is of great significance to this dissertation, and will be discussed in detail in Chapter II. Examples of health status indices include the Quality of Well-Being Scale and the McMaster Health Utility Indices.

It is important to note that there are some health status measurement instruments that do not fit exclusively into one of these classes (i.e., profiles or indices). For example, the Sickness Impact Profile allows the user report
scores in different ways, a separate score for each dimension, a partial aggregation of some dimensions into an aggregate score of physical dysfunction and psychosocial dysfunction, or an overall aggregate score of all 12 dimensions.

**Scaling Focus**

The level of aggregation of a health status measurement instrument is usually a function of the scaling focus adopted in its development, which in turn is dictated by the purpose for which the instrument was developed. In this regard, there are two general purposes for developing health status measurement instruments -- to scale people or to scale stimuli. In scaling people, the objective is to compare people by identifying their location along a continuum. On the other hand, in scaling stimuli, the objective is to compare stimuli (e.g., health states) along some continuum. Accordingly, there are two different types of scaling foci used in health status measures -- the person-scaling focus and the stimulus-scaling focus. Within each focus, there are several scaling methods which can be used for generating the scores. Both foci involve assigning numbers to each level of each dimension of health status measured by the instrument. The difference lies in the meaning of the numbers and how they are generated. This difference is important in terms of the potential uses of the scores obtained.

In the person-scaling focus, the numbers assigned to the levels are usually based on their relative frequency of occurrence or relative severity in the population. No attempt is made to provide a measure of the relative worth or preferences associated with the levels of each dimension. In other
words, although the numbers express increasing (or decreasing) frequency of occurrence and severity of the level in question, they do not contain any information about how much being in or belonging to a specific level of a dimension distresses or is appreciated by an individual. Scaling methods used in the person-scaling focus include summated (e.g., Likert's method of Summated Ratings) and cumulative scaling (e.g., Guttman scaling). Examples of health status measurement instruments which have been scaled using the person-scaling focus include the RAND HIE health status instruments and the SF-36. The person-scaling focus has mainly been applied in the construction of health status profiles and batteries (the Nottingham Health Profile is a notable exception).

On the other hand, in the stimulus-scaling focus, the numbers assigned to the levels represent the relative worth or preferences associated with the level in question. Scaling methods used in this focus are of two basic types -- psychophysical and decision-theoretic methods. Included within the psychophysical methods are category scaling, magnitude estimation, and paired comparison methods. Decision-theoretic methods include the time tradeoff and standard gamble methods. Each of these different methods will be discussed extensively in Chapter III. Examples of health status measurement instruments which have been scaled using the stimulus-scaling focus include the Quality of Well-Being Scale, the McMaster Health Utility Indices (Marks I, II, and III), and the EuroQol instrument. The stimulus-scaling focus has mainly been applied in the construction of health status indices.
It should be noted that health status measurement instruments developed using the person-scaling focus implicitly assume that each dimension of health status has equal importance. This is not so for instruments developed using the stimulus-scaling focus, which explicitly weight the different dimensions of health status and combine them into an overall index of health status. This issue of weighting and the relative importance of health status dimensions is important to this dissertation, and is discussed in greater detail in Chapter II.

**Importance of Health Status Measurement**

Given the historical perspective and current state of health status measurement, one can appreciate its importance. The three major uses of health status measurement are public policy research and decision making, clinical research, and routine clinical decision making.

**Public Policy Research and Decision Making**

At the public policy level, health status measurement is important in program evaluation and resource allocation decisions. The use of these measures in large population studies has a long history. Beginning with the National Health Interview Survey, functional status assessments were included in establishing national health norms. The RAND Health Insurance Experiment and the Medical Outcomes Study examined the impact of alternative financing and systems of care on the patient's health (Brook, Ware, Rogers, et al., 1983; Tarlov, Ware, Greenfield, et al., 1989). The Quality
of Well-Being Scale and the McMaster Health Utility Indices are particularly attractive for use in public policy research and decision making on account of the fact that they are amenable for incorporation into cost-utility analyses. Both these instruments have been extensively used for this purpose. The use of the QWB and the General Health Policy Model in the Oregon Medicaid program (see Kaplan, 1994; Strandberg, 1991) has already been discussed. The latest version of the McMaster Health Utility Index (Mark III) has been used in the Ontario Health Survey, a population-based health survey of over 50,000 residents on Ontario, Canada and in the Statistics Canada General Social Survey, a Canadian population survey of over 11,000 residents. The Mark III Version of the McMaster Health Utilities Index is also currently being used in the National Population Health Survey and National Longitudinal Health Survey of Children in Canada (Feeny, Furlong, Boyle, et al., 1995).

Kaplan and Bush (1982) have suggested guidelines for the interpretation of cost-utility ratios. Although these guidelines were stated in terms of US dollars, the year of the currency was not specified (i.e., 1980 or 1981 etc. US dollars). According to these guidelines, if the cost per QALY or well-year is less than $20,000, the alternative under evaluation is cost-effective by current standards. If the cost per QALY or well-year of the alternative under evaluation is between $20,000 and $100,000, it is possibly controversial, but may be justifiable. However, if the cost per QALY or well-year is over $100,000 of the alternative under evaluation, it is questionable in comparison with other health-care expenditures (Kaplan and Bush, 1982). These guidelines need to be interpreted with extreme caution. The
guidelines were forwarded in 1982, and need to be adjusted for the time value of money in their application today. Moreover, there are several issues to consider in the assessment of alternatives via "league tables," i.e., ranking different alternatives in terms of their cost per QALY or well-year (see, for e.g., Drummond, 1987). For instance, the studies used to construct these league tables may have been conducted at different times and places using different methodologies. Hence, it is necessary to evaluate the external validity of each of the studies in a given league table before using the studies for decision making. Further, cost-utility ratios do not address the issue of equity, and therefore should not be the sole criterion used for public policy decision making. In spite of these concerns, however, the importance of incorporating health status measurement in public policy research and decision making remains unchanged, because it is necessary to evaluate the effects of proposed alternatives on the health of the population.

Clinical Research

Health status measurement is increasingly being used in clinical research today in both, pre-marketing clinical trials as well as post-marketing studies of drugs in a variety of settings. The recommendation of the FDA Oncologic Drugs Advisory Committee that health-related quality of life data be considered along with survival data in the approval decision of anticancer drugs (Johnson and Temple, 1985) has prompted pharmaceutical manufacturers to sponsor and conduct pre-marketing clinical trials for these and other drugs. As noted earlier, most pharmaceutical manufacturers have established entire departments to be responsible for such studies (Revicki,
Rothman, and Luce, 1992). The importance of including health status measures in clinical trials is illustrated by the following two studies, which were among the earliest studies measuring the health status of patients in clinical trials to be published.

Croog, Levine, Testa, et al. (1986) conducted a multicenter, randomized, double-blind clinical trial of 626 men with mild to moderate essential hypertension to demonstrate the effects of captopril, methyldopa, and propranolol on blood pressure control and health status. Health status was measured using a battery of ten unidimensional scales -- general well-being, physical symptoms, sleep dysfunction, sexual function, work performance and satisfaction, emotional status, cognitive function (2 scales), social participation, and life satisfaction. It was found that all three drugs were equivalent in terms of blood pressure control, but patients taking captopril had better health status scores than patients taking either methyldopa or propranolol. More specifically, patients taking captopril scored significantly higher (p<0.05 to p<0.01) on measures of general well-being, had fewer side effects, and had better scores for work performance, cognitive function, and life satisfaction than patients taking methyldopa. Compared to patients taking propranolol, patients taking captopril reported fewer side effects, less sexual dysfunction, and a greater improvement in measures of general well-being (p<0.05 to p<0.01). This study demonstrated the need to look beyond traditional clinical measures in clinical trials, as all three drugs were equivalent in terms of blood pressure control but showed marked differences in their effect on health status in patients (Croog, Levine, Testa, et al., 1986).
Bombardier, Ware, Russell, et al. (1986) conducted a multicenter, randomized, double-blind study of 303 patients with classic or definite rheumatoid arthritis to determine the difference between auranofin and placebo in terms of four outcomes composites -- traditional clinical measures (e.g., number of tender joints), measures of functional performance, pain, and global assessments of health status. Auranofin was found to be significantly better than placebo in the traditional clinical (p=0.003), functional performance (p=0.001), and global assessment (p=0.007) composites. Both trended similarly in the pain composite. Individual measures within the composites also favored auranofin. This study demonstrated that nontraditional outcomes measures, when included in clinical trials along with the traditional clinical measures, provide a more meaningful assessment of the alternatives under consideration in a clinical trial (Bombardier, Ware, Russell, et al., 1986).

Routine Clinical Decision Making

The third major use of health status measures is in routine clinical decision making. Quantifiable assessments of health status have been shown (see Kazis, Anderson, and Meenan, 1989) to be at least as reliable and valid as many standard clinical parameters (e.g., blood pressure) and laboratory tests (e.g., total cholesterol). Many health status measurement instruments (e.g., the SF-36 and the COOP Function Charts) are easy to administer and score. These instruments can be used to routinely assess patients' health status in clinical practice. Profiles and batteries may be more useful for this purpose than indices, since they are able to obtain more detailed information
about specific health status dimensions as compared to indices. The results could be displayed as a quantitative score on each dimension and compared with a reference score (e.g., a population norm) expressed as a mean and its standard deviation. The scores can be examined by the clinician on each visit and changes in scores tracked over successive visits. Specific flags and norms can also be established to signal significant variations (Kazis, 1991).

Although this may sound overly idealistic, recent studies have supported the clinical applicability of this approach. Kazis, Anderson, and Meenan (1988) have noted that health status summary reports could be evaluated by clinicians in a reproducible manner. Rubenstein, McCoy, Cope, et al. (1989) found that giving clinicians feedback about their patients' health status in the form of computerized summary reports was associated with a significant improvement in patients' psychological status and social functioning. The issue of developing population norms is currently being investigated for the SF-36 (Ware, Snow, Kosinski, et al., 1993).

Another approach would involve providing such information to patients, in addition to their clinicians (Kazis, 1991). This might allow patients to understand their current health status better in terms of how they compare to not only their previous health status, but also a reference group. This may lead to better patient outcomes via the "sense of control" feeling identified earlier. It should, however, be noted that the benefits of including health status measures in routine clinical decision making have yet to be realized to the extent that the benefits of including these measures in the other areas (i.e., public policy research and decision making and clinical research) have been realized. Some of the possible reasons for this are
addressed in the next section, which identifies the outstanding or unresolved issues in health status measurement which need to be addressed before the full benefits of these measures can be realized.

**Outstanding Issues in Health Status Measurement**

In spite of the progress made in the measurement of health status, there are yet many issues which need to be resolved. These outstanding or unresolved issues can be broadly classified into three groups -- conceptual, methodological, and application-oriented. Specific points to consider under each of these issues are introduced in this section. Some of these points to consider are discussed in greater detail in this section, while discussion of other points to consider is deferred until the subsequent chapters of this dissertation. The decision about which points to consider should be discussed in greater detail in this section versus subsequent chapters was based on considerations of the logical flow of issues -- for instance, discussion of points to consider which were relevant to the nature of the problem investigated in this dissertation were deferred until Chapter II, discussion of points to consider which were instrumental in making decisions about the methodology adopted in this study were deferred until Chapter V, while discussion of points to consider in the application of health status measures to their potential uses are discussed in this section. It should be noted that the points to consider introduced in this section are by no means exhaustive -- rather they reflect the ones which are mentioned most frequently in the published literature and have relevance to this dissertation.
Conceptual Issues

One of the major unresolved conceptual issues in the measurement of health status is the identification of the dimensions of health status. As noted earlier, there is no agreement on what are the specific dimensions of health status among health status researchers. For instance, the Quality of Well Being (QWB) Scale has only four dimensions -- mobility, physical activity, social activity, and a symptoms/problems complex (Kaplan and Anderson, 1990), the Sickness Impact Profile (SIP) has 12 dimensions -- sleep and rest, eating, work, home management, recreation and pastimes, ambulation, mobility, body care and movement, social interaction, alertness behavior, emotional behavior, and communication (Bergner, 1988), while the revised version of the EuroQol instrument has five dimensions -- mobility, self-care, usual activities, pain/discomfort, and anxiety/depression (Rosser and Sintonen, 1993). In the face of such diversity, Kirshner and Guyatt’s (1985) advice on letting the purpose of measurement dictate the content validity of the health status measurement instrument seems appropriate. The particular health status dimensions investigated in this study will be discussed in Chapter V. Apart from the identification of the dimensions themselves, there is the question of whether the individual dimensions should be viewed as orthogonal to each other or do they covary in predictable ways? On a related note, the issue of prognosis or expected outlook in terms of quantity and quality of life at a given point in time has also not been settled. The issue of prognosis will be revisited in Chapter II, when the effect of time on health status preference measurement is discussed.
Another conceptual issue, which is particularly important to this dissertation, is the relative importance of dimensions of health status. The relevant question in this regard is whether each dimension should be viewed as contributing equally to health status or should one assign weights to each dimension, representing their relative importance? Moreover, how does the possibility of interactions among health status dimensions affect their relative importance? These issues are discussed further in Chapter II. Further, if such weights are desirable, how should they be generated? This particular issue is the subject of Chapter III. A related issue is the extent to which these weights differ in different population groups (including geographic, age, gender, and therapeutic groups) and even within an individual as a function of time and life stages. This dissertation will focus on only the identification of the relevant weights and their methods of generation in one particular population group -- therefore this last issue will not be discussed further.

Methodological Issues

Methodological issues include both, issues involved in the design as well as implementation of health status measurement instruments. One of the more salient issues relevant to the design stage is the expression of the different levels of each dimension (i.e., in terms of functional capacity or performance). In this regard, individual items in health status measurement instruments can be expressed either in terms of functional capacity or performance. When phrased in terms of functional capacity, individual items stress on the ability of the respondent to engage in the specific function
in question. On the other hand, when phrased in terms of performance, individual items stress on the *extent* to which the specific function in question was actually carried out by the respondent. Since the two forms of expression stress different aspects of functioning, it is possible that they could lead to different results for the same respondent's assessment of his or her functioning at the same point in time, e.g., respondents may have the ability to do several different activities but may choose for a variety of reasons to do actually perform these different activities. In other words, by not carefully distinguishing between functional capacity versus performance expressions of individual items in health status measurement instruments, one runs the risk of confounding inherent ability to function with motivation to function.

Given this, the question that arises is which particular form of expression should be used in health status measurement instruments, or should both be used? This is an area which has not subjected to much research -- either conceptual or empirical -- in the published health status literature [although it should be noted that an unpublished report on this issue was referenced by Kaplan (1982)]. The few health status researchers who have recognized the distinction between these two forms of expression have only stated a preference for one instead of justifying the choice of one over the other. A notable exception is the work of the McMaster group of researchers, who, while explaining their choice of expressing levels of function for the Mark II version of the Health Utilities Index in terms of functional capacity as opposed to performance noted that their intent was "... to document the extent to which deficits in health status for each attribute
inhibit or prohibit normal functioning rather than to report the level at which an individual chooses to function, as would be reflected in a measure of performance" (Feeny, Furlong, Barr, et al., 1992 p.924).

At present, the different health status measurement instruments use both types of expression forms. For instance, the Mark II and Mark III versions of the McMaster Health Utilities Indices phrase the individual items in terms of functional capacity, while the Quality of Well Being Scale phrases the individual items in terms of performance. Given the potential to obtain different results (i.e., responses to questions) by both phrases of expression, it is necessary to empirically determine: (a) whether this potential to obtain different results is actually realized in practice and (b) if there is practical significance to the functional capacity versus performance distinction, which would the preferred form of expression be for health status measurement and further, does the preferred form of expression depend on the particular purpose of measurement of health status.

As opposed to the case of the functional capacity versus performance distinction, the issue of labeling diseases or conditions has been subjected to some empirical testing in the published health status literature. Indeed, the behavioral consequences of the effects of labeling patients as hypertensives has been recognized in the clinical literature for over a decade (for e.g., see Alderman, Charlson, and Mercher, 1981; Taylor, Haynes, and Sackett, 1981). Bulpitt and Fletcher (1988) have suggested that the act of labeling a patient as having a particular disease or condition (which is essentially the same as diagnosing a patient with a particular disease or condition) could lead to decreased psychological well-being.
The issue of labeling has special implications if preferences for health status need to be measured in any study. This is because several applications of health status preference measures require the general population to evaluate different states of health (as in the Oregon Medicaid project described earlier in this chapter). Therefore, it is possible that describing a particular health state as resulting from that of a patient with one disease state may result in a different evaluation of the same health state described as resulting from another disease state, e.g., the "for-a-person-with-disease X-that-is-good" effect.

Another important issue which needs to be considered in the design stage is the aggregation of individual preferences, if health status preferences are to be measured. This issue, of course, is only a factor for the cases of public policy and clinical policy decision making as opposed to individual clinical decision making. Relevant to the issue of the aggregation of individual preferences are two questions -- should individual preferences be aggregated and if so, how? The first question has been the subject of contentious debate among economists for over a century. Perhaps the most frequently cited argument against the aggregation of individual preferences was Arrow's Impossibility Theorem (Arrow, 1951a). In this classic piece, Arrow (1951a) considered the expected group decision based on the individual preferences of the members of the group. After laying out a set of reasonable assumptions about how aggregate decisions should not contradict the apparent preferences of individual group members, Arrow (1951a) demonstrated how aggregate decisions can violate the apparent preferences of individual decision makers.
However, Kaplan, Feeny, and Revicki (1993) have noted two reasons why Arrow’s Impossibility Theorem may not be applicable to the aggregation of all health status preference measures. First, this theorem was directed at decisions under conditions of certainty as opposed to decisions under conditions of uncertainty. Although the distinction between these two environmental conditions of decision making is explained in Chapter II, at this point it should be noted that not all health status preference measures reflective of preferences under conditions of certainty -- therefore, Arrow’s Impossibility Theorem may not apply to such measures. Second, Arrow (1951a) assumed that the metric underlying preferences was not meaningful and not standardized across individuals. In other words, this theorem was formulated with preferences measured on an ordinal scale of measurement, not an interval or ratio scale of measurement. As explained earlier, a major requirement of cost-utility analyses is that the preferences used in the analyses be measured on at least an interval scale of measurement. Therefore, as Kaplan, Feeny, and Revicki (1993) have noted, Arrow’s Impossibility Theorem may not be applicable to the aggregation of health status preferences as they are currently measured.

Although Arrow’s Impossibility Theorem was formulated with ordinal preferences in mind and is therefore not strictly applicable to the case of cardinal (i.e., interval- or ratio-scale) preferences, the issue of aggregation of cardinal preferences has also been debated in the literature. For example, Dyer and Sarin (1979b), Harsanyi (1975), Hildreth (1953), and Keeney (1976) have all argued that aggregation of cardinal preferences is valid while Kalai and Schmeidler (1977) have argued the opposite. In the
face of such equivocality, Torrance, Boyle, and Horwood (1982) have insightfully noted that "... in order to make social decisions, and in the very process of making those decisions, individual preferences must be and are compared. The question, then, is not whether to make such comparisons, but how to make them" (p.1054; emphasis in original).

If one accepts this argument, the second question -- i.e., what should the basis of aggregation be -- needs to be addressed. In this regard, the alternative bases of aggregation that have been suggested in the literature include the arithmetic mean, geometric mean, or median (see Torrance, Boyle, and Horwood, 1982). On this issue, Torrance, Boyle, and Horwood (1982) have recommended the arithmetic mean as the aggregation method. It should be noted that the arithmetic mean is also the aggregation method recommended by Harsanyi (1975), Hildreth (1953), and Keeney (1976) if an egalitarian premise is adopted.

Apart from these methodological issues relevant to the design stage, one also needs to consider the evaluation time frame and the essence of this dissertation -- which methodology should be used to measure health status preferences? These issues will be discussed further in subsequent chapters -- the evaluation time frame will be discussed in the section of Chapter II which deals with the effect of time on the measurement of health status preferences, while the issue of the particular methodology used to measure health status preferences is addressed in several chapters, given its centrality to the purpose of this dissertation.

Relevant to the implementation stage, the more salient issues include whom the respondent should be (i.e., patient, family member/partner,
physician, or society), the mode of data collection, and the site of administration of the instrument. Each of these issues is largely influenced by the specific constraints and purposes of different studies, and will therefore be discussed with respect to how they are handled in this particular study in Chapter V.

**Application-oriented Issues**

The outstanding issues relevant to this class mainly pertain to the barriers encountered in the application of health status measures to their potential uses. Deyo and Patrick (1989) have provided a lucid description of the barriers to the use of health status measures in public policy research and decision making, clinical research, and clinical decision making. Application-oriented barriers mainly include attitudinal and knowledge barriers of users of such measures and practical barriers in the application setting.

Regarding attitudinal and knowledge barriers, clinicians usually have little or no training in the methods and philosophy of the social sciences and survey research methodology, which form the basis of health status measurement. Indeed, some clinicians have mentioned the existence of a "credibility gap" as a reason why they find these methods unappealing (Deyo and Patrick, 1989). A related problem is the perception that subjective information is too "soft" to be used as a basis for drawing definite policy, research, or clinical decisions. Even when investigators and clinicians acknowledge the need to assess health status, they are usually partial to the older, clinician-rated instruments (Deyo and Patrick, 1989). As noted earlier,
it is desirable to have patients assess their own health status so as to get a more complete assessment of physiological status (i.e., signs as well as symptoms) in particular and health status in general.

Regarding practical barriers in the application setting, these mainly include issues that arise in the logistics of measuring and using health status information. Primary among these is the issue of respondent burden on account of the long and sometimes complicated instruments. It is understandable that the last thing respondents want is yet another form to fill out. This problem is compounded if the forms are lengthy and time-consuming to complete. Apart from the important threat of upsetting respondents, there is also the issue of nonresponse error (both total and item nonresponse) that must be considered in such situations. Fortunately, the introduction of shorter health status measurement instruments which are reliable and valid but can be easily filled out in 10-15 minutes offers a potential solution to this problem. Especially notable in this regard are the COOP Function Charts and the SF-36. The introduction of an even shorter version of the SF-36, called the HSQ-12 -- where HSQ stands for Health Status Questionnaire -- since it has only twelve items, is particularly promising (Radosevich and Husnik, 1995). Further, although there were some concerns initially about the ability of patients to reliably assess their own health states, this view has been discredited in the published research on this topic (Breslow, 1972; Patrick and Erickson, 1988).

A related issue is staff burden. In a survey of investigators active in the field of psychological oncology, the most prevalent problem reported in conducting health status measurements was obtaining cooperation from the
medical staff (Aaronson, van Dam, Polak, et al., 1986). It is possible that this is due to the attitudinal and knowledge barriers discussed earlier. However, the additional workload on the staff imposed by the data collection and coordination requirements of health status measurement instruments is also a factor to consider. While this may seem like a relatively minor issue, it has serious implications for the validity of the data collected, e.g., if staff give the impression to respondents that they do not think the surveys are useful or that they do not have time to answer any questions, respondents may not take the surveys seriously or may answer the questions incorrectly, thus compromising the validity of the data collected.

It had been noted previously that the benefits of including health status measures in routine clinical decision making have yet to be realized to the extent that the benefits of using such measures in public policy research and decision making and clinical research have been realized. An important issue in this regard is the financing of such measures in routine clinical practice. In other words, who should pay for routine health status measurement in clinical practice? Should it be the patient, the clinical setting, the patient's insurance company (or other third-party payor), or a combination of these sources? As Deyo and Patrick (1989) have noted, the costs of gathering, scoring, and presenting survey data are amplified in clinical practice, because information must be provided promptly to be of most use. In clinical research, by contrast, it is often feasible to collect data at specified times but score and analyze them batchwise whenever it is convenient. Given the pressures for cost-containment discussed earlier, it would require persuasive evidence of benefit and substantial educational
efforts in order to obtain third-party reimbursement for routine health status measurement in clinical practice. It is the premise of this dissertation that such persuasive evidence lies in providing meaning to health status scores so that they are easily interpretable by all interested stakeholders. This issue is related to the issue of presenting the scores in a more useful manner. In this regard, the emphasis is on the format in which the scores are presented. The format is important because it influences the interpretation and subsequent use of health status scores. The current research to obtain population norms for the SF-36 is noteworthy (Ware, Snow, Kosinski, et al., 1993), as it will allow clinicians to compare their patients to these norms.

Other issues include the actual data collection procedures, i.e., how does a clinician incorporate the health status measurement procedure in a patient encounter of limited duration and when should such measures be taken? Both these issues depend on the individual clinical setting, since every practice setting operates somewhat differently. In some instances, physicians may be directly involved in the data collection procedures, while in other cases, the primary responsibility for data collection may be entrusted to the clinic staff. In either case, educational intervention may be necessary to ensure that the physicians and staff buy into the importance of health status measurement, which will mean overcoming the attitudinal and knowledge barriers discussed earlier.

In spite of the issues discussed in this section, it should be noted that there are some settings in which health status measures are regularly being collected and utilized for clinical decision making (see, for e.g., Lansky, Butler, and Waller, 1992; Wasson, Keller, Rubenstein, et al., 1992). However,
as Deyo and Patrick (1989) have noted, these settings are more the exception than the rule.

In summary, although the methodology of health status measurement has shown significant improvement over the years, there are still many issues which need to be resolved before such measures can achieve their full potential. In order for them to used widely by all relevant stakeholders, health status scores must be considered meaningful. In this sense, it is necessary for scores to be interpretable from the provider’s, payor’s, and patient’s points of view (Greenfield, 1989). As Greenfield and Nelson (1992) have noted, the fundamental issue for providers and payors is the degree to which health status scores reflect the care that was received by the patient and rendered by the individual provider or the larger health-care system. For patients, the fundamental issue is the relevance of health status scores to them. It is this issue -- providing meaning to health status scores -- which is the focus of investigation in this dissertation.
CHAPTER II

INTRODUCTION TO THE STUDY

Chapter Overview

While the previous chapter reviewed the field of health status measurement in general, this chapter focuses on the particular outstanding issue in health status measurement which was investigated in this dissertation. In describing the nature of the issue that motivated the present study, this chapter discusses some of the problems with the interpretation of current health status scores. It is argued that in order to provide meaning to health status scores, it is necessary to measure preferences for health states. Specific issues that need to be considered in measuring health status preferences are discussed and the methodology for health status preference measurement is introduced. After discussing the assumption of procedural invariance and the phenomenon of preference reversal, the chapter introduces discrete choice conjoint methodology as a potential health status preference measurement methodology that may overcome some of the drawbacks of existing preference measurement methodologies. The topic of validating this new methodology is then presented, and specific bases for validation via the principle of convergent validation are discussed. The
specific research questions which were investigated in this study are then presented, with a brief discussion behind the rationale behind each particular research hypothesis. The chapter ends with a discussion of the assumptions and significance of the study.

**Nature of the Problem**

Given the interest in measuring health status and the importance of health status measurement, it may be necessary for clinicians to incorporate such measures into routine clinical decision making. A necessary prerequisite to this, however, -- and indeed for the acceptance of health status measures in general -- is that the meaning of health status scores be clearly understood by clinicians, payors, and patients. Although the field of health status measurement has shown significant progress over the past few decades, especially since the 1970s, there are still many unresolved issues which need further research. One such issue, which served as the motivation behind this dissertation and is the focus of this chapter, is the meaning of health status scores.

**Problems with Interpretation of Current Health Status Scores**

All that can be obtained from most current health status measures is an absolute numerical score. In case of profiles and batteries, it is also possible to get scores for each individual dimension. Although these numerical scores are important as indicators of the patient’s health status,
they are not easy to interpret. For instance, what does a score of 60 on a 0-100 point scale of physical functioning mean? It may be above the midpoint of the scale (i.e., 50), but how does it compare to the scores of individuals who are in better, worse, or the same physical functional status? In this sense, most current health status scores are rather limited in terms of the information obtainable from them, since they indicate how the patient is doing relative to an arbitrary statistical midpoint of the scale. Therefore, it may be said that most current health status measures can only be evaluated relative to a statistical norm.

Although the current research in progress to develop population norms for the SF-36 (Ware, Snow, Kosinski, et al., 1993) may answer some of the questions raised above, the scores will only contain meaningful information in a relative sense. In this regard, the resulting health status scores will at best provide one more piece of information, i.e., how the patient is faring with respect to a given population. Therefore, these resulting health status measures can be evaluated with respect to statistical as well as population and inter-individual norms.

Although this information may be important, it totally ignores the patient's perspective. The relevant question in this regard is what does the patient think about the score of 60 on the physical functioning scale? If the patient is content with his or her level of physical functioning, it may be a moot point to compare the score to that of a given population. Further, if the patient is not content with his or her level of physical functioning but yet his or her physical functioning score falls within or above the population norm, reliance on such relative measures will obscure this fact. In such cases, the
clinician may not consider any therapeutic intervention to help the patient achieve his or her goals for physical functioning. In other words, the clinician would not consider the patient's preferences for health outcomes in making a therapeutic decision. As explained in the previous chapter, failure to incorporate patients' preferences for health outcomes in clinical decision making may lead to negative outcomes such as decreased health-related quality of life for the patient.

Therefore, what is needed is a measure of health status which can be evaluated with respect to not only statistical and inter-individual norms, but also intra-individual norms. Such a measure should provide information about the patient's value of that score, defined as the worth or preference of that particular score (i.e., health state) compared to another. In other words, such information would tell the user how much better one health state is compared to another, as assessed by the patient. This information is useful for not only clinicians but also third-party payors, who may then appreciate the importance of using health status measures in routine clinical practice, thus addressing some the financing issues raised earlier. Finally, such information would put the patient in his or her rightful place in the health-care system -- at the center, since these health status scores would reflect the preferences of the patient. While it is acknowledged that the scores obtained from health status indices reflect individuals' preferences for various health states, there are certain issues in the way these scores were generated which may be cause for concern.

Therefore, in order to provide meaning to health status scores, it is necessary to determine patients' preferences for health states. Since each
individual health state is nothing but a combination of specific levels of each dimension of health status, patients' preferences for health states are a function of the combination of their preferences for health status dimensions and each individual level within each health status dimension. Inherent in this determination process is the issue of the relative importance of health status dimensions.

**Relative Importance of Health Status Dimensions**

In interpreting the results of research which has multiple indicators of the same construct, it is important to make decisions about which indicators are more important than others. This is as true for public policy decisions as for decisions involving the care of individual patients and groups of patients. When the evidence from one indicator conflicts that from another, there must be some way of deciding which indicator carries the most weight. The same argument holds for health status measurement instruments, which are usually multidimensional.

Ware and Young (1979) have observed that the World Health Organization or WHO definition of health (World Health Organization, 1947) assumes equal importance of physical, mental, and social health; however, not all people value these components equally. Therefore, if health status scores are to be considered meaningful, there must be some way of deciding the relative importance of the different health status dimensions.

The need for understanding the relative importance of different health status dimensions is best illustrated by an example. Suppose there are
two antihypertensive drugs a physician is considering prescribing to a patient. Both drugs are equivalent in terms of blood pressure control and both offer the convenience of once-daily dosing. The only difference between them is in terms of their effect on health status. The first drug has beneficial effects on physical functioning and social functioning, but detrimental effects on mental functioning. On the other hand, the second drug has beneficial effects in mental and social functioning, but detrimental effects on physical functioning. For the sake of simplicity, assume that the change in each individual dimension score on account of these drugs (as measured by effect sizes) is the same. Which drug should the physician choose? This choice involves weighing the different health status dimensions against each other to see which is the most important. Since the outcomes of each drug are multiple and move in different directions, tradeoffs have to be made -- and making tradeoffs involves preferences.

Physicians face such decisions regularly in clinical practice. They ultimately make a general interpretation of the profile of health status scores by applying some sort of internal weighting scheme, based on "clinical expertise." The point to note is that evaluations about the relative importance of health status dimensions is common. Typically, however, as Kaplan and Coons (1992) have noted, such evaluations are made implicitly, arbitrarily, and in an idiosyncratic way. Moreover, these evaluations are usually based on only the physician's preferences on the basis of his or her "clinical expertise."

Given the need to incorporate patient preferences for health states into the therapeutic decision, it is necessary that the process by which the
relative importance of health status dimensions is evaluated be made explicit. Further, it is necessary that the resulting relative importance or preference weights reflect the evaluation of the patient, and not the physician or any other individual. In this way, the health status scores obtained will have meaning as described earlier. There have been some attempts in the published literature to make this process more explicit and incorporate the patient’s point of view, most notably the work by the San Diego group of researchers (see Kaplan and Anderson, 1990), the McMaster group of researchers (see Torrance, Furlong, Feeny, et al., 1995), and the EuroQol group of researchers (see Rosser and Sintonen, 1993). Indeed, the issue of weighing the relative importance of health status dimensions is critical in the construction of health status indices. The issue of relative importance of health status dimensions is further complicated by the potential for interactions among the dimensions of health status.

**Interactions Among Health Status Dimensions**

Keppel (1982) has noted that an interaction is said to be present when the effects of one independent variable on the dependent variable change at different levels of other independent variables. Cohen and Cohen (1975) have used the term “conditionalized” to express interactions, i.e., an interaction is present when the effects of one of the independent variables are “conditionally related” to the other independent variables. In the case of preferences for health states, an interaction is said to be present when the patient’s preference for a particular health status dimension changes at
different levels of the other health status dimensions. The important issues in this regard are how likely is it that health status dimensions can interact with one another and what are the implications of such interactions.

Likelihood of Interactions Among Health Status Dimensions

Before addressing the likelihood of interactions among health status dimensions, it is necessary to specify what are these dimensions. As noted in the previous chapter, there is no agreement in the published literature on what dimensions need to be included in a health status measure. For the purpose of this section, however, the point of the likelihood on interaction among health status dimensions can be made with respect to the basic dimensions identified in the section on clarification of terminology, i.e., functional status (including physical function, mental function, and social function), physiological status, and general health perceptions. It will be shown that there is reason to believe that there is a likelihood of interaction among these basic health status dimensions. Evidence of correlations among these basic dimensions from published research will then be presented to support this argument. Further, since interactions are expected among the basic health status dimensions, it can be expected that the same would hold true with a more detailed breakdown of health status dimensions, at least as operationalized in the existing health status measurement instruments.

Although physical, mental, and social function may have been conceptualized as distinct dimensions of health, many descriptions of their interrelationships have been offered, as illustrated by the following examples. A person who suffers a loss of a limb may decrease his or her
physical activities, which may subsequently result in decreased social activities and more emotional stress. A person experiencing emotional distress may be more likely to decrease his or her social activities or even be more likely to experience physical injury (Ware, Brook, Davies-Avery, et al., 1980).

This line of reasoning can readily be extended to include physiological status and general health perceptions. For instance, if an antihypertensive patient experiences adverse effects such as dizziness and sexual impotence due to his drug therapy (i.e., decreased physiological status), this may cause him to decrease his social activities and also cause emotional stress. If one includes general health perceptions as a separate dimension in a health status measure it will be highly correlated with all the other dimensions, since these general health perceptions are nothing but an overall assessment of a patient's health status. These examples suggest that changes in any one dimension of health status may be related to changes in other health status dimensions, and that the dimensions may both directly and indirectly affect each other.

To present these arguments in relation to specific health status measurement instruments, consider the following instances. In the case of the SF-36, which has eight dimensions [i.e., physical functioning, social functioning, role limitations due to physical problems, general mental health (including psychological distress and well-being), role limitations due to emotional problems, vitality (including energy and fatigue), bodily pain, and general health perceptions], role limitations due to physical and emotional problems could be expected to interact with physical and general
mental health respectively, which in turn could interact with general health perceptions. In the case of the latest version of the McMaster Health Utility Index (i.e., Mark III), which has eight dimensions [i.e., vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain], ambulation may be expected to interact with pain. Finally, in the case of the revised version of the EuroQol instrument, which has five dimensions [i.e., mobility, self-care, usual activities, pain/discomfort, and anxiety/depression], mobility and self-care may be expected to interact with one another, as might self-care and usual activities.

On the other hand, the Quality of Well-Being Scale, which has four dimensions [i.e., mobility, physical activity, social activity, and symptom/problem complex], specifically omits mental function as a separate health status dimension, arguing that the effect of mental function on health status is manifest through the other dimensions. This manifestation is arguably the result of interactions of mental function with the other health status dimensions.

These expected interrelationships between health status dimensions have been shown to exist in the published literature. For example, physical and mental function (Bradburn, 1969; Gilson, Bergner, Bobbit, et al., 1978; Social Psychiatry Research Unit, 1977), physical and social health (Fine, 1975; Greenblatt, 1975; Jeffers and Nichols, 1961; Palmore and Luikart, 1972), and also mental and social function (Bradburn, 1969, Fine, 1975; Greenblatt, 1975; Klemmack, Edwards, and Carlson, 1974) have all been shown to be positively correlated. Moreover, the growing literature on
psychoneuroimmunology demonstrates the conditional nature of physical and mental health outcomes (see, for e.g., Biondi and Pancheri, 1985).

Further, all eight dimensions of the SF-36 have been shown to be positively correlated with one another. Indeed, the developers of the SF-36 have noted, "... despite the complexity of interpretation inherent in measures of role and social disability, vitality, and perceptions of health, they are essential qualities to measure and obtain a synergistic (italics added) and comprehensive assessment of the burden of disease and/or treatment on patients' everyday functioning and well-being (McHorney, Ware, and Raczek, 1993 p.261)." By noting the synergy between the dimensions of health status measured by the SF-36, the developers are acknowledging the possible existence of interactions between health status dimensions.

Although research has not been published on the latest version of the McMaster Health Utility Index (i.e., Mark III), results of tests on the type of underlying model -- additive, multiplicative, or multilinear -- for the previous two versions (i.e., Marks I and II) are available. Additive models do not allow interactions, multiplicative models allow limited interactions, and multilinear models allow all types of interactions among model components. In both versions, the additive model was not found to be representative of the data. Instead, the data were best represented in either case by a multiplicative model (Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992). This is powerful evidence for the existence of interactions among health status dimensions, since both additive and multiplicative models were tested and the multiplicative model was found to be better than the additive model. In either case, the multilinear
model was not tested because it was thought that it would be impractical and unlikely, on the basis of existing literature.

There have also been some attempts to model the EuroQol data. The results of two such attempts have been presented at the EuroQol Plenary Meetings and are available from the EuroQol group. Hence they shall be reviewed here. Both attempts involved the revised version of the EuroQol. The first attempt, by van Hout and McDonnell (1992) used data collected from a mail survey of the EuroQol in Rotterdam in 1991 (Bonsel, Essink-Bot, van Hout, et al., 1992; Stouthard and Essink-Bot, 1992) and modeled the EuroQol data using stepwise linear regression. These researchers tested both the additive main-effects only model form as well as a type of multiplicative model, specifically a main-effects model with first order (i.e., two-way) interactions. On the basis of their analysis, they rejected the additive main-effects only model in favor of the main-effects model with first order interactions. Specifically, they found that the interacting dimensions were: (a) mobility and self care, (b) mobility and usual activities, (c) mobility and anxiety/depression, (d) self-care and pain/discomfort, (e) usual activities and pain/discomfort, and (f) pain/discomfort and anxiety/depression (van Hout and McDonnell, 1992).

The second attempt at modeling the EuroQol data, by Dolan and Kind (1994) is more recent, and is based on data collected from mail surveys in Finland (Ohinmaa, Sintonen, and Pekurinen, 1994) and Frome, UK (Gudex, 1992). This attempt at modeling the EuroQol data used two types of analytical techniques -- least squares regression (linear as well as non-linear regression) and iterative regression analysis. The least squares regression
analysis was performed on the Finnish data (Ohinmaa, Sintonen, and Pekurinen, 1994), while iterative regression analysis was performed on the Frome data (Gudex, 1992). The results of the linear least squares regression analysis rejected the additive main-effects only model in favor of a multiplicative model which included main-effects and first order interactions. In this regard, the general model form estimated from these data was similar to that estimated by van Hout and McDonnell (1992).

However, there were differences observed in which specific dimensions interacted with one another. The study by Dolan and Kind (1994) found the following dimensions interacted with one another: (a) mobility and self-care, (b) mobility and usual activities, (c) mobility and pain/discomfort, (d) self-care and usual activities, (e) self-care and pain/discomfort, and (f) usual activities and pain/discomfort (Dolan and Kind, 1994). It is particularly noteworthy that the dimension of anxiety/depression was not found to interact with any other health status dimension in this study. However, the parameter estimates obtained from this model were sometimes counter-intuitive, in the sense that the intermediate levels for mobility as well as for anxiety/depression received lower preference scores than the worst levels for these two dimensions.

These counterintuitive results prompted Dolan and Kind (1994) to fit a nonlinear model to the data, in which the drop from a ‘good’ state to a ‘not so good’ state would be greater than the drop from a ‘bad’ state to an ‘even worse’ state. Accordingly, they used nonlinear least squares regression with the natural log of the observed health status values as the dependent variables. While the model gave a reasonably good fit with only main
effects, there were still some counterintuitive results with respect to the anxiety/depression dimension. In this regard, the major purpose for which a nonlinear model was fit to the data was not achieved. Finally, the iterative regression analysis yielded no single unique solution to the data. On the basis of these results for the EuroQol, it seems appropriate to conclude that interactions among the health status dimensions of the EuroQol are likely.

Further evidence for the existence of interactions among health status dimensions has been presented by Rose (1980) and Veit, Rose, and Ware (1982). In her dissertation research on patient preferences for health states defined in terms of physical, mental, and social function, Rose (1980) found interactions among all three dimensions of health status. Veit, Rose, and Ware (1982) studied university students' preferences for health states defined in terms of only two dimensions of health status -- physical and mental function. Although these researchers hypothesized an additive model, they found systematic interactions between physical and mental function. This study was particularly noteworthy in that it used the functional measurement approach (which will be discussed in Chapter III) to measure health state preferences, which provides a stringent test of the underlying model form.

The evidence provided by these studies on the presence of interactions among health status dimensions seems compelling. Indeed, Frost (1973) has specifically noted, "In the health setting, the behavioral assumptions implicit in the simple additive ... function are overly restrictive (p.212)." Finally, it should be noted that although the Quality of Well-Being Scale was tested for interactions between its dimensions, interactions were
found to be absent (Kaplan, Bush, and Berry, 1976). Given the method of construction of this instrument and the nature of the health status dimensions included in it, this is not very surprising. This latter issue will be addressed in Chapter III, when the problems with the currently used methodologies for measuring health state preferences are discussed.

**Implications of Interactions Among Health Status Dimensions**

Given the evidence for the existence of interactions among health status dimensions, it is necessary to consider the implications of such interactions on patient preferences for health states. Moreover, it is also important to consider the effect of not accounting for such interactions while assessing patient preferences for health states.

The presence of interactions among health status dimensions often requires more complexity in theoretical explanations of data than would be the case if no interaction were present. In the case of preferences for health states, this would mean that it is less useful to consider a patient's preference for a single health status dimension, since the effect of that particular dimension depends on the particular levels of the other dimensions. For instance, there is little use in considering a patient's preference for physical function only, because this preference is dependent on the patient's level of well-being on the other dimensions of health status. Indeed, if a patient's preference for physical function is considered in isolation to the other dimensions of health status, it is possible that the preference score obtained is biased. This observation underscores the need for not only multidimensional health status measurement instruments, but also
measurement methodologies which can simultaneously manipulate more than one variable in measuring preferences.

Moreover, health state preferences which are assessed explicitly recognizing interactions between the health status dimensions could provide better understanding of the true structure of patients' preferences for health states. This is essential if such preferences are to be incorporated in clinical decision making because if the preferences assessed are not actually representative of the patient's true preferences, it does little good to assess them since they may not lead to the positive outcomes described in the previous chapter. Indeed, incorrectly assessed preferences may have the opposite effect on patient outcomes.

The issue of interaction among attributes or dimensions of constructs under evaluation is not a new one. Concerns about representative study design in the presence of such interactions have been raised repeatedly in the psychometric and marketing literature (see for e.g., Brunswick, 1955; Cattin and Wittink, 1982; Green, Helsen, and Shandler, 1988; Huber and McCann, 1982). Several possible approaches to incorporate interactions in the study design have also been suggested (Brunswick, 1955; Carmone and Green, 1981; Goldberg, Green, and Wind, 1984; Louviere and Woodworth, 1988). However, empirical studies on the effects of not incorporating interactions in study designs which have correlated attributes or dimensions have yielded mixed results.

For instance, Green, Helsen, and Shandler (1988, 1989) have reported two relevant studies based on student apartments. The first study used three calibration designs -- the first with orthogonality among all attributes (i.e.,
all attributes were assumed to be uncorrelated with each other), the second with price as a function of orthogonal nonprice attributes, and the third with price as a function of Pareto-optimal non-price attributes. No significant differences were found between these models in terms of predicting preferences on a set of Pareto-optimal price covarying validation profiles (Green, Helsen, and Shandler, 1988). The second study compared a design in which all attributes were orthogonal to one in which the nonprice attributes were orthogonal, but price covaried with the quality of apartments. In other words, the interaction between price and quality of apartments was modeled as a separate variable. In this case, the fully orthogonal design yielded better results than the second design in terms of predicting preferences on a validation set (Green, Helsen, and Shandler, 1989). One salient point about both these studies should be made explicit -- only one attribute (i.e., price) was assumed to interact with the other attributes. If the researchers had allowed all the attributes to interact with each other, the results might have been different.

Moore and Holbrook (1990) compared orthogonal main-effects only (i.e., no interactions allowed) and interactive designs for predicting students' preferences and likelihood of purchase of cars. They found that both designs were equivalent in terms of predicting preferences and likelihood of purchase on a validation set. They concluded that although plausible conceptually, interactions may be less relevant in practice. Accordingly, they recommended using orthogonal main-effects only designs in order to capitalize on their statistical properties (Moore and Holbrook, 1990). However, it should be noted that this study included only a subset of
possible interactions among attributes and incorporated lower than actually-occurring correlations in the real environment. Further, the product profiles evaluated as part of the interactive design were perceived by respondents to be more realistic than those of the orthogonal main-effects only design.

On the other hand, other empirical studies have shown designs incorporating interactions to have better predictive validity than orthogonal main-effects only designs. DeSarbo, Mahajan, and Steckel (1985), using a procedure similar to that of Moore and Holbrook (1990) -- i.e., incorporating lower than actually-occurring correlations among attributes than in the real environment -- found that interactive designs are better able to predict apartment choices than orthogonal main-effects only designs. In a separate study, these results were replicated by Steckel, DeSarbo, and Mahajan (1988). Finally, Finkbeiner and Lim (1991) demonstrated the importance of incorporating interactions in study designs with correlated attributes as well as an indication of the magnitude of the resulting noise and bias in the parameter estimates obtained if interactions were not incorporated into such designs. An especially noteworthy point about the Finkbeiner and Lim study was that it was done on a pharmaceutical product.

Johnson and Mai (1979) conducted a simulation study to examine the suitability of additive models in situations which are known to be non-additive and found that in some cases additive models could misleadingly differentiate between alternatives. In other words, using an additive model in such situations led to falsely differentiating between alternatives which were actually similar. This study has implications for the phenomenon of preference reversal (which will be discussed later in this chapter), i.e., if
some alternatives are truly so similar to each other in terms of preference scores, at any given opportunity an individual might select any one, depending on several aspects of the decision situation. By constraining a certain alternative to be the most preferred on all occasions, the linear model would not account for such variations in individual behavior.

Hagerty (1986) made an important distinction between the predictive accuracy of models estimated at the individual level versus models estimated at the group level. In his study, he showed that while simpler (i.e., additive) models may yield more accurate results at the individual level (even if these additive models are "false") in terms of predictive accuracy, this is not the case at the group level. He explained his results by showing that the additional parameters that are required to be estimated when interactions are added to a model are estimated with larger variance, which tends to decrease the predictive accuracy of multiplicative models at the individual level. However, at the group level, the Law of Large Numbers suggests that some types of errors would cancel out when aggregated over several respondents, thus yielding more accurate predictions on account of the specification of the true model form of the respondents (Hagerty, 1986).

In light of these mixed findings from empirical studies, it is advisable to consider each case for incorporating interactions into study designs separately. For the present study, the issue on hand is the interactions among health status dimensions. The evidence for the existence of interactions among health status dimensions -- especially the tests for the additive and multiplicative models for the Mark I and Mark II McMaster Health Utility Indices and the EuroQol -- is compelling. Further, there is the
issue of what use these health status scores are going to being put to, and how can interactions affect this use. As noted, the major use of health status scores is to understand patient preferences for health states for application in clinical and public policy decision making. In both these applications, the numerical score itself is important. For instance, if these scores are to be used in program evaluation as the weights for Quality Adjusted Life Years, a slight error in their estimation could make a marked difference in the Quality Adjusted Life Years of a program under evaluation, thus making the program in question look unduly favorable or unfavorable depending on the direction of the error. Given the potential for obtaining biased parameter estimates (i.e., health status scores) from orthogonal designs, it is necessary to incorporate interactions between health status dimensions explicitly in study designs. This is in spite of the statistical advantages of orthogonal main-effects only designs, i.e., simpler estimation and interpretation of parameter weights.

This conclusion is supported by the Working Group on Quality of Life Research in Oncology, which observed, "Future research should be based on conceptual models of quality of life that incorporate the multidimensionality of the construct and that attempt to explicate the interrelationships among quality of life domains (Aaronson, Meyerowitz, Bard, et al., 1991 p.842)." The importance of assessing interactions in providing meaning to health status scores was noted by Ware (1991), "To understand patient health outcomes, different health components need to be measured and interpreted separately and in combination (the latter when trade-offs are involved)" (p.776 emphases added).
The Role of Risk in Measuring Health Status Preferences

Apart from interactions between health status dimensions, it is also desirable to incorporate the notion of attitude toward risk while measuring health status preferences. In presenting this argument, the role of risk in general decision making will first be discussed, followed by a discussion of why it is important to use risky decision making contexts in measuring health status preferences.

Risk in Decision Making

There are two basic classes of environmental conditions under which decisions on health status preferences can be made -- certainty and uncertainty. Certainty refers to an environment in which the individual has full knowledge of the occurrence of the consequences of each of the alternatives under consideration. Decisions made under certainty are known as riskless decisions. Uncertainty, on the other hand, refers to an environment in which the individual has less than complete knowledge of the occurrence of the consequences of each of the alternatives under consideration.

Within the broad class of uncertainty, Knight (1921) has drawn a distinction between two types of uncertainty -- risk (or measurable uncertainty) and unmeasurable uncertainty. According to this distinction, a decision is said to be risky when the decision maker is uncertain about the occurrence of the consequences of each of the alternatives under consideration but is able to express this uncertainty in the form of a probability distribution over the possible consequences of each alternative. Knight (1921) defined unmeasurable uncertainty as a condition when the decision maker is
uncertain about the occurrence of the consequences of each of the alternatives under consideration but is unable to express this uncertainty in the form of a probability distribution over the possible consequences of each alternative.

Ellsberg (1961) further distinguished between two types of unmeasurable uncertainty -- ambiguity and ignorance. According to him, ambiguity was a state when the decision maker is uncertain about the occurrence of the consequences of each of the alternatives under consideration and is able to express this uncertainty in the form of a variable probability distribution over the possible consequences of each alternative. Therefore, the distinction between risk and ambiguity is in terms of the "degree of confidence" the decision maker has in his or her estimation of the probability distribution over the possible consequences of each alternative. Finally, ignorance according to Ellsberg (1961) is exactly was unmeasurable uncertainty was to Knight (1921) -- a condition when the decision maker is uncertain about the occurrence of the consequences of each of the alternatives under consideration but is unable to express this uncertainty in the form of a probability distribution over the possible consequences of each alternative.

Of all these conditions, there is a large body of literature on decision making under certainty and risk. The literature on decision making under ambiguity is growing, but is still at a nascent stage in its development, while the area of decision making under ignorance is relatively unexplored (for a review of this literature, see Camerer and Weber, 1992). Therefore, this
discussion will confine itself to the conditions of decision making under certainty and decision making under risk.

It should further be noted that in all cases where probability distributions may be estimated over the possible consequences of each alternative, the probabilities in question may be subjective or objective. Specifically, for the purpose of this dissertation, the subjective probability approach is adopted. This is in accordance with the Bayesian approach to decision making and views probability as a measure of the decision maker's knowledge or beliefs about states of the world (Raiffa, 1968a; Ramsey, 1931; Savage, 1954).

As noted previously, decisions made under certainty are known as riskless decisions, while decisions made under risk are known as risky decisions. Accordingly, two types of health state preferences may be distinguished -- value-based and utility-based preferences. Value-based preferences are choices or judgments regarding health states made under conditions of certainty, while utility-based preferences are choices or judgments regarding health states made under conditions of risk (Keeney and Raiffa, 1976). Strict decision theorists believe that every decision has some amount of risk associated with it; therefore they argue that utility-based preference functions are theoretically superior than their value-based counterparts.

The logical relationship between these two types of preference functions has been examined by some researchers, who conclude that severe limitations constrain such relationships and only a few forms of such relationships may exist (Barron, von Winterfeldt, and Fischer, 1984; Dyer
and Sarin, 1979; Sarin, 1982; von Winterfeldt, 1979 — this body of research is reviewed in Chapter IV). However, von Winterfeldt and Edwards (1986) have stated that the distinction between utility-based and value-based preferences is spurious on account of the following arguments: (a) in practice, the conditions of certainty are never satisfied (i.e., there is no such thing as a "true" value-based function), (b) risk aversion can frequently be explained by marginally decreasing value-based functions and/or regret attributes of a value-based function, (c) in the case of repetitive choices, risk aversion will be eliminated, and an argument can be made that all choices are repetitive, and (d) error and method variance within value and utility measurement procedures obscure any differences in practice which may be present in theory.

However, the few studies measuring both value-based and utility-based preferences conducted in the context of health status preference measurement have not supported von Winterfeldt and Edwards' (1986) argument (see, for e.g., Torrance, 1976a; Torrance, Zhang, Feeny, et al., 1992; Wolfson, Sinclair, Bombardier, et al., 1982). Each of these studies found significant differences in value-based and utility-based preference functions. An especially noteworthy point about these studies is that they used different measurement strategies to measure health state preferences — three of the studies used holistic preference measurement strategies and the other used the conditional utility-based function procedure, which is one of the explicitly decomposed preference measurement strategies. (These different preference measurement strategies are discussed in Chapter III). Given this finding, it may be concluded that the notion of attitude toward risk in
decision making is an important factor to consider in the measurement of health status preferences.

Risk and Health Status Preferences

As noted above, it is also desirable to incorporate the notion of attitude toward risk while measuring health status preferences. This is because an individual's attitude toward risk taking may affect his or her preferences for health states. Indeed, the importance of the notion of risk in understanding consumer behavior in general has received increasing attention in the marketing and psychology literature (Mitchell, 1992; Taylor, 1974). To illustrate, consider the following example: if asked to choose between two alternative therapies, of which the first would greatly improve the respondent's current health status and the second would noticeably improve the respondent's current health status (although not as much as the first option), most respondents would choose the first option. However, a different alternative might be chosen if the notion of risk is introduced. Specifically, if the first option had a 50% probability of producing the desirable outcome (mentioned above) but a 50% probability of worsening the respondent's current health status through some adverse effect, while the second option had an 90% probability of producing the desired health outcome (mentioned above) and a 10% probability of worsening the respondent's current health status by the same adverse effect and to the same extent as the first option, respondents may choose the second option.

This change in respondents' preferences may be explained by the fact that they are risk-averse (i.e., wary of taking chances) with respect to their
health status. In the context of the example given above, respondents would prefer a smaller improvement in their current health status if it meant that they would have a lesser likelihood of suffering some adverse effect. Therefore, by not considering respondents' attitudes toward risk, the researcher may not gain complete understanding of patients' preferences for health states. The implications of this observation are that health status preference measurement should occur under risky conditions.

Indeed, Mehrez and Gafni (1990) have recommended that since individuals usually do not make decisions under conditions of certainty, methods for the measurement of individual preferences between alternative uncertain consequences should capture individuals' attitudes toward risk. These methods are also appropriate for public policy health-care decision making, as noted by Ben-Zion and Gafni (1983). Such methods are usually introduced as part of the scaling method used to measure health status preferences, and will be discussed in Chapter III.

Time and the Issue of Prognosis in Measuring Health Status Preferences

Another important factor which needs to be considered in measuring health status preferences is time. Indeed, as Gafni and Torrance (1984) have noted, "... the health of an individual has a time aspect inextricably bound to it (p.440)." One particular scaling method used to measure health status preferences, the time tradeoff method, specifically uses time in a given health state as the response measure. The effect of time on health status preferences is multifaceted, and includes: (a) the timing effect -- the point in
time when the individual will experience the health state under evaluation, (b) the duration effect -- the extent of time for which the given health state will be experienced by the individual, and (c) the sequence effect -- the preceding health states experienced by as well as the future health states expected for the individual. It should be noted that although these effects are distinct from one another in theory, they may be confounded in the measurement process. Indeed, Gafni (1995) has argued that current methodologies for measuring health status preferences do not allow for the separation of the timing and the sequence effect. Related issues include the stability of health status preferences over time in respondents who, subsequent to the process of evaluation, experience versus do not experience the evaluated health state(s). Each of these effects and issues will be discussed in turn in this section.

The Timing Effect

The timing effect deals with the specific point in time when the individual will experience the health state under evaluation. The economic concept of time preference is central to the timing effect (Olsen and Bailey, 1981). Specifically, for a given quantity of an economic "good" (defined as something yielding pleasure), a person with a positive time preference will prefer current to future consumption. On the other hand, for a given quantity of an economic "bad" (defined as something yielding displeasure), a person with positive time preference will prefer future over current consumption (it is acknowledged that the definition of "good" and "bad" is inherently subjective and may be relative). This is the concept underlying
the use of the principle of discounting in pharmacoeconomic evaluations. Although the assumption of positive time preference is the usual one employed in most economic analyses, it has been reported that some individuals exhibit negative time preferences, i.e., prefer future over current consumption for economic "goods" and current over future consumption for economic "bads" (Loewenstein and Prelec, 1991).

The implications of the timing effect for health status preference measurement depend on the nature of time preference of the individual (i.e., positive or negative) and the nature of the health state under evaluation (i.e., "good" or "bad"). For instance, an individual with positive time preference may express a lower overall preference for a "good" health state simply because it would be experienced at some point in the future as opposed to the present. In other words, if the same "good" health state were to be experienced in the present, this individual would express a higher overall preference. On the other hand, an individual with negative time preference may express a higher overall preference for a "bad" health state simply because it would be experienced in the present as opposed to some time in the future. If this same "bad" health state were to be experienced some time in the future, this individual would express a lower overall health status preference. In this way, the expressed overall health status preference may be a function of when the health state under evaluation is to be experienced by the respondent.
The Duration Effect

As noted above, the duration effect is nothing but the extent of time for which the given health state will be experienced by the individual. The implications of this effect on health status preference measurement are straightforward -- when evaluating "good" health states, individuals would express higher overall preferences for longer durations and lower overall preferences for shorter durations. The opposite would be the case for evaluations of "bad" health states. Of all the effects of time on health status preferences, the duration effect is the easiest to isolate, as it can be introduced as part of the question stem, e.g., the instructions to respondents may be, "Assume you will experience this health state for a period of ten years."

The Sequence Effect

The sequence effect refers to the order of health states ("bad" and "good") experienced before and after the health state under evaluation and its effect on the individual's preference for the health state under evaluation. Gafni (1995) has defined the sequence effect as an "... individual's attitude toward the order of good and bad outcomes over his/her life time" (p.9)." The concept of sequence effect represents the phenomenon that consumption at one point in time can influence the marginal utility of consumption at another, or that the utility assigned to a package of consequent events is not equal to the sum of discounted utilities of each event separately (Gafni, 1995). Loewenstein and Prelec (1993) have argued and demonstrated that such effects can and do occur in practice.
Applied to health status preference measurement, the sequence effect could lead to higher or lower expressed overall health status preferences. For instance, if the preceding health state(s) were “bad” and the health state under evaluation was “good,” the respondent may express a higher overall preference (i.e., an “improvement” effect). An example of this case may be a mild headache following a migraine attack. Further, if the future expected health states are considered “good” by the respondent, the health state under evaluation may be the beneficiary of a higher overall preference than would be obtained if the future expected health states were “bad” — this is because respondents may be more willing to endure a given health state in expectation of better health states in the future (i.e., an “anticipation” effect). An example of this case may be nausea as a side effect of a drug which would prevent epileptic attacks in the future.

An important component of the sequence effect is the concept of prognosis, or the expected future health states an individual may experience. Patrick, Bush, and Chen (1973a) have cogently argued for the inclusion of prognosis in any measure of health status. They have observed, “... if two individuals are confined to bed because of gastrointestinal discomfort, one with a mild case of diarrhea and the other with a malignancy, the immediate level of function and comfort may be the same. But the expected course of the two disorders, and therefore, the health status of the two persons, is vastly different. Thus, no precise statement of health status can be made for an individual or a group without knowledge of the expected transitions among the function levels over time (Patrick, Bush, and Chen, 1973a p.7).”
Related Issues

An important related issue is the stability of health status preferences over time. In this regard, there are two different cases of interest. The first case is the likelihood of a change in health status preferences in respondents as a function of natural time, i.e., an "unprompted" change. The second case is the effect of experiencing the health state under evaluation on the overall preference (i.e., hypothetical versus actual health states) for that health state, i.e., a change "prompted" by experiencing the evaluated health state.

The unprompted change is nothing but the effect of the passage of time on the evaluation of a given health state. Lerner (1973) has noted that individuals' preferences for health may change as a function of life stage. Generally, age has -- implicitly or explicitly -- been used as a proxy for life stage. Weinstein and Stason (1977) have also noted the importance of age in health status preference measurement in their definition of Quality Adjusted Life Years (QALYs): "A health-status index is essentially a weighting scheme: each definable health status, ranging from death to coma to varying degrees of disability and discomfort to full health, and accounting for age differences (emphasis added), is assigned a weight from zero to one, and the number of years spent at a given health status $Y_s$, is multiplied by the corresponding weight, $\lambda_s$, to yield a number $Y_s\lambda_s$, that may be thought of as an equivalent number of years with full health -- a number of quality adjusted life years (QALYs)."

However, several empirical studies have found no significant differences in health status preferences attributable to age (see, for e.g., Busschbach, Hessing, and de Charro, 1993; Carter, Bobbitt, Bergner, et al.,
1976; Kaplan, Bush, and Berry, 1978; Rosser and Kind, 1978). One notable exception was the study by Sackett and Torrance (1978), which found that the preferences associated with six of their 15 disease-specific health states were associated with age. It is possible, though, that this observation could have been confounded with the fact that several of the older participants were also experiencing the health state under evaluation. However, because some of the studies contained small sample sizes and had a high degree of variability in the distribution of preferences, it is possible that there might not have been sufficient statistical power to detect differences between age groups.

Apart from these cross-sectional studies which have compared different age groups on the basis of their expressed health status preferences, there have been a few longitudinal studies which have tracked changes in the same respondents' health status preferences over time. For instance, recent investigations conducted in oncology patients have used such longitudinal designs (Llewellyn-Thomas, Sutherland, Ciampi, et al., 1984; O'Connor, Boyd, Warde, et al., 1987). In these studies, the preferences of a single group of respondents for a particular short-term acute health state encountered during the process of treatment were repeatedly assessed while the respondents' own health status changed. The results of these studies indicate that patients' expressed health status preferences do not change as they proceed through treatment (Llewellyn-Thomas, Sutherland, Ciampi, et al., 1984; O'Connor, Boyd, Warde, et al., 1987). Therefore, it is likely that individuals' preferences for health states are not likely to change with only the passage of time. This is encouraging, since it implies health status
preferences measured at one point in time may be valid at some future time. Therefore, it may not be necessary to repeatedly measure individuals' health status preferences.

Regarding the prompted changes, it is possible that when patients experience a particular health state, they learn to cope with it and therefore the general public's fear of and disutility for a given health state may be different than that of patients experiencing that health state. Empirical studies on this issue have yielded mixed results. Like the studies for the "unprompted" changes, these studies are of two types -- cross-sectional and longitudinal.

Sackett and Torrance's (1978) cross-sectional study found that the health state of the respondent was related to the preferences for some but not all health states -- for instance, home dialysis patients expressed higher preference scores for kidney dialysis than did the general public. However, Llewellyn-Thomas, Sutherland, Tibshirani, et al., (1984) have reported that respondents' own health states did not influence their evaluations. However, the comments about small sample sizes and variability of preference distributions made earlier also apply to these studies.

Longitudinal studies on this issue have also yielded mixed results. For instance, Christensen-Szalanski (1984) found that pregnant women's preferences for avoiding pain and using anesthesia during childbirth varied with the passage of time. Specifically, the respondents preferred to not use anesthesia during childbirth when asked one month before labor and during early labor. However, during active labor, their preferences shifted toward using anesthesia. Their preference shifted again toward not using anesthesia.
during childbirth when evaluated at one month postpartum (Christensen-Szalanski, 1984). However, this study may have suffered from a selection bias, since the respondents were women who had paid to enroll in a privately sponsored childbirth class.

In another longitudinal study, Llewellyn-Thomas, Sutherland, and Theil (1993) evaluated laryngeal cancer patients' preferences for different health state descriptions they were likely to experience following a course in radiation therapy. Patients were interviewed before and after they underwent a four-week course in radiation therapy. Their health status at each of these times was also measured, in order to ensure that they had experienced the health state they were evaluating. It was found that these patients' evaluations of imagined short-term health states encountered during radiation therapy for laryngeal cancer remain consistent when those states are experienced at a later time. Thus the results of this study do not agree with those of the Christensen-Szalanski (1984) study. It would therefore seem that the issue of stability of health status preferences after experiencing a given health state needs to be approached on a case-by-case basis.

Therefore, it can be seen that patients' preferences for health states are affected by a multitude of factors, including the relative importance of health status dimensions, interactions between these dimensions, patient attitudes toward risk, the point in time patients are expected to experience the health state, the length of time patients are expected to stay in the particular health state, and the sequence of health states preceding and following the health
state under evaluation. In order to provide true meaning to health status scores, the role of each of these factors and their effect on patients’ preferences for health states must be assessed. At this point, it is necessary to review the existing research on providing meaning to health status scores to see if (and how) these different issues have been addressed.

**Previous Research on Providing Meaning to Health Status Scores**

Although there has been some research conducted on providing meaning to health status scores, this research does not address all the issues raised above. Previous research on providing meaning to health status scores has generally taken three approaches. The first approach involves providing meaning to changes in health status scores, the second tries to compare health status scores with traditional clinical measures, while the third involves providing meaning to the actual score itself.

**Providing Meaning to Changes in Health Status Scores**

The first approach (i.e., providing meaning to changes in health status scores) is of recent origin and is exemplified by the research of Kazis, Anderson, and Meenan (1989), whose premise is that changes in health status scores can have meaning to clinicians and patients only if the changes have broader application than the raw health status scores. Kazis, Anderson, and Meenan (1989) have proposed that a measure of change in health status scores should allow for the translation of results into a numerical quantity that facilitates comparisons with benchmarks of change and provides a basis
for deciding whether a given change in health status is clinically significant or not. Accordingly, they propose using effect sizes as measures of the meaning and magnitude of the changes in health status scores (Kazis, Anderson, and Meenan, 1989).

An effect size represents a standardized measure of change in a group or a difference in changes between two groups. Although a number of approaches are used for calculating effect size, all basically involve taking the mean change found in a particular variable and dividing it by the standard deviation of that variable (Cohen, 1977). Kazis, Anderson, and Meenan (1989) calculated effect size by taking the difference of the health status score means before and after treatment and dividing this difference by the standard deviation of the same measure before treatment. This approach treats the effect size as a standard measure of change in health status scores in a “before-after study” context (Kazis, Anderson, and Meenan, 1989).

As noted, Kazis, Anderson, and Meenan (1989) were interested in using effect sizes to give meaning to health status scores in terms of comparisons with benchmarks of change and deciding if a particular change is clinically significant or not. When used as benchmarks for understanding changes in health status scores, effect sizes can be used at both a general as well as instrument-specific level. At both levels, it is necessary to determine some cut-off points for assessing the relative magnitude of change. In their study, Kazis, Anderson, and Meenan (1989) adopted Cohen's (1977) guidelines for this purpose, i.e., an effect size of 0.20 or less is small, an effect size between 0.20 and 0.80 is moderate, and an effect size over 0.80 is large. Although it is not necessary to use these guidelines for all purposes, it is the
idea that is noteworthy. By interpreting each observed change according to some chosen set of guidelines, it is possible to get some understanding about the meaning of health status scores in terms of the magnitude of the change in health status scores before and after the treatment. These same guidelines are helpful in determining the degree of clinical significance (in terms of mild, moderate, and large clinical significance) of an observed statistically significant change. Effect sizes are useful for this purpose because unlike standard statistical tests of significance, they are not influenced by sample size considerations.

This approach is noteworthy because it attempts to provide universal or standardized meaning to the admittedly abstract health status scores. It is especially suited to health status measurement instruments which are intended to measure changes in health status over time. However, the approach does little to provide meaning to a given individual health status score by itself.

Comparing Health Status Scores with Traditional Clinical Measures

There have been two different methods used to compare health status scores with traditional clinical measures. Both methods are based on the notion that since clinicians are more familiar with traditional clinical measures, it would be useful to express health status measures in terms of their relationship to these traditional measures. By understanding the relation between these measures, it is expected that clinicians would better appreciate health status measures and use them to a greater extent in clinical decision making.
The first of the two methods used to compare health status scores with traditional clinical measures is the effect size approach (Kazis, Anderson, and Meenan, 1989), as described above. In this case, changes in both measures are expressed in terms of effect sizes and comparisons are made between them. In the Kazis, Anderson, and Meenan (1989) study on arthritic patients, it was found that the same treatment which produced moderate to high effect size estimates for health status measures such as pain and physical activity also produced moderate effect size estimates for tenderness and swelling scores. Hence, this study showed that the pain and physical activity dimensions of the health status instrument used in the study tapped into the same dimension that the traditional tenderness and swelling scores did.

The other method involves predicting one set of scores from the other. Kaplan (1987) investigated the relationship between these two sets of measures in diabetic and hypertensive patients. Health status measures included measures of physical functioning, role functioning, and perceived health. Traditional clinical measures included blood sugar levels for diabetic patients and blood pressure levels for hypertensive patients. Physical functioning and role functioning scores but not perceived health were found to be significantly correlated with elevated blood sugar and blood pressure levels. Using ordinary least squares regression, however, the best predictor of blood sugar and blood pressure levels at the end of the study were the enrollment levels of these same clinical measures, not any of the health status measures. Still, all health status measures in this study made significant and independent contributions to the prediction of the traditional
clinical measures at the end of the study (Kaplan, 1987). Patrick (1987) has questioned the direction of causality tested in this study, arguing that the traditional clinical measures, which actually measure physiological status, could predict functional status and general health perceptions. Patrick's (1987) argument is well taken, since the traditional clinical measures are only intermediate and not final outcomes. Health status measures, however, are final outcomes.

There have been two published studies which compared health status as operationalized by the SF-36 and glycemic control (Nerenz, Repasky, Whitehouse, et al., 1992; Weinberger, Kirkman, Samsa, et al., 1994). These studies have yielded mixed results. The first of these studies found an inverted U-shaped relationship between glycemic control and SF-36 scores for insulin-dependent diabetic patients, while a positive linear relationship between these two variables for non-insulin-dependent diabetic patients (Nerenz, Repasky, Whitehouse, et al., 1992). On the other hand, the second study only included non-insulin-dependent diabetic patients and could not find either a linear or a curvilinear relationship between these two variables (Weinberger, Kirkman, Samsa, et al., 1994). It should be noted that the two studies used different samples of patients -- the first study excluded patients above 60 years of age, while the second had a mean age of 63.7 years. Moreover, the patients in the second study had much lower health status scores on the SF-36 dimensions than the patients in the first study. Therefore, it is possible that the relationship between health status and clinical measures may be dependent on age or level of functioning.
Although the idea of comparing health status measures to traditional clinical measures is useful for helping clinicians understand the relationship between the two sets of measures, it can be argued that this approach is less useful than the others. According to the clarification of terminology provided earlier, both types of measures need to be included in any instrument which purports to measure health. Given this, the rationale for comparing them is questionable, i.e., if both sets of scores need to be included since they measure different dimensions of the same construct, why compare them? Moreover, for clinicians to accept and use the new health status measures, they must be shown to provide information over and above that given by the traditional clinical measures. It is in this respect that the third approach is the most promising for, unlike the other approaches, it provides information on the actual meaning of a given health status score.

Providing Meaning to Actual Health Status Scores

This approach is actually the oldest of the three. It was pioneered by the work of Bush and his colleagues, when they developed the Quality of Well-Being Scale (see Kaplan and Anderson, 1990). Since then, Torrance and his colleagues have also done much work in this area (see Torrance, Furlong, Feeny, et al., 1995). The EuroQol group of researchers have also been very active in this area of research (see Rosser and Sintonen, 1993). This is the only approach which provides true meaning to health status scores from the patient's perspective, because it explicitly measures patient preferences for being in a particular health state. Hence, the scores obtained not only reflect the health status of the patient, they also consider the measure of worth the
patient associates with the particular health state. Moreover, depending on the particular model used to represent the relationship among health status dimensions (i.e., additive, multiplicative, or multilinear), it is possible to include interactions among the dimensions as part of this approach. Further, depending on the particular scaling method used, it is possible to incorporate the effect of risk and/or time on patient preferences for health states. Details about the work of these groups of researchers will be provided in Chapter III, when the methodology for measuring health status preferences is discussed in detail. However, at this point, it is necessary to briefly introduce the methodology for measuring health status preferences in order to define the problem investigated in this dissertation more clearly so that specific research questions may be framed.

A Brief Introduction to Health Status Preference Measurement Methodology

Froberg and Kane have provided an excellent review of the methodology for measuring health status preferences (Froberg and Kane, 1989a, 1989b, 1989c, 1989d). They have conceptualized the methodology of health status preference measurement as consisting of four basic steps: (a) defining a health status classification system, (b) deciding upon a measurement strategy, (c) choosing a scaling approach, and (d) implementing the selected methodology (Froberg and Kane, 1989a). At this point, only the middle two steps (i.e., deciding upon a measurement strategy
and choosing a scaling approach) will be introduced since these are the steps pertinent to the problem that was investigated in this dissertation.

Froberg and Kane refer to the overall structure for posing questions to respondents (i.e., having respondents evaluate multiattribute or multidimensional health states versus evaluating each attribute or dimension separately) and the corresponding method of analyzing the data (i.e., multiple linear regression analysis, analysis of variance, logit, or probit) as the measurement strategy. On the other hand, they consider the scaling approach to be the specific task required of the respondent to obtain scale scores for health states (Froberg and Kane, 1989a).

In health status measurement research, the major measurement strategies which have been used are the statistically inferred and the conditional utility function-based strategies. Details about these strategies will be provided in Chapter III. Statistically inferred strategies attempt to develop an algebraic model of respondents' preferences for health states from a series of evaluations of multiattribute descriptions of health states via inferential statistics. The approach adopted in this strategy is therefore "top-down," in the sense that the preference scores for individual attributes or dimensions of health status and the levels used to operationalize these attributes or dimensions are estimated on the basis of the overall evaluations of multiattribute or multidimensional descriptions of health states. On the other hand, the conditional utility function-based strategy adopts a "bottom-up" approach, in which individual attribute preference functions are estimated first and subsequently aggregated into an overall preference
function for health states. The basis for aggregation in the latter strategy is mathematical.

There are two basic types of scaling approaches which have been used in health status preference measurement — direct and indirect. Although each basic type of scaling approach will be discussed in Chapter III, it is necessary to briefly introduce them at this point. In direct scaling, respondents are instructed to make evaluations at a certain level of measurement and the resulting data are treated as such. For instance, respondents may be asked to perform a ratio-level scaling task such as assigning a number which represents the magnitude to which a certain health state under evaluation is preferred to the health state of death. On the other hand, in indirect scaling approaches, respondents are instructed to make evaluations at a certain level of measurement and the data are later converted to a different level of measurement by the researcher(s). For instance, in the method of paired comparisons (which is a particular type of scaling method), pairs of health states are presented to respondents, who indicate which of the two states they prefer, i.e., an ordinal judgment. In order to convert these ordinal-level judgment data into interval-level or ratio-level measures of preference, it is necessary to apply a set of theoretical assumptions based on the variability of the subjects' responses (Froberg and Kane, 1989b). One such set of assumptions is Thurstone's Law of Comparative Judgment (Thurstone, 1927a), which is based on the idea that stimulus differences which are detected equally often are subjectively equal.

Within each basic type of scaling approach, there are several different scaling methods which can be used. These include the rating scale,
magnitude estimation, standard gamble, time tradeoff, contingent valuation, equivalence, paired comparison, and pick-one methods. Details about each of these scaling methods will be provided in Chapter III. As Froberg and Kane (1989b) have noted, each of these scaling methods has its own strengths and weaknesses. For instance, the rating scale scaling method is considered by some researchers as the easiest to use. The standard gamble method is based on the axioms of expected utility theory and, thus far, is the only scaling method which explicitly measures health status preferences under risk. On the other hand, the contingent valuation method is based on the theoretical foundations of welfare economics. Further, the pick-one scaling method is specifically designed to infer respondents' preference functions using choices (as opposed to judgments) as the response mode. Depending on the research priorities, different scaling method(s) may be appropriate for measuring health status preferences.

As noted previously, health status measurement researchers have employed a variety of measurement strategy/scaling method combinations to measure health status preferences. For instance, the researchers at McMaster University have used the conditional utility function-based strategy along with the rating scale, time tradeoff, and standard gamble scaling methods (Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992). Researchers at the University of California at San Diego have used a statistically-inferred measurement strategy along with a rating scale method (Kaplan, Bush, and Berry, 1978; Kaplan and Ernst, 1983), as have the EuroQol group of researchers (Dolan and Kind, 1994; van Hout and McDonnell, 1992).
It should be noted that regardless of the methodology used to measure preferences, the construct of preference is usually an abstract variable that is given empirical interpretation through specific processes of operationalization of scaling methods called response modes. At this point, in order to avoid any confusion in the following discussion on the effect of response mode on health state preferences, it is necessary to clarify the three terms just introduced, i.e., scaling approach, scaling method, and response mode. As noted previously, a scaling approach is the specific task required of the respondent to obtain scale scores for health states. This specific task can be of two types -- direct or indirect. Within each type of scaling approach, there are several scaling methods which can be used to measure health status preferences. Each of these methods is operationalized by asking the respondent to make a particular type of response -- judgments about each of the alternatives under evaluation or choice among the alternatives under evaluation. This specific operationalization of scaling methods (i.e., the specific nature of the response required of respondents) is referred to as the response mode. It is the premise of this dissertation that the effects of response mode on health status preferences have not been sufficiently investigated in the published health status preference measurement literature.

In spite of the variety of measurement strategy/scaling method combinations used, the vast majority of the published studies of health status preference measurement have used the same basic type of response mode for the measurement of health status preferences -- judgmental responses. As noted above, apart from judgments about alternative health states, health status preferences may also be measured using choices among
alternative health states as the response mode. Therefore, preferences may be inferred from data obtained via two basic types of response modes -- judgments or choices.

Louviere (1988a) has noted the difference between judgmental and choice data. According to him, judgmental data are evaluative rankings or ratings of a set of multiattribute or multidimensional alternatives obtained from individuals or groups of individuals, which are assumed to be at least ordinal in level of measurement. On the other hand, choice data are responses that identify one and only one of a set of alternatives as the "highest," "best," etc. or responses which involve the allocation of fixed sets of resources to a set of alternatives so the individual allocations sum to the total fixed amount (Louviere, 1988a). Choice data need only be measured at a nominal level of measurement.

The distinction between judgment and choice is important to note, since judgmental data may not contain sufficient information to explain and/or predict choice behavior. This is true for all types of judgmental data - rank-order, rating scale, indifference (also referred to as matching), and pricing. Rank-order judgmental data are collected when the respondent ranks the various alternatives in order of increasing or decreasing preference. Such data are usually collected when the scaling method of paired comparisons is used. Rating scale judgmental data are collected when the respondent rates the various alternatives on some scale such as a visual analog scale. Such data are usually collected when the scaling methods of rating scales and magnitude estimation are used. Indifference judgmental data are collected when the respondent indicates indifference between two
alternatives. Such data are usually collected when the gamble, time tradeoff and equivalence (or person trade-off) scaling methods are used. Finally, pricing judgmental data are collected when the respondent sets a price on the alternatives under evaluation -- either in terms of willingness to pay or willingness to accept. Such data are usually collected when the scaling method of contingent valuation is used. (Each of these scaling methods will be discussed in some detail in Chapter III) By primarily using only judgments as the response mode, the published studies on health status preference measurement have (at least implicitly) assumed "procedural invariance."

The Assumption of Procedural Invariance

The assumption of procedural invariance simply means that different response modes used for measuring health status preferences should end up ranking the alternative health states under evaluation in the same order. In other words, the preferential ranking of alternative health states under evaluation should not depend on the particular procedure used to measure health status preferences. Thus, it is assumed that individuals have a well-defined preference order (or, more specifically, a well-defined preference function) which does not change according to the response mode used to measure preferences. Therefore, different response modes for measuring these preferences should produce the same ordering of alternatives. Indeed, most theories of measurement usually require the ordering of alternatives to be independent of the particular response mode of assessment.
It should be noted that different response modes may be used to operationalize different scaling methods which in turn may be based on different theoretical rationales, and therefore may not give rise to the same preference score. This is NOT contradictory to the assumption of procedural invariance -- the assumption of procedural invariance is only with respect to the ordering of preferences measured with different response modes and does not pertain to the specific preference scores obtained.

The questions that arise, then, are: (a) does the assumption of procedural invariance hold in practice? (b) what is the outcome if the assumption of procedural invariance does not hold in practice? (c) what is the exact mechanism by which the breakdown of the assumption of procedural invariance leads to this outcome? and (d) what are the implications for health status measurement if the assumption of procedural invariance is violated in practice? Each of these questions needs to be addressed. In this section, however, only questions (a), (b), and (d) will be addressed; question (c) will be addressed in Chapter IV.

Specifically, this section will begin by noting why the assumption of procedural invariance may not hold in practice. The phenomenon of preference reversal will then be introduced as a possible consequence of the breakdown of the assumption of procedural invariance in practice. Since there have been other plausible explanations forwarded in the published literature for the phenomenon of preference reversal (apart from the breakdown of the assumption of procedural invariance), these other explanations will be reviewed. Comparative studies designed to determine the true cause of preference reversal will also be reviewed in order to
understand the role of the breakdown of the assumption of procedural invariance in the phenomenon of preference reversal. It will be shown that in spite of the other plausible explanations, the breakdown of the assumption of procedural invariance is a major causative factor in the phenomenon of preference reversal. Although this conclusion begs the question of how exactly does the breakdown of the assumption of procedural invariance lead to the phenomenon of preference reversal, the answer to this particular question shall be deferred until Chapter IV, when the published literature on response mode effects will be reviewed. This section will end with a discussion of the implications of the breakdown of the assumption of procedural invariance in general and the phenomenon of preference reversal in particular for the measurement of health status preferences.

Does The Assumption of Procedural Invariance Hold In Practice?

Given the centrality of the assumption of procedural invariance in the measurement of health status preferences, it is important to examine the tenability of this assumption in practice. To reiterate, the assumption of procedural invariance holds that the preferential ordering of alternatives under evaluation should not depend on the particular response mode used as part of the preference measurement methodology. Tversky, Sattath, and Slovic (1988) have noted that the assumption of procedural invariance is likely to hold when respondents have well-articulated preferences and beliefs, as is commonly assumed in the classical theory of decision making, i.e., expected utility theory. However, empirical research on decision making
and human judgment has indicated that people generally do not have well-defined values and beliefs (see, for e.g., Fischhoff, Slovic, and Lichtenstein, 1980; March, 1978). In these situations, observed or stated preferences are not simply read off from some pre-existing preference function present in the respondents' mind -- they are actually constructed during the measurement process. In other words, decision making behavior is contingent upon the specifics of the measurement process (Payne, 1982).

Empirical research in behavioral decision making has supported the contingent nature of decision making behavior, wherein the factors of the measurement process which most greatly affect decision making behavior are the response mode and the framing of alternatives (Payne, 1982; Slovic and Lichtenstein, 1983; Tversky and Kahneman, 1981). Thus, instead of invariance, human decision making behavior has been shown to exhibit procedural (i.e., effects of response mode) as well as descriptive (i.e., effects of framing of alternatives and presentation of information about alternatives) variance (Payne, Bettman, and Johnson, 1992). Indeed, the demonstration of the breakdown of both descriptive as well as procedural invariance in practice has been called the "... most fundamental result of three decades of empirical investigation [in behavioral decision making]" (Shafer, 1986 p.464).

Although the importance of the failure of the assumption of descriptive invariance is undeniable, it is not the focus of this dissertation. This is simply because the implications of the breakdown of the assumption of descriptive invariance on the measurement of health status preferences have been recognized in the health status preference measurement literature.
Indeed, the issue of the framing of alternatives was recognized in the early 1980s (see McNeil, Pauker, Sox. et al., 1982) and there have been several empirical investigations of the effects of framing on the measurement of health status preferences (see O'Connor, Boyd, and Till, 1985). Therefore, for the purpose of this dissertation, the description of each alternative health state was standardized across methodologies. This standardization was achieved by describing each alternative health state in the form of a listing of attribute levels (see Chapter V and Appendix B for details). Since the variable of description of health state was thus controlled, the literature on this issue will not be reviewed here. Instead, the discussion will focus on the assumption of procedural invariance, since it has been relatively unaddressed in the health status preference measurement literature.

Payne, Bettman, and Johnson (1992) have observed that procedural and descriptive invariance in human decision making may fail in practice on account of at least three factors. First, decisions usually involve conflicting values, wherein one must decide how much one values one attribute or dimension relative to another. This is nothing but the issue of the relative importance of dimensions, as discussed earlier. In trying to deal with such conflicts, individuals often adopt different strategies in different situations, potentially leading to variance in preferences. Second, decisions are often complex, containing several alternatives, each described on the basis of multiple attributes or dimensions. Since these problems are simplified by individuals in different ways, failures of invariance may be related to task complexity (Payne, Bettman, and Johnson, 1992). Finally, as noted by
Tversky, Sattath, and Slovic (1988), one might not have well-defined preferences -- i.e., uncertainty in preferences.

The conclusion to be drawn from this empirical evidence is that different response modes may have different effects on the structure, process, and outcome of decision making. More specifically, different response modes may highlight different aspects of the alternatives under evaluation (i.e., effect on structure) and suggest different heuristics to decision making (i.e., effect on process), which in turn may give rise to inconsistent responses (i.e., effect on outcome) -- different ranking of alternatives measured using different measurement methodologies (Tversky, Sattath, and Slovic, 1988). This is known as the phenomenon of preference reversal.

The Phenomenon of Preference Reversal

The first published observation of preference reversal was over a quarter of a century ago in an article by Paul Slovic and Sarah Lichtenstein in 1968 (Slovic and Lichtenstein, 1968), who observed differences between the preferential ordering of gambles as implied by: (a) buying and selling prices for the gambles (i.e., pricing judgments using willingness to pay and willingness to accept judgments as the response mode) on the one hand and (b) attractiveness ratings of gambles (i.e., judgmental ratings as the response mode) and choices between the gambles (i.e., choices as the response mode) on the other hand (Slovic and Lichtenstein, 1968).

In particular, they found that both buying and selling prices of gambles were primarily determined by the payoffs or amounts to be won or
lost, whereas attractiveness ratings of gambles and choices between gambles were primarily influenced by the probabilities of winning and losing. Specifically, when their subjects found a bet attractive, their bid prices correlated predominantly with the amount to win; when they disliked a bet, their bid prices correlated predominantly with the amount to lose. Accordingly, they hypothesized that, if the method used to elicit preferences affected the weighting of the gambles' components, it should be possible to construct pairs of gambles such that the same individual would choose one member of the pair but set a higher price for the other (Slovic and Lichtenstein, 1968).

Lichtenstein and Slovic (1971, 1973) demonstrated exactly such reversals in a series of studies designed specifically to test the above hypothesis, in both the laboratory (Lichtenstein and Slovic, 1971) as well as a real-world setting (Lichtenstein and Slovic, 1973). In their first study, Lichtenstein and Slovic (1971) conducted three experiments using the same general framework: subjects (university students in all experiments) were presented with a number of pairs of bets. All of the bets had positive expected value and were viewed by subjects as bets they would like to play. Every pair was composed of two bets with the same (or nearly the same) expected value -- a 'P bet' (i.e., a bet with a high probability of winning a modest amount and a low probability of losing an even more modest amount) and a '$ bet' (i.e., a bet with a modest probability of winning a large amount and a large probability of losing a modest amount). For each pair, subjects indicated which bet they would choose (i.e., choice as a response mode). After subjects made their choices, they made a bid (using willingness
to accept in two experiments and willingness to pay in one experiment) for each of the bets which were presented one at a time. The results strongly supported their hypothesis -- in all three experiments, subjects frequently chose one bet from the pair and subsequently bid more for the bet they did not choose. The frequency of such reversals varied by experiment, but was always greater than could be explained by unreliability alone (Lichtenstein and Slovic, 1971).

This study was significant in that it was one of the first to empirically test for the preference reversal phenomenon. However, a few limitations of this study need to be noted. First, it did not control for the order in which response modes were presented to respondents -- choices were always presented before pricing judgments. It is possible that the act of making a choice between bets influenced the information-processing strategies of respondents as they decided on selling or buying prices for the bets. Second, the bets in all three experiments had positive expected values. In order to establish the phenomenon of preference reversal, it is necessary to replicate these findings with bets with negative expected values. Finally, the experiments were conducted in a laboratory setting, and it is possible that the subjects exhibited 'hypothetical bias' or did not take the experiment seriously.

Lindman (1971) published a very similar study about the same time, with exactly the same results. In Lindman's (1971) study also, subjects when choosing among gambles, tended to prefer gambles with high probabilities of winning but gave higher selling prices to gambles with small probabilities of winning larger amounts. A noteworthy point of Lindman's (1971) study is
that it used gambles with both, positive as well as negative expected values. On the other hand, it suffered from a few drawbacks. Similar to the Lichtenstein and Slovic (1971) study, it was conducted in a laboratory setting, and was therefore also susceptible to 'hypothetical bias.' Moreover, it did not control for any order effects -- on this point it should be noted that the order of presentation of response modes was exactly opposite that of the Lichtenstein and Slovic (1971) study, i.e., pricing judgments for willingness to accept were elicited first, followed by choice tasks. This switched order did not have an effect on the existence of the phenomenon of preference reversal.

Although these two studies demonstrated the existence of the phenomenon of preference reversal in the laboratory setting, the question of external validation of the phenomenon of preference reversal still remained unresolved. In this regard, the notion of personal relevance was important -- if respondents did not feel it was worthwhile for them to 'behave optimally,' they might be inconsistent in their responses to judgmental and choice tasks. Lichtenstein and Slovic's (1973) second study was designed to test whether the phenomenon of preference reversal which was observed under laboratory conditions could be replicated in a real-world setting. To this end, they replicated their 1971 study in a Las Vegas casino with real money at stake. The results remained unchanged. Lichtenstein and Slovic concluded, "The widespread belief that decision makers can behave optimally when it is worthwhile for them to do so gains no support from this study. The source of the observed information-processing bias appears to be cognitive, not motivational" (Lichtenstein and Slovic, 1973 p.20).
Lichtenstein and Slovic (1971, 1973) explained these results on the basis of differences in information-processing procedures brought about by different response modes. They attributed the pricing judgment results to a starting point (anchoring) and adjustment procedure -- subjects setting a price on an attractive gamble appeared to start with the amount to win and adjust it downward to take into account the probability of winning and losing as well as the amount that could be lost. Lichtenstein and Slovic (1971) conjectured that the adjustment process was relatively imprecise on account of human information-processing constraints -- a conjecture which was subsequently supported empirically by Slovic (1972) and Tversky and Kahneman (1974) -- leaving the pricing judgments greatly influenced by the starting point payoff. Therefore, they reasoned that using pricing judgments as a response mode prompted an alternative-based evaluation procedure by respondents. However, using choice as a response mode prompted a dimension-based evaluation procedure by respondents, wherein each dimension or attribute of an alternative is compared with the same dimension or attribute of the other alternatives. Such a dimension-based evaluation procedure was previously suggested by Tversky (1969), and forms the basis for his proposed theory of choice known as Elimination by Aspects (Tversky, 1972).

These seminal studies formed what Tversky, Slovic, and Kahneman (1990) have classified as the first wave of preference reversal studies to be published in the literature. Taken together, this first wave of studies introduced the phenomenon of preference reversal into the published
literature, demonstrated its existence under a variety of conditions, and provided a rationale for its existence.

The second wave of preference reversal studies, conducted during the late 1970s and the early 1980s, consisted of critical replications of these early studies and attempts to eliminate preference reversals by procedural variations and/or increased incentives (Tversky, Slovic, and Kahneman, 1990). All the published studies which comprise this wave were prompted by a healthy skepticism of the phenomenon of preference reversal and a belief that -- under the right circumstances -- it might disappear. Rather than review each of the studies which make up this second wave of preference reversal studies, attention will only be focused on a few -- the most cited studies. This is because all of these studies were designed for the same purpose (i.e., to replicate and/or eliminate the phenomenon of preference reversal) and all reached the same conclusion (i.e., the phenomenon of preference reversal is robust and cannot be eliminated by a wide variety of incentives).

Perhaps the most famous of these studies was the one by Grether and Plott (1979), who conducted a series of experiments "... designed to discredit the psychologists' work as applied to economics" (Grether and Plott, 1979 p.623). Approaching the phenomenon of preference reversal from the viewpoint of economists who were skeptical of the relevance of the significance of the first wave of studies to economics and economists, Grether and Plott (1979) investigated whether the phenomenon of preference reversal existed in "... situations where economic theory is generally
applied" (p.624) and whether it could be accommodated by standard economic theory or some immediate extension of standard economic theory.

Grether and Plott (1979) generated a list of objections and potential artifacts that might render the preference reversal phenomenon irrelevant to economic theory or explain the phenomenon in terms of standard economic theory. They classified these objections and artifacts into three broad categories — economic-theoretic hypotheses, psychological-theoretic hypotheses, and experimental method artifacts. Economic-theoretic hypotheses included misspecified incentives for respondents, income effects, and lack of use of indifference judgments as response modes. Psychological-theoretic hypotheses included strategic responses by subjects, changes in subjective probabilities during the experiments, use of noncompensatory decision making strategies by respondents such as elimination by aspects and lexicographic semiordering, decision costs (i.e., use of heuristics in decision making), response mode effects, and justification effects. Experimental method artifacts included respondent confusion and misunderstanding, low frequency of observed phenomenon (of preference reversal) so as to be irrelevant in practice, use of unsophisticated subjects, and the fact that the experimenters were psychologists (Grether and Plott, 1979).

In their experiments on replicating/eliminating the existence of the phenomenon of preference reversal in "... situations where economic theory is generally applied" (Grether and Plott, 1979 p.624), Grether and Plott explicitly controlled for all the economic-theoretic explanations and all the psychological-theoretic explanations but the effect of response mode and
justification. Specifically, they used two monetary incentive systems to heighten motivation, substituted a different probability device of deciding the outcomes of the bets, controlled for income and order effects, and tested for indifference and the influence of strategic or bargaining effects.

Further, in order to avoid respondent confusion and misunderstanding, the instructions were elaborate and included a rationale behind the study, a practice gamble, a demonstration of procedures, and a written test. The correct answers to the test were discussed and subjects questions were addressed. Finally, subjects were allowed to work at their own pace so as to avoid rushing their responses. Regarding their objections to the previous studies on the issues of unsophisticated subjects and psychologists as experimenters, Grether and Plott themselves conducted the experiments (they were economists) and they used subjects from economics and political science classes -- subjects who might be expected to have some training in the principles of decision theory and economics and therefore be 'sophisticated'. Lastly, Grether and Plott let the results of their experiments speak for the frequency with which the phenomenon of preference reversal was observed (Grether and Plott, 1979).

In spite of these careful controls, Grether and Plott were not able to eliminate the phenomenon of preference reversal in their experiments. Indeed, a significant proportion of their subjects exhibited the phenomenon of preference reversal. Since they controlled for all other explanations, they attributed the existence of the phenomenon of preference reversal in their experiments to response mode and justification effects: "The one theory which we cannot reject ... is in many ways the least satisfactory of those
considered since it allows individual choice to depend on the context in which choices are made. For example, if the mode of response or the wording of the question is a primary determinant of choice, then the way to modify [standard economic] theory is not apparent” (Grether and Plott, 1979 p.634; emphasis added). It is noteworthy that Grether and Plott, without explicitly stating it, identified the results of the breakdown of two major assumptions of standard economic theory -- procedural invariance (i.e., mode of response) and descriptive invariance (i.e., wording of the question).

In spite of the rigor exercised by Grether and Plott (1979) in their experiments, some economists were not satisfied with the stringency of the controls. Pommerhne, Schneider, and Zweifel (1982) had three objections to the Grether and Plott (1979) study -- the subjects were not sufficiently motivated to make careful decisions, the incentives provided were not strong enough for subjects to make a motivated and rational choice over the entire duration of the experiments, and the conclusions drawn from a one-shot experiment could not be fully valid. Accordingly, Pommerhne, Schneider, and Zweifel (1982) attempted to increase motivation by raising the face value of the payoffs and creating differences in expected value between the component bets in a pair. Even in the face of these stronger incentives, the phenomenon of preference reversal was still observed in a substantial proportion of the subjects in this study -- although less than that observed by Grether and Plott (1979).

Reilly (1982) was also skeptical of the adequacy of Grether and Plott's (1979) controls. In order to maximize subjects' understanding of the task, he conducted his study within small groups where questions could readily be
asked of the experimenter. Further, the money at stake was placed within eye-view of respondents and the size of potential losses in the gambles was increased to enhance motivation. Finally, some subjects were shown the expected values for all gambles and were given a description of the expected value concept. Although the rate of preference reversals was somewhat lower in this study as compared to the rate observed in the Grether and Plott (1979) study, the phenomenon of preference reversal still persisted to a substantial extent (Reilly, 1982). This is quite remarkable, given the rather elaborate (and quite contrived) controls of this study. Reilly (1982) conceded that these results provided "... further confirmation of preference reversal as a persistent behavioral phenomenon in situations where economic theory is generally applied" (p.582).

Hamm (1979) was another researcher who tried unsuccessfully to make the phenomenon of preference reversal disappear. In his dissertation, Hamm (1979) examined the stability of reversals over time, in the face of experience, practice, forced introspection or discussion, and advice to adopt an intuitive or analytic approach to the task. The order of stimulus sets and tasks was carefully counterbalanced in this study. Hamm (1979) found that the preference reversal phenomenon was replicated under all these conditions. Task order had no effect, nor did emphasis on analytic or intuitive processes. Discussion about one's decision strategies actually increased the tendency toward reversals, countering the hypothesis that if people were given greater opportunity to think about their strategies, the preference reversal phenomenon would disappear (Hamm, 1979).
The existence of the phenomenon of preference reversal was also found in other studies by Mowen and Gentry (1980) who observed preference reversal in both individual as well as group choices concerning product development, by Knez and Smith (1986) who allowed their subjects to trade bets in an experimental market, and by Berg, Dickhaut, and O'Brien (1985) who introduced arbitrage. Slovic and Lichtenstein (1983) aptly summarized the results of the second wave of preference reversal studies taken as a whole: "... the most striking result of these studies is the persistence of preference reversals in the face of determined efforts to minimize or eliminate them" (p.599).

The third wave of published articles about preference reversals were prompted by the observed robustness of the phenomenon, and as a group attempted to provide an explanation for the phenomenon of preference reversal rather than try to attempt to deny its existence. Three classes of models for preference reversal were introduced: (a) nontransitive decision-making models were developed both by Loomes and Sugden (1982, 1983) and Fishburn (1984, 1985), (b) generalized utility models which maintain transitivity but relax the independence axiom were applied to preference reversal by Holt (1986) and Karni and Safra (1987), and (c) response mode models were proposed by Tversky, Sattath, and Slovic (1988) and Goldstein and Einhorn (1987). In essence, these three classes of models attribute the phenomenon of preference reversal to violations of the axioms or assumptions of expected utility theory -- the transitivity axiom, the independence axiom, and the procedural invariance assumption respectively.
The nontransitive decision-making and generalized utility models were attempts by economists to explain the phenomenon of preference reversal as an act of a rational decision maker. In other words, these models were developed for the purpose of explaining behavior which was at variance with that predicted by the expected utility model without sacrificing the tenets of rationality that the expected utility model was built upon and without giving up the assumptions of procedural and descriptive invariance. These models differed with respect to the specific axiom of the expected utility model which they allowed to be violated in order to explain non-expected utility behavior, such as preference reversal -- the transitivity axiom or the independence axiom respectively.

The transitivity axiom simply entails individuals to be perfectly transitive in their ordering of preferences -- i.e., if an individual prefers alternative A to alternative B and alternative B to alternative C, he or she should then prefer alternative A to alternative C. As noted before, the transitivity axiom is one of the key axioms of expected utility theory. Bell (1982), Loomes and Sugden (1982, 1983) and Fishburn (1984, 1985) independently developed theories of rational decision making without requiring the transitivity axiom. In other words, these theories retained the other axioms of expected utility theory but allowed the decision maker to be nontransitive in his or her preferences while still being considered ‘rational.’ Although none of these theories was specifically formulated to explain the phenomenon of preference reversal, they could accommodate the phenomenon. Indeed, in certain cases, these theories could also predict when the phenomenon of preference reversal would occur.
For example, using regret theory (Loomes and Sugden, 1982) as an exemplar of these types of models, the phenomenon of preference reversal between pricing judgments and choice can be reconciled with rational choice theory by taking into account the contribution to subjective utility of alternatives not chosen, i.e., the regret-utility hypothesis. When choice is used as a response mode subjects not only choose a given alternative, they also reject the other alternatives. The parameters of the rejected alternatives may weigh into the calculation of the subjective utility of the chosen alternative (i.e., the I’m-so-glad-I-did-not-pick-that-other-alternative feeling). However, when pricing judgments are used as the response mode, rejected alternatives are absent and therefore may not be expected to enter into the calculation of subjective utility of the chosen alternative. By making certain assumptions about the underlying utilities, Loomes and Sugden (1983) showed that subjects may rationally choose an alternative for which they gave a lower judged price.

As opposed to the transitivity axiom, the generalized utility models relax the independence axiom (also referred to as the substitution axiom) or specific components of the independence axiom of expected utility theory in order to explain various types of observed non-expected utility behavior, including the phenomenon of preference reversal. The independence axiom is an assumption that any compound (i.e., two-stage) lottery is equivalent to a simple lottery with the same final consequences and with probabilities computed from the component probabilities according to the ordinary probability calculus (i.e., in any lottery $L$, the lottery $L_{\text{new}}$ can replace the
consequence A -- provided the two are indifferent to the decision maker -- thus making the original simple lottery a compound lottery).

Holt (1986) and Karni and Safra (1987) developed explanations of preference reversal by relaxing the independence axiom of expected utility theory, thus showing that the phenomenon of preference reversal did not necessarily have to involve the violation of the transitivity axiom (which in their opinion was a more critical axiom for expected utility theory than the independence axiom). It is important to note that neither Holt nor Karni and Safra formally developed generalized utility models themselves -- they only applied the generalized utility models developed by other researchers to explain the phenomenon of preference reversal. [For an example of a generalized utility model without the independence axiom, see Machina (1982)]. Holt (1986) explained the operationalization of how the violation of the independence axiom could lead to preference reversal: "If the independence axiom of expected utility theory is not satisfied, then the lottery choice and the selling price elicitation decisions are no longer separable as would be the case in an expected utility analysis. In this case an individual might value one lottery more highly than another, but choose the less preferred lottery in an experiment in which the consequences of this choice are diluted by the common prospect associated with the elicitation of selling prices. As a consequence, decisions that appear to violate transitivity can result from direct violations of the independence axiom or of some other axiom such as 'reduction of compound lotteries'" (p.514). The reduction of compound lotteries can be considered to be a particular component of the independence axiom (Segal, 1988). It is interesting to note that Segal (1988),
in his explanation of the phenomenon of preference reversal, identified the cause of preference reversal as the violation of the reduction of compound lotteries 'sub-axiom.'

Explaining the reason behind 'sacrificing' the independence axiom instead of the transitivity axiom, Holt (1986) noted: "... to abandon transitivity would be a drastic step that would make it difficult to construct a formal choice theory with empirical content. The transitivity assumption is needed for the existence of a utility functional that represents preferences over lotteries; independence is a strong assumption about the functional form of this utility functional ..." (p.514).

The response mode models, on the other hand, were proposed by psychologists who appeared to be less keen on retaining the framework of the expected utility model in their explanations of the phenomenon of preference reversal. These models generally attributed the phenomenon of preference reversal to the changes in information-processing strategies of respondents brought about as a result of changes in response modes used as part of measurement methodologies. In other words, they explained preference reversal as the result of the failure of procedural invariance. These models essentially built upon the explanations offered by Lichtenstein and Slovic (1971, 1973) about the cause of the phenomenon of preference reversal. Before providing any more details about the response mode models, it is necessary to review the literature on experiments which have directly compared the different proposed explanations for the phenomenon of preference reversal in order to determine which explanation seems most likely to account for preference reversal.
Comparative Studies on the Causes of Preference Reversal

There are only a few published studies which have compared different explanations of preference reversal with respect to experimental data. It was noted that different economists had identified the independence axiom and the reduction of compound lotteries 'sub-axiom' as potential causes of the phenomenon of preference reversal. Unfortunately, most of the experiments conducted in the second wave of studies on preference reversal by economists [e.g., Grether and Plott (1979); Pommerhne, Schneider, and Zweifel (1982); Reilly (1982); Berg, Dickhaut, and O'Brien (1985)] could not differentiate between -- indeed were not designed to differentiate between -- violations of the independence axiom, the reduction of compound lotteries 'sub-axiom,' and the transitivity axiom as potential causes of the phenomenon of preference reversal. This was rather unfortunate, since the different explanations about preference reversal offered by economists [e.g., Loomes and Sugden (1983); Holt (1986); Karni and Safra (1987); Segal (1988)] as part of the third wave of preference reversal studies used these very studies to argue their respective cases. Therefore, each of the potential explanations could have been responsible for the preference reversals observed in these studies.

In recognition of this fact, Cox and Epstein (1989) designed some experiments that did not contain any compound lotteries and did not require the independence axiom to interpret the results. Therefore, any reversals of revealed preferences for lotteries that were observed in their study could not be attributed to violations of the independence axiom and reduction of compound lotteries 'sub-axiom.' Explicitly controlling for wealth, outcome,
and framing effects, they found a substantial number of preference reversals between choices and selling prices, the order of which was randomized across subjects (Cox and Epstein, 1989). The pattern of preference reversals observed in their study did not support the anchoring and adjustment hypothesis forwarded by Lichtenstein and Slovic (1971, 1973). Further, the preference reversals observed in their study were attributed to a violation of the asymmetry axiom which, as Sonnenschein (1971) has noted, is more critical to expected utility theory than even the transitivity axiom (Cox and Epstein, 1989). However, Tversky, Slovic, and Kahneman (1990) have noted, subjects in this study priced each pair of lotteries concurrently (as opposed to separately in most other studies) and then played the higher-priced lottery -- this procedure allowed the subject to generate a price ordering by comparing the two lotteries without performing the more difficult task of assessing their selling prices. Indeed, Cox and Epstein (1989) acknowledge -- on the basis of the subjects' responses -- that their procedure did not elicit proper cash equivalents.

Loomes, Starmer, and Sugden (1989) conducted three experiments to discriminate between the response mode and nontransitive choice models (specifically regret theory -- see Loomes and Sugden, 1982) explanations of the phenomenon of preference reversal. Instead of using the standard experimental design used in preference reversal studies, wherein subjects offer bids (i.e., willingness to pay or willingness to accept judgments) for gambles and make choices between the gambles, these researchers used a "choice-only" experimental design, wherein subjects were made to undergo three pairwise choice tasks. The alternatives under consideration were: (a) a
gambles: (a) a gamble with a high probability of winning a moderate sum of money, (b) a gamble with a small probability of winning a rather large sum of money, and (c) a sure-thing (i.e., 100% chance of winning) of a very small sum of money. It should be noted that the first two alternatives are the same alternatives which have been used in the classic preference reversal studies. These alternatives were placed two at a time in choice sets in order to form the "choice-only" experimental design. The three pairwise choice tasks corresponded to the three combinations of the alternatives, taken two at a time. For each task, subjects were asked to choose the alternative they preferred (Loomes, Starmer, and Sugden, 1989).

As noted previously, these researchers conducted three experiments. The first two experiments required direct choices between alternatives. The purpose of these two experiments was to demonstrate and replicate the phenomenon of preference reversal in the context of a "choice-only" experimental design. In other words, respondents used only choices as the response mode; therefore, if preference reversals were observed, it could not have been to a response mode effect, only due to intransitivity. Further, regret theory (Loomes and Sugden, 1982) predicts that violations of transitivity can occur in only one direction; in contrast, response mode models provide no reason for expecting systematic violations of transitivity in "choice-only" experiments. The results of both experiments strongly supported these hypotheses (Loomes, Starmer, and Sugden, 1989).

In the third experiment reported by Loomes, Starmer, and Sugden (1989), subjects were randomized into matched pairs of subsamples. One subsample from each pair was assigned to the "choice-only" experiment,
while the other was assigned to the standard preference reversal experiment [i.e., asked to provide willingness to accept pricing judgments for (a) and (b)] using the same monetary amounts and probabilities as the (a) and (b) alternatives in the "choice-only" experiment. The purpose of the third experiment was to compare the frequency of observed preference reversals in "choice-only" versus standard preference reversal experimental designs. Since this could not be done by means of a direct comparison of the two subsamples, the researchers made a few assumptions: on the basis of the pricing judgments provided for the (a) and (b) alternatives by the subjects in the standard preference reversal experiment, the researchers imputed the choices that subjects would be expected to make when faced with the prospect of a direct choice between (a) and (c) on the one hand and between (b) and (c) on the other hand (the value of the sure-thing for (c) was the same as that for the "choice-only" experiment). It should be noted that this process rested on two salient assumptions -- subjects would have actually confronted the "choice-only" experiment and, more importantly, would have acted on their valuations provided in their willingness to accept judgments when confronted with the "choice-only" experiment.

Combining these imputed choices with the choices actually made by subjects in the standard preference reversal experiment, the researchers compared the frequency of observed preference reversals in the two subsets of data. Their line of reasoning was as follows: if preference reversals were caused solely by intransitivity of preferences, the method of imputing choices would be legitimate and, therefore, one would expect to find no systematic differences between the number of observed preference reversals
in the imputed choices and the actual choices of the "choice-only" subsample. If, on the other hand, there was a response mode effect, it would not be legitimate to impute choices from pricing judgments and, therefore, if response mode effects were contributing to preference reversal one would expect to find a greater tendency to observe preference reversals in the imputed choices as opposed to the actual choices. The results supported the hypothesis that intransitivity of preferences was the likely cause of preference reversals, since there were no significant differences between the number of observed preference reversals in the imputed choices and the actual choices of subjects (Loomes, Starmer, and Sugden, 1989).

Loomes, Starmer, and Sugden (1989) interpreted the results of their three experiments taken together as supportive of nontransitive choice models in general and regret theory in particular. Although they noted that the response mode effect did not apply to their results, they did not rule out such effects from having a role to play in the causation of the phenomenon of preference reversal (Loomes, Starmer, and Sugden, 1989). Still, it must be noted that this study provided strong evidence of the role of intransitivity in explaining preference reversals.

However, Bostic, Herrnstein, and Luce (1990) reached a different conclusion in their study on the causes of preference reversal. In their study, Bostic, Herrnstein, and Luce (1990) drew a distinction between judged indifference points (JIP) and choice indifference points (CIP). According to them, JIP was assumed to be the judged certainty equivalent of a gamble, while CIP was the boundary -- in money -- above which monetary amounts would be preferred to the gamble and below which the gamble would be
preferred. Thus, the CIP procedure would yield a monetary amount that could be likened to a JIP, but it would be obtained using choice as a response mode instead of some judgmental response (Bostic, Herrnstein, and Luce, 1990).

Bostic, Herrnstein, and Luce (1990) explained the phenomenon of preference reversal on the basis of the distinction between JIP and CIP. They noted that most explanations of preference reversal imply that reversal occur when the JIPs elicited by some procedure using judgment as the response mode do not actually represent their CIPs. At the CIP, both the gamble and the monetary equivalent are equally likely to be chosen -- at least in theory. A pilot study by Pearson (1986) provided empirical support for the proposition that JIPs may differ from CIPs. By empirically determining CIPs and JIPs, the Bostic, Herrnstein, and Luce (1990) study was designed to replicate the results of this pilot study, determine whether using choices as response modes would decrease the incidence of observed preference reversals as compared to using judgments as the response mode, and also to see if transitivity is sustained when indifference points are determined using choice as the response mode instead of using judgment as the response mode.

In order to test their hypotheses, Bostic, Herrnstein, and Luce (1990) conducted two experiments, each using a different method for determining CIPs. The first method was a version of the classical psychophysical up-down method, while the second was called PEST (or Parameter Estimation by Sequential Testing). In both experiments, subjects were also made to provide JIPs in terms of willingness to accept pricing judgments as well as
choose between two gambles. Therefore, in both experiments, each subject provided data relating to CIPs, JIPs, and direct choices between gambles. This allowed tests between JIPs and CIPs (i.e., changes in indifference points due to response mode effects), JIPs and direct choices (i.e., preference reversal due to indifference points determined using judgments as a response mode), and CIPs and direct choices (i.e., preference reversal due to indifference points determined using choices as a response mode).

With both CIP methods, Bostic, Herrnstein, and Luce (1990) found that the frequency of preference reversals was significantly reduced as compared to JIPs, especially with the PEST method of estimating CIPs. Further, subjects who reversed their preferences frequently usually judged the gambles to be equal in value, while those who did not reverse their preferences at all judged them as different. This suggests that random error could have a part to play in preference reversal. Finally, the typical subject exhibited JIP > CIP; this bias appeared to underlie the observed intransitivity when JIP was used (Bostic, Herrnstein, and Luce, 1990).

The results of this study are significant for several reasons. First, with regard to the different explanations for the phenomenon of preference reversal, neither the nontransitive choice models nor the generalized utility models could explain the results of this study. On the other hand, the study results support the response mode models. Specifically, the results are congruent with the compatibility hypothesis (which will be explained later) forwarded by Tversky, Sattath and Slovic (1988) and by Slovic, Griffin, and Tversky (1990). In this regard, the CIP procedures, by using choice as a response mode, are more compatible with the choice phase of the procedure.
as compared to the JIP procedures; therefore the CIP procedures are less likely to introduce systematic inconsistency and, consequently, preference reversal. Moreover, the study pioneered the use of choice-based procedures for assessing indifference points, an idea subsequently taken up by Daniels and Keller (1992).

The final study which empirically compared the different explanations of the preference reversal phenomenon in terms of their ability to explain experimental data which shall be reviewed here is the seminal one by Tversky, Slovic, and Kahneman (1990). These researchers tested the relative contributions of the different plausible explanations of the phenomenon of preference reversal to the true cause of preference reversal. From a design point of view, their study was very progressive. First, they extended the traditional experimental design by adding a sure-thing monetary amount as an alternative in a similar manner as done by Loomes, Starmer, and Sugden (1989). The reason provided for this extension was as follows: if preference reversal was due to an intransitivity, it should be observable in binary choice data without the use of a pricing (or, more generally, judgmental) task. On the other hand, choice could be transitive if preference reversal is due to setting a high price for a lottery with a low probability of winning a large monetary amount [alternative (b) in the description of the Loomes, Starmer, and Sugden, (1989) study] or setting a low price for a lottery with a high probability of winning a modest amount of money [alternative (a) in the description of the Loomes, Starmer, and Sugden, (1989) study]. Therefore, in order to determine whether preference reversal was caused by intransitivity or procedural variance (or non-
invariance, as they termed it), they included the sure-thing monetary amount (Tversky, Slovic, and Kahneman, 1990).

In this study, subjects provided three choices [like the "choice-only" experimental design in the Loomes, Starmer, and Sugden, (1989) study] as well as two pricing judgments for the two lotteries, as described above. Subjects first provided the pricing judgments and then the direct choices; the order of the alternatives in the choice task was randomized. On the basis of this extended experimental design, they developed a new diagnostic procedure which classified all test patterns according to whether they violate the transitivity axiom but not procedural invariance or whether they violate procedural invariance but not the transitivity axiom. The relative frequencies of these patterns were then used to test preference reversal models that assumed procedural invariance but not transitivity (e.g., the regret theory model of Loomes and Sugden, 1982) against response mode models which assumed transitivity but not procedural invariance (e.g., the contingent weighting model of Tversky, Sattath, and Slovic, 1988). In this way, Tversky, Slovic, and Kahneman (1990) differentiated between intransitivity and procedural invariance as possible causes of preference reversal in their experiments.

The results of their experiments showed that 90% of test patterns violated the assumption of procedural invariance, while only 10% of test patterns violated the assumption of transitivity. Tversky, Slovic, and Kahneman (1990) interpreted this to support their contention that the true cause of preference reversal was response mode effects, not the violation of transitivity. However, they were careful to note that this conclusion did not
imply that intransitivity had no role to play in the phenomenon of preference reversal. Indeed, they replicated some of the observed patterns of choice which were reported by Loomes, Starmer, and Sugden (1989).

Having considered two possible causes of the phenomenon of preference reversal, Tversky, Slovic, and Kahneman (1990) turned their attention to the third, i.e., violation of the independence axiom and the reduction of compound lotteries 'sub-axiom' of expected utility theory. Accordingly, they introduced an incentive compatible procedure which they called an ordinal payoff scheme, to provide a test for the role of violations of the independence axiom and the reduction of compound lotteries 'sub-axiom' as the cause(s) of preference reversal. In accordance with this procedure, subjects were first presented with each lottery separately and asked to judge the minimum amount they would be willing to accept in order to sell the lottery. Subjects were then presented with pairs of lotteries and asked to choose the most preferred prospect in each pair. Subjects knew that one of these pairs would be selected at random at the end of the session, and that a random device would determine whether they would play the lottery they chose in the choice task or the lottery they priced higher. Since the prices were used only to order the bets within each pair, choice and pricing judgments should be strategically equivalent -- at least in the normative sense. In such cases, a reversal in preference amounts to an inconsistency wherein the subject expresses a preference for playing the lottery with the high probability of winning a modest monetary amount in the direct choice but a preference for playing the lottery with the low probability of winning a larger amount of money in the pricing judgment. In
order to explain this pattern within a generalized utility model, it is necessary to assume that subjects prefer an even chance to play either of the lotteries over the option of playing each lottery by itself. However, such a mixed strategy could only explain random reversals of preference, not systematic preference reversals. Therefore, the presence of systematic preference reversals in the ordinal payoff scheme -- which cannot be explained by violations of the independence axiom or the reduction of compound lotteries 'sub-axiom' of expected utility theory -- is a test for the falsification of the hypothesis that violations of the independence axiom or the reduction of compound lotteries 'sub-axiom' of expected utility theory can explain the phenomenon of preference reversal (Tversky, Slovic, and Kahneman, 1990).

The results of their experiments indicated a high incidence of preference reversal (45%) obtained using the ordinal payoff scheme. Tversky, Slovic, and Kahneman (1990) interpreted this to mean that their results could not be explained by generalized utility models -- therefore relaxing the independence axiom or the reduction of compound lotteries 'sub-axiom' was neither necessary nor sufficient to account for the phenomenon of preference reversal.

In general, Tversky, Slovic, and Kahneman (1990) explained their results in terms of the scale compatibility hypothesis: the weight of any aspect or attribute or dimension (e.g., probability, payoff) of an object under evaluation is enhanced by compatibility with the response (e.g., choice, pricing). Specifically with regard to their experiments, Tversky, Slovic, and Kahneman (1990) noted that because willingness to accept pricing judgments
were expressed in monetary amounts, compatibility entails that the payoffs (which are expressed in the same units) would be weighted more heavily in pricing judgments that in choice tasks. These results are explained within the purview of the contingent weighting model (Tversky, Sattath, and Slovic, 1988), which is a particular type of response mode model which will be reviewed in Chapter IV of this dissertation. In conclusion, Tversky, Slovic, and Kahneman (1990) noted that the failure of procedural invariance was the most likely cause of preference reversal, at least in their particular experimental condition.

The net result of all these comparative studies on the causes of preference reversal taken together is that although violations of the transitivity axiom of expected utility theory may contribute to preference reversal, the phenomenon of preference is most likely the result of the failure of the assumption of procedural invariance in practice. Therefore, the particular response mode used as a part of the preference measurement methodology can exert a significant influence on the preferential ordering of alternatives under evaluation. Having established this, the next logical question is: “How do response mode effects exert their effects on preference measurement?” However, as noted earlier, this particular question shall be addressed in Chapter IV. At this point, it is necessary to discuss the implications of the above findings for the problem investigated in this study. In other words, the immediate question which needs to be addressed is, “What are the implications of the failure of procedural invariance in general
and the phenomenon of preference reversal in particular for the measurement of health status preferences?"

The Implications of Preference Reversal for Health Status Measurement

Given the dependence of rank ordering of health states on the nature of the response mode used to measure health status preferences, it becomes important to observe the judgment versus choice distinction in health status preference measurement. Unfortunately, the distinction between judgment and choice data has been largely unaddressed in the health status preference measurement literature. This maybe because all the published studies of health status preference measurement have used judgments as the response mode, thus -- at least implicitly -- operating on the "highest-preference-equals-choice" rule, i.e., the alternative which has the highest preference score as inferred from judgmental data will be the one chosen by the respondent.

This "highest-preference-equals-choice" rule is consistent with the tenets of the expected utility model and its variants (Schoemaker, 1982), which have usually formed the theoretical basis for the modeling of health status preferences in the published literature. In other words, as Einhorn and Hogarth (1981) have noted, these models treat judgment and choice as equivalent, i.e., judgment is necessary and sufficient for choice. However, the same authors also pointed out that, "... from a psychological point of view, it might be more accurate to say that while judgment is generally an aid to choice, it is neither necessary nor sufficient for choice. That is, judgments serve to reduce the uncertainty and conflict in choice by processes
of deliberative reasoning and evaluation of evidence. Moreover, taking action engenders its own sources of conflict so that judgment may only take one so far; indeed, at the choice point, judgment can be ignored” (Einhorn and Hogarth, 1981 p.73).

However, Einhorn and Hogarth (1981) continued by emphasizing that the distinction they drew between judgment and choice should not be construed to mean that judgment and choice are unrelated. Indeed, they acknowledge that sometimes judgment and choice are “... inseparable” (Einhorn and Hogarth, 1981 p.73). They conclude their discussion of the judgment versus choice issue by stressing the importance of considering when preferences inferred from judgmental data are similar to preferences inferred from choice data, and when they are different (Einhorn and Hogarth, 1981).

As mentioned previously, this is an area which has been left unexplored in health status preference measurement. Indeed, since choice has never been used as the response mode in published applications of health status preference measurement, the judgment versus choice issue could not possibly have been explored in any of the published studies on health status preference measurement. By using judgment as well as choice as response modes of different methodologies, this study can address this issue in the context of health status preference measurement.

On the basis of the empirical evidence on preference reversal reviewed above, it seems reasonable to conjecture that preferences inferred from judgmental data may rank health states in a different order than preferences inferred from choice data. If this is indeed so, the next question
that arises is which type of response mode is best suited for measuring health status preferences -- judgment or choice. To answer this question, it is necessary to note the relative merits and demerits of judgments and choices as response modes in health status preference measurement methodologies and then revisit the purpose of measuring health status preferences.

**Relative Pros and Cons of Judgments and Choices as Response Modes**

The distinction between judgmental and choice data has already been established. As noted, preferences may be inferred from either judgmental or choice data. However, due to the breakdown of the procedural invariance assumption in practice, it is possible that the preferential ordering of alternatives based on judgmental data might not agree with that based on choice data. Therefore, the next question is, "Which type of data, judgmental or choice, represent people's 'true' preferences?" Unfortunately, in the absence of a gold standard for comparison, one cannot truly know which type is a better representative of a person's true preferences. However, the decision between these two types of data can be made on logical and empirical grounds, by considering the relative strengths and weaknesses of each response mode.

Although the information content contained in a single choice is clearly less than that obtained from a judgment, Daniels and Keller (1992) have noted that a preference model calibrated from choices may still provide a more accurate representation of an individual's preferences. This is because the quality of the elicited choice data may be significantly higher, in
terms of replicability, confidence, and consistency than the corresponding elicited judgmental data. Further while judgmental data usually require respondents to make relatively few evaluations, it is possible that these judgments may be difficult to provide reliably. On the other hand, a choice-based measurement methodology requires a large number of evaluations which are usually easier to elicit from respondents. As Daniels and Keller (1992) have noted, all of these factors -- information content, information quality, and effort required in the elicitation process -- must be considered in selecting an appropriate assessment response mode.

Conclusions regarding the descriptive validity of preference models can also be influenced by the response mode used in the preference measurement methodology. Currim and Sarin (1989, 1992) and Daniels and Keller (1990) evaluated utility models assessed via indifference judgments by counting the number of correct predictions over a holdout sample of choice scenarios. The indifference-based response mode used to calibrate the models in these studies differed fundamentally from the choices used to test the models, thus confounding the evaluation process. Choice-based preference measurement methodologies avoid this problem by using the same response mode for model calibration and evaluation (Daniels and Keller, 1992).

Another important consideration is predictive validity. Daniels and Keller (1992) conducted an empirical comparison of the predictive validity of preference models assessed using judgments versus choices as response modes. Specifically, these researchers compared the predictive performance of the expected utility (EU) theory model and the lottery dependent
expected utility (LDEU) theory model (Becker and Sarin, 1987) as assessed by indifference judgments as well as choices. (LDEU is a type of generalized utility model which is more descriptively valid than the classical EU model). Predictive performance was based on comparisons with a holdout sample of respondents' actual choices between lotteries. It was found that the predictive performance of both, EU and LDEU models improved significantly when choice data were used to calibrate these models as compared to indifference judgmental data (which is usually what is used). The results of this empirical study generally supported the use of choice-based response modes for the measurement of preferences, at least in terms of predictive validity of the preference models assessed using choice versus judgmental data (Daniels and Keller, 1992).

Finally -- but perhaps most importantly -- there is the issue of respondent evaluations. Since it is the respondents who are providing the judgment or choice data, their evaluations of the relative merits and demerits of the different types of response modes must be considered. The reason for this is simple: if the respondents are unfavorably inclined to a particular response mode, the responses they give using that response mode might not be valid and reliable. In other words, respondent evaluations of response modes have implications for the very quality of the data collected using the response modes in question.

Although there are many aspects of a given response mode which can be evaluated by respondents, at a minimum the response mode should be clear, easy to understand, easy to do, hold the respondent's interest for the appropriate period of time, and be sensible to the respondent. Further goals
to aspire toward are success in achieving the intentions of the measurement procedure via the particular response mode as well as confidence of respondents that the response mode used fulfilled its intended purpose. It should be emphasized that these attributes of response modes should be asked of respondents themselves and not inferred from overt respondent behavior via observation.

A study reported by Tversky, Sattath, and Slovic (1988) offers some guidance on the relative confidence the respondents of that study had in their judgmental and choice responses. The subjects provided data via both choice and indifference judgment response modes. The majority of respondents (21 out of 32) gave inconsistent responses that conformed to the prominence hypothesis. After the session, each participant was interviewed and confronted with his or her answers. The subjects were surprised to discover their inconsistencies and proceeded to offer a variety of explanations, most of which corresponded to the prominence hypothesis. When asked to reconsider their answers, most of the respondents modified the judgmental response in the direction of the choice response and only a few did the opposite (Tversky, Sattath, and Slovic, 1988). While it is acknowledged that this study was not expressly designed to test the direction of changes in responses, it does show that the majority of respondents of the study were more confident in their choice-based responses as opposed to their judgmental responses. An indication of the relative strengths of conviction of both types of responses can be seen by the fact that most of the respondents changed their judgmental responses, not their choice responses.
The Purpose of Measuring Health Status Preferences

With regard to the measurement of health status preferences, unlike judgment, choice is an outcome. In light of the discussion in the opening section of Chapter I of this dissertation about outcomes management, it would seem most appropriate to focus on choice as opposed to judgment as a response mode. Further, there is the issue of justification or accountability. As noted earlier, the process of making a choice has an inherent aspect of justification to it, even if that justification be to oneself. As Slovic (1975) has noted, “Compensation rates or trade-off functions derived from matching techniques may be unaffected by considerations of justifiability, which are likely to become relevant only within the context of choice. Thus subjective preference functions derived from matching may imply choices different from those an individual would actually make” (p.287).

This raises the question of the relevance of justification or accountability in measuring health status preferences. As noted in the opening section of the preamble to this dissertation, the US health-care system has been described as currently in an “Era of Assessment and Accountability” (Relman, 1988). Indeed regulatory agencies, managed care organizations, third party payors, and patients are vociferously demanding accountability from health-care providers and pharmaceutical companies. Further, on the matter more relevant to this dissertation, true meaning will be imparted to health status scores only if these scores reflect the demands of the health-care environment -- which is one of accountability. Indeed, if health status scores are to realize their true potential and be used for public policy as well as clinical decision making, they need to be justifiable or
accountable. Therefore, it might be necessary to infer preferences for health states by using a methodology which would incorporate this element of justification or accountability. In other words, in such a case, choice might be preferable as a response mode over judgment for measuring health status preferences.

More importantly, a major reason for measuring health status preferences is to provide true meaning to health status scores. It is necessary to provide meaning to health status scores so that they may be used in making therapeutic decisions which would take the patients’ preferences into account. It is believed that such “patient-centered” therapeutic decisions would lead to better patient outcomes, such as better compliance with therapy, increased patient satisfaction, and improved health-related quality of life.

The salient point in this chain of events is the therapeutic decision. A decision ultimately involves selecting one out of many alternatives as the “best” according to some criterion. In other words, a decision ultimately involves a choice among alternatives, not a judgment about alternatives. Given this line of reasoning, it may be a fallacy to rely on judgmental data to be used for determining choices, especially when choice data can be collected as easily as judgmental data. Indeed, the rationale and methodology for using choices as the response mode for measuring preferences was clearly stated over thirty years ago by Restle (1961):

“Any adequate theory of how motivations combine to produce choice behavior can be used to measure utility. The method is to have the person choose under a variety of conditions, and to record these choices. Then one can hypothesize a certain set of utilities and see whether those utilities, along with the theory
of choice, lead to the choices actually made by the person. The experiment has the appearance of circularity but is not circular. The person is given some restricted set of choices, and from these choices his utilities are estimated, using the theory. When the set of utilities is computed, the theory can be used to predict choices among combinations of alternatives which were not used in measurement. Thus the theory of choosing becomes testable” (p.60).

All this assumes that there exists a valid and reliable methodology for inferring preferences from choice data. Ideally, such a methodology should include the strengths of currently used methodologies for measuring health status preferences based on the expected utility model or its variants (such as mathematical rigor and tractability in estimation procedure) as well as overcome some of the drawbacks of the currently used methodology. Fortunately, such a methodology has been developed -- discrete choice conjoint methodology.

**Discrete Choice Conjoint Methodology**

Although a detailed description of discrete choice conjoint methodology will be deferred until Chapter III, where the different health status preference measurement methodologies are discussed, at this point it is necessary to give a brief overview of this methodology -- especially as it pertains to some of the issues raised in this chapter.

Discrete choice conjoint methodology represents the fusion of preference measurement methodologies from two separate fields -- discrete choice modeling from transportation economics and conjoint measurement/analysis from mathematical psychology/marketing...
respectively. It uses a statistically inferred measurement strategy (i.e., conjoint measurement/analysis) and a scaling approach which emphasizes choice as the response mode (i.e., discrete choice modeling).

The distinguishing characteristic of discrete choice modeling is the proposition that the total utility (or value, depending on whether the preference measurement methodology incorporated the notion of risk or not) of any alternative is composed of two components -- a systematic component and a random component. The random component is very important to discrete choice modeling, because this is the component that differentiates it from the classical expected utility model (which has been -- and still is -- the dominant preference model used in decision analysis). In spite of its mathematical elegance, rigor, and popularity, the expected utility model lacks an error theory. This means that it assumes that the preference scores obtained as a result of the application of the expected utility model are exactly representative of the respondent's preferences and do not contain the slightest amount of systematic or random error. However, this assumption seldom holds in practice. Although there have been some efforts over the course of the past decade to incorporate random error into the expected utility model (see, for e.g., Eliashberg and Hauser, 1985; Laskey and Fischer, 1987), these are still in the nascent stage of their development. More importantly, from a theoretical point of view, such efforts may contradict some of the very basic assumptions and axioms upon which the expected utility model was built. In other words, such efforts to make the expected utility more descriptive do so at the expense of its normative appeal.
By including the random component, discrete choice modeling does not involve the restrictive assumption that individuals' preferences are well-articulated (unlike the expected utility model and its variants), and thus allows for the existence of preferences which are neither stable, consistent, nor precise. Such preferences could, then, change according to the specifics of the question(s) asked of respondents (i.e., the particular question frame) as well as the task required of respondents (i.e., judgment or choice). In other words, discrete choice modeling allows for the existence of the phenomenon of preference reversal. It is important to note that this strength of discrete choice modeling does not come at the expense of the benefit of analytical tractability, because, by assuming a particular distribution of the error term, it is possible to specify different choice models (such as logit, probit, dogit, etc.). These different choice models are necessary to estimate the parameters of the preference function of individuals.

On the other hand, the distinguishing characteristic of conjoint methodology is the simultaneous manipulation of multiple levels of attributes or dimensions, usually in a factorial or fractional factorial design. Such designs, being experimental designs, allow cause-and-effect conclusions to be drawn about the manipulated factors. In this way, it is possible to rigorously test for particular specifications of individuals' or group preference functions (with the help of the necessary statistical estimation techniques, like logit, probit, analysis of variance, or multiple regression analysis) -- thus it is possible to test for the presence of interactions among health status dimensions using such a methodology.
In summary, the discrete choice conjoint methodology affords the opportunity to test for the presence of interactions among health status dimensions and uses choice as the response mode. At least in theory, it affords a means to deal with the preference reversal phenomenon identified earlier. Further, it affords a good balance between mathematical elegance and analytical tractability (this advantage will be fully explained in Chapter III). On the other hand, a drawback of the methodology is that the current state of the art does not allow for the explicit inclusion of risk in the measurement of preferences using such a methodology. Moreover, such models are usually estimated only at the aggregate level, since the parameter estimates of choice models are usually biased and unstable at the individual level.

Further, this methodology is rather new, having been developed a little over ten years ago. No published applications of such a methodology to the measurement of health status preferences were found in the literature which was reviewed in designing this study (see Bibliography). Thus, it is in need of validation as a measure of health status preferences. Validation, of course, is an ongoing process. This study provides one aspect of validation for discrete choice conjoint methodology as a measure of health status preferences.

Validation of Discrete Choice Conjoint Methodology

To paraphrase Thorndike and Hagen (1969), a measure is said to be valid if it measures what it is intended to measure, all of what it is intended
to measure, and nothing but what it is intended to measure. Nunnally has observed that a measure is valid only for "... some specifiable functions with specific groups under specific conditions" (Nunnally, 1967). This suggests that validity is associated with the data collected, not the instrumentation per se. Therefore, one tests the validity of the data collected by the measurement methodology, not the measurement methodology itself. However, if the data collected are indeed found to be valid, a strong case may be made for the validity of the measurement methodology itself -- at least in a context similar to that of the present study. Validity is generally thought to be of three types -- content, criterion, and construct. Of these three, construct validation is the most comprehensive, and some measurement experts view it as encompassing the other two types of validation (Froberg and Kane, 1989b).

Content validity refers to the representativeness of the content of the measure as related to the entire domain of the content desired to be represented by the measure. Applied to health status measurement, content validity refers to the adequacy of the health status descriptions in representing health status. It is usually achieved by careful selection of health status dimensions or attributes and detailed instructions to respondents for the specifics of the scaling approach. Regarding the selection of health status dimensions, this study conceptualized and operationalized health status using the EuroQol. As noted earlier, the EuroQol is a standardized health status measurement instrument. Content validity was assessed in this study by means of a panel of experts.

Strictly speaking, criterion validity does not apply to the measurement of health status preferences. This is because there is no single
criterion embodying an individual's "true preferences." Moreover, there are no external objective measures which are in perfect agreement with health status preferences. However, Torrance has noted that health status preferences measured with the standard gamble scaling approach be viewed as the "gold standard" or criterion according to which other scaling approaches should be evaluated (Torrance, 1976a). The basis for his argument is that the standard gamble is valid by definition since it is based directly on the intuitively appealing axioms of expected utility theory. However, on account of the drawbacks of expected utility theory (as noted previously), it was decided not to consider the standard gamble as the "gold standard" for this study.

In this study, the discrete choice conjoint methodology was validated using construct validation. This was because in measuring health status preferences, one is concerned with an abstract variable or "construct" rather than an observable variable. Peter (1981) has defined a construct as a term specifically designed for a specific scientific purpose, generally to organize knowledge and direct research in an attempt to describe or explain some aspect of nature. Constructs have at least two types of meaning -- systemic and observational (Kaplan, 1964). Systemic meaning refers to the fact that the interpretation of what a construct stands for depends upon the theory in which the construct is encountered (Peter, 1981). On the other hand, observational meaning refers to the notion that a construct must be capable of being directly or indirectly operationalized if it is to have explanatory power (Torgerson, 1958). Therefore, as Peter (1981) has noted, if a construct has no observational meaning, it is merely a metaphysical term, and if a
notion has no systemic meaning, it is not a construct but an observational term. The term “construct validity” is generally used to refer to the vertical correspondence between a construct which is at an unobservable, conceptual level and a purported measure of it which is at an operational level (Peter, 1981).

As noted by Miller, Kaplan, and Edwards (1967), “The basic idea of construct validity is that a test should make sense and the data obtained by means of it should make sense. One form of making sense is that different procedures purporting to measure the same abstract quantity should covary” p.367). It should be noted that this is only one of several ways in which construct validity can be assessed. In the construct validation of health status preference measures, mainly two approaches have been taken: (a) examining the degree to which the results of different methodologies converge, and (b) examining the degree to which predicted relationships between preferences and other variables are empirically supported (Froberg and Kane, 1989b).

This study assessed construct validity by means of convergence of the health status preferences obtained as a result of the discrete choice conjoint methodology with the preference scores obtained by two other commonly used methodologies in health status preference research -- a conditional utility function-based strategy/standard gamble scaling approach and a conditional utility function-based strategy/rating scale scaling approach.

The choice of these two particular methodologies reflects their popularity in the published literature on health status preference measurement. Moreover, these particular methodologies provide interesting
theoretical contrasts to discrete choice conjoint methodology. In this regard, the discrete choice conjoint methodology is a statistically inferred (i.e., top-down) measurement strategy with choice as the response mode. On the other hand, the conditional utility function-based measurement strategy is a bottom-up strategy. This brings up the issue of the difference between statistically derived preference scores versus subjectively estimated preference scores. This is an area which is receiving increasingly greater attention recently. The published literature suggests that -- all else being equal -- the statistically inferred measurement strategies would result in higher and more peaked preference scores being given to alternatives as compared to the subjectively estimated measurement strategies. The difference has been attributed to the fact that since statistically inferred strategies evaluate holistic alternatives, they are better able to capitalize on any interactions among the dimensions or attributes which describe the alternatives (see Green and Srinivasan, 1978, 1990).

Further, both the rating scale as well as standard gamble scaling approaches use judgments as the response mode. This stands in contrast to the choice-based response mode used by discrete choice conjoint methodology. Hence, it would be possible to test for the effects of the judgment versus choice distinction, which was discussed earlier.

Moreover, there is the issue of risk. The standard gamble scaling approach explicitly considers risk in evaluation. None of the other scaling approaches do so -- in other words, the standard gamble is the only scaling approach which measures utilities, as opposed to values. Therefore, it would be possible to model the risk attitude of respondents on the basis of the
estimation of a functional relationship between health status preference scores obtained using the standard gamble and the other scaling approaches used in this study.

Finally, the selection of these particular methodologies is also related to the issue of replication, i.e., the relationship between rating scale and standard gamble scaling approaches has been theoretically and empirically established, not only in the decision analysis literature but also in the health status preference measurement literature (see, for e.g., Barron, von Winterfeldt, and Fischer, 1984; Torrance, Zhang, Feeny, et al., 1992), as reviewed in Chapter IV. Therefore, replicating the results of previous published studies comparing these two methodologies would give one additional confidence that one is measuring what one intends to measure, i.e., health status preferences. In other words, this replication will serve to establish the health status preference measurement measures against which the discrete choice conjoint methodology will be validated. Then, if can be shown that the measures obtained as a result of discrete choice conjoint methodology converge with these “standard” measures, the construct validity of this new methodology as a measure of health status preferences would be supported.

Issues to Consider in Comparing Preference Measurement Methodologies

Given that the construct validity of discrete choice conjoint methodology in this dissertation will be investigated by comparing the measures (i.e., preference scores) produced by it to the measures obtained
using two other preference measurement methodologies, the question that arises is what are the issues which need to be considered in such comparative analyses of preference measurement methodologies. This section will briefly discuss the issues which were considered important for the purpose of this dissertation, with the view of setting the stage for the section of this chapter which presents the specific research questions which were investigated in this study (which follows the section after this one).

Before actually comparing preference measurement methodologies, it is useful to evaluate the accuracy of each methodology as a measure of the construct it is intended to measure. An often used estimate of accuracy is predictive ability or predictive validity, which refers to the accuracy with which measures obtained as a result of a methodology represent the actual scores of the variables under measurement. However, since there are no actual scores of health status preferences, it is necessary to modify the approach of predictive ability in its application to health status (or generally any type of) preference measurement. For this purpose, the sample is divided into two subsections, called the estimation and the holdout samples. The estimation sample is used for modeling purposes, in order to determine the model form or combination rule of the preference function for health status (or whatever construct for which preferences are being measured). Preference scores are then calculated for the alternatives in the holdout sample on the basis of the model form which was determined from the estimation sample, and compared to the observed preference scores of the alternatives in the holdout sample. The mean absolute difference between the two sets of scores gives an estimate of the accuracy of the preference
measurement methodology under evaluation. In this way, it is possible to evaluate the accuracy of each preference measurement methodology as a measure of health status preferences.

Once the accuracy of each methodology as a measure of the construct it is intended to measure is established, the issue of comparing the methodologies can be addressed. The logic behind this two-step procedure to comparison is as follows -- first make sure that each methodology accurately measures a particular construct; then determine the extent to which the different methodologies are measuring the same construct. Since each methodology is intended to measure health status preferences, if there is significant convergence among the measures obtained using the different methodologies, they all must be measuring the same construct, i.e., health status preferences.

The first aspect on which to compare the different preference measurement methodologies is the order in which they rank the alternatives (i.e., health states) under evaluation. This is necessary to investigate the occurrence of the phenomenon of preference reversal, which was explained earlier in this chapter.

Further, in comparing different measurement methodologies, it is important to differentiate between correlation and concordance of measures. A correlation coefficient is a measure of linear association between two variables. On the other hand, a concordance coefficient measures the extent of agreement between two or more variables. The distinction between the two types of coefficients is important because correlation coefficients do not provide any information about agreement between variables. For instance,
one may observe a correlation coefficient of nearly 1.0 (i.e., nearly perfect linear association) when one measure is approximately twice the other, but actual agreement in this case is non-existent. Good agreement is only obtained when the pairs of readings closely follow the line of equality. Further, and more importantly, correlation coefficients totally ignore systematic bias — in other words, if there is theoretical justification for two measurement methodologies to give separate preference scores for health states, correlation coefficients do not take this into account in the calculation, while concordance coefficients indicate the strength of agreement between the methodologies in recognition of the theoretical differences between them.

Since the different preference measurement methodologies are all intended to measure the same construct (i.e., health state preferences), they should be positively correlated with one another. However, on account of their inherent differences (e.g., value versus utility, judgment versus choice), they may be expected to give significantly different preference scores for the same state of health. Therefore, the measures or preference scores obtained as a result of the different preference measurement methodologies also need to be compared, in order to determine whether they are equal to one another or if there are significant differences among them.

However, even though the measure or preference scores obtained using the different preference measurement methodologies may be significantly different (on account of their inherent theoretical differences), it would still be expected that they would be functionally related. In other words, it is expected there would be some type of functional relationship
between each of them, taken two at a time, defined as the expression of one score as a mathematical function of another. The logic behind this expectation is that since the different methodologies measure the same construct (i.e., preferences), the scores obtained by one methodology will be a function of those obtained by another. It should, however, be noted that this function need not be linear. The specific nature of the functional relationship would be dictated by theoretical considerations. Intuitively, it is useful to think of functional relationships as a sort of conversion formula to convert scores obtained using one particular methodology to those that would be obtained using another methodology. Therefore, the establishment of a functional relationship comes closest among all the bases of comparison discussed in this section to the notion of true convergence (i.e., the same scores) among different measurement methodologies intended to measure the same construct; thus, it provides a stringent test of the convergent validation approach to construct validation.

To summarize, the relevant questions to ask when comparing health status preference measurement methodologies include: (a) do the methodologies agree in terms of the order in which they rank alternative states of health? (b) do the methodologies give significantly different preference scores for the same health states? (c) is there a significant correlation between the measures or preference scores obtained as a result of using them? (d) what is the degree of concordance between the preference scores obtained as a result of using them? (e) what is the nature of the functional relationship between the preference functions estimated as a result of using them?
Apart from these questions, it is also necessary to consider how the subjects who are used to provide responses evaluate each of the preference measurement methodologies. This is mainly because of two reasons: (a) different preference measurement methodologies are based on different theoretical principles; therefore, there may not be any single "correct" measurement methodology. In this case, as long as each methodology is shown to be able to provide reliable and valid data, the decision as to which one to use in any particular situation should consider the perspective of the subjects who will be made to perform the tasks necessary to provide data for the measurement methodologies, and (b) respondent evaluations may be indicative of the degree to which each preference measurement methodology poses an undue burden on respondents. Respondent burden is an important aspect to consider in evaluating preference measurement methodologies, since it has direct implications for the validity of the data collected -- if the respondent is unduly burdened by the task requirements of the methodology, the quality of the data collected may suffer. Therefore, when comparing different preference measurement methodologies, it is important to consider respondents' evaluations of each of the methodologies.

**Statement of the Problem**

Given the need to measure patient preferences for health states, it is imperative that such measurements are based on sound methodological principles. What is needed is a health status measurement methodology which is feasible for routine clinical and population studies, has the
capability of incorporating interactions between health status dimensions, and does not require making untestable assumptions about the relationship between respondents' stated preferences and subsequent choices. Further, the preferences so obtained should at least be measured on an interval scale in order to permit the use of various statistical techniques and pharmacoeconomic analyses. Moreover, the effects of risk and time (including prognosis) on the preferences for health status need to be addressed in the scaling approach used for measuring health status preferences. Unfortunately, a review of the published literature (see Bibliography) reveals that the methodology for collectively addressing all these issues in health status preference measurement has not been developed.

However, the discrete choice conjoint methodology offers significant potential advantages over existing methodologies for measuring health status preferences -- at least in principle. As noted earlier in this chapter, there have been no published applications of this methodology to the measurement of health status preferences. The broad objective of this study was to develop and validate discrete choice conjoint methodology as a measure of health status preferences. As noted previously, the discrete choice conjoint methodology was validated against two commonly used health status preference measurement methodologies, as summarized below.

Three methodologies for measuring health status preferences were investigated in this study: (a) a conditional utility function-based measurement strategy along with a rating scale scaling approach (i.e., bottom-up strategy, measures value with judgmental data) -- denoted as
CUF-RS, (b) a conditional utility function-based measurement strategy along with a standard gamble scaling approach (i.e., bottom-up strategy, measures utility with judgmental data) -- denoted as CUF-SG, and (c) a statistically inferred measurement strategy along with a choice-based scaling approach (i.e., top-down strategy, measures value with choice data) -- denoted as DCCM.

Research Questions

This section presents the specific research questions that were investigated as part of this dissertation. For each of the questions, a short discussion on the rationale behind the question, the specific hypothesis used to test the question, and the analytical procedure used for the purpose of hypothesis testing are provided. The questions are organized into three categories: (a) modeling the data obtained using the different preference measurement methodologies investigated in this study, (b) comparing the preference scores obtained using the different preference measurement methodologies investigated in this study, and (c) comparing the respondent evaluations of the different preference measurement methodologies investigated in this study.

(a) Modeling the Data

The first six research questions that were investigated in this dissertation pertain to the modeling of the data obtained using the three different preference measurement methodologies. As explained earlier in
this chapter, modeling the data obtained using the different preference measurement methodologies is a necessary preliminary step before comparing them. Specifically, there were two aspects of modeling the data which were investigated in this study -- the model form or combination rule of the preference function and the predictive ability of the preference measurement methodology.

The need for accounting for interactions among health status dimensions in measuring health status preferences was discussed earlier in this chapter. The existence of interactions among health status dimensions is reflected in the specific model form or combination rule of the preference function for health status. As will be explained in Chapter III, both the conditional utility function-based procedure and the discrete choice conjoint measurement strategies allow for the testing of different model forms, and therefore can be used to verify the existence of interactions among health status dimensions.

The EuroQol instrument was used to operationalize health status in this study. As discussed earlier in this chapter, previous empirical research has supported the existence of a multiplicative model form for the EuroQol health status preference function (Dolan and Kind, 1994; van Hout and McDonnell, 1992). However, before forwarding any hypotheses about the nature of the model form determined by the preference measurement methodologies investigated in this dissertation, it is instructive to consider the results of previous research on modeling health status preferences using these methodologies. This is because according to change of process theory (Mellers, Ordonez, and Birnbaum, 1992) -- which is reviewed in Chapter IV -
- model forms of preference functions may change as a function of the measurement methodology used to measure preferences.

The published literature empirically supports the existence of a multiplicative model form for health status preference functions as determined by the conditional utility function based procedure with a rating scale scaling method or CUF-RS (for e.g., see Torrance, Boyle and Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992). Therefore, it was hypothesized that the model form of the preference function as determined by this particular methodology would be multiplicative.

Although Torrance, Zhang, Feeny, et al. (1992) reported a multiplicative model form for health status preferences determined using the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG, this conclusion was not based on data which were collected using this particular preference measurement methodology. These researchers empirically determined the model form of their respondents' health status preference functions using the CUF-RS methodology, and converted the values obtained using the CUF-RS methodology into utilities that would be obtained if the CUF-SG methodology were used (by means of a power curve transformation). The implicit assumption these researchers operated on was that the model form of health status preference functions would remain invariant across preference measurement methodologies. However, change of process theory (Mellers, Ordonez, and Birnbaum, 1992) posits that the model forms of preference functions may change as a function of the measurement methodology used to measure preferences. Therefore, it was decided to not
forward any specific research hypothesis for this question. Along the same lines, since there have been no published applications of discrete choice conjoint methodology or DCCM to the measurement of health status preferences, no specific research hypothesis was forwarded for the model form of the preference function determined by this particular methodology.

The second aspect of modeling the data which was investigated in this study involved the predictive ability of the preference measurement methodologies. As explained earlier in this chapter, this was necessary to evaluate the accuracy of each preference measurement methodology as a measure of health status preferences. The statistic used to evaluate the predictive accuracy of each methodology was the mean absolute residual (i.e., the mean of the absolute difference between the observed score of an alternative and its score predicted by the preference function determined using the model form estimated in the first aspect of the modeling procedure). On account of the paucity of published literature on the predictive ability of these methodologies as applied to the measurement of health status preferences, no specific research hypotheses were formulated for these set of questions.

Since no inferential statistics were computed for these six research questions, there were no statistical null hypotheses. Regarding the particular analytical procedures used to address each of these research questions, the determination of the model form of the health status preference function followed the specifics of each measurement strategy (as explained in Chapters III and V), while the estimation of the predictive ability of each methodology involved the calculation of the mean absolute residual for the
holdout samples in each methodology. A detailed description of the analytical procedures is provided in Chapter V.

The first two research questions pertain to the modeling of the data obtained using the conditional utility function-based procedure with the rating scale scaling method (CUF-RS) methodology. They can be stated as:

**Research Question #1:** What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?

**Research Question #2:** What is the predictive ability of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology as a measure of health status preferences, as indicated by the mean absolute residual?

The next two research questions pertain to the modeling of the data obtained using the conditional utility function-based procedure with the standard gamble scaling method (CUF-SG) methodology. They can be stated as:
Research Question # 3: What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Research Question # 4: What is the predictive ability of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology as a measure of health status preferences, as indicated by the mean absolute residual?

The last two research questions in this category pertain to the modeling of the data obtained using the discrete choice conjoint methodology (DCCM). They can be stated as:

Research Question # 5: What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the discrete choice conjoint methodology or DCCM?

Research Question # 6: What is the predictive ability of the discrete choice conjoint methodology or DCCM as a measure of health status preferences, as indicated by the mean absolute residual?
(b) Comparing Preference Scores

The next category of research questions was concerned with comparing the preference scores of multiattribute health states obtained using the different preference measurement methodologies investigated in this study in terms of the different bases of comparison explained earlier in this chapter, i.e., preferential ordering, magnitude of preference scores, correlations, concordances, and functional relationships. Each of these questions will be addressed in turn in this section.

(i) Preferential Ordering

The first aspect on which the different preference measurement methodologies were compared was in terms of the order in which they ranked the alternative health states under evaluation. Since each of the three methodologies measures the same construct, from a normative point of view they should agree in terms of the relative order in which they rank states of health. This is, of course, the assumption of procedural invariance. However, from a descriptive point of view, on account of possible response mode effects, it is possible that there might not be perfect agreement among the methodologies in terms of the order in which they rank alternative states of health. Therefore, the particular research questions investigated as part of this aspect of the comparison among the methodologies were intended to investigate the occurrence of the phenomenon of preference reversal.

Since these research questions required only ordinal-level information, they were addressed using two different types of statistical measures of ordinal agreement -- Kendall's Coefficient of Concordance (or
Kendall's $W$ and Kendall's Tau (or Kendall's $\tau$). The difference between these two statistics is in the number of measures among which they can describe agreement -- Kendall's $W$ describes the relationship among three or more measures simultaneously using a single number, while Kendall's $\tau$ describes the relationship between only two measures simultaneously using a single number. It was decided to use both statistics to address this particular research question in order to give an overall indication of agreement in preferential ordering among all three preference measurement methodologies (i.e., Kendall's $W$) as well as to provide more specific indications of agreement between each of the preference measurement methodologies taken two at a time (i.e., Kendall's $\tau$). The specific research questions and hypotheses tested to address each particular research question will be presented first for the case of all three preference measurement methodologies taken together, and then for the preference measurement methodologies taken two at a time.

**Research Question # 7:** Do the three health status preference measurement methodologies (i.e., the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) agree in terms of the order in which they rank alternative states of health?
As noted above, this question was addressed using Kendall’s Coefficient of Concordance or Kendall’s W. Kendall’s W measures the extent of agreement in the rank ordering of a variable measured by three or more different measures. It is bounded between 0 and 1, with 0 indicating no agreement among the three measures and 1 indicating perfect agreement among the three measures (since there is no such thing as perfect disagreement with more than two measures, Kendall’s W cannot take on a value of less than 0). Kendall’s W can also be used to test the null hypothesis of no agreement among the rankings produced by the three or more measures. The only possible alternative hypothesis is that agreement exists among the rankings produced by the three or more measures. The analytical procedure followed to test this hypothesis involved computing Kendall’s W and determining its significance using the appropriate statistical tables. Details of the analytical procedure are provided in Chapter V.

**Research Question # 8(a):** Do the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology agree in terms of the order in which they rank alternative states of health?

**Research Question # 8(b):** Do the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice
conjoint methodology or DCCM agree in terms of the order in which they rank alternative states of health?

**Research Question # 8(c):** Do the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM agree in terms of the order in which they rank alternative states of health?

As noted earlier, the issue of agreement in preferential ordering of alternative health states between preference measurement methodologies taken two methodologies at a time was addressed using Kendall’s τ. Kendall’s τ is bounded between -1 and +1, with -1 indicating perfect disagreement between the two measures, +1 indicating perfect agreement between the two measures, and 0 indicating no agreement between the two measures. Kendall’s τ can also be used to test the null hypothesis of no agreement among the rankings produced by the two measures. It was decided to conduct a one-tailed significance test for hypothesis testing purposes, since the motivation behind this research question was to determine whether there was agreement -- not disagreement -- between the preference measurement methodologies. As explained earlier, since all three methodologies measure the same construct, it is important to measure the extent of agreement among them. Further, use of a one-tailed test provides one with more statistical power for hypothesis testing. Therefore, the alternative hypothesis was that agreement exists among the rankings
produced by the two measures. The analytical procedure followed to test this hypothesis involved computing Kendall's τ and determining its significance using the appropriate statistical tables. Details of the analytical procedure are provided in Chapter V.

(ii) Magnitude of Preference Scores

The second aspect on which the different preference measurement methodologies were compared was in terms of the magnitude of the preference scores they ascribed to the alternative health states under evaluation. The specific research question investigated in this regard was stated as follows:

**Research Question # 9:** Do the three health status preference measurement methodologies (i.e., the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) agree in terms of the magnitude of the preference scores they ascribe to alternative states of health?

Since each of the three preference measurement methodologies incorporated different information and underlying assumptions, it was hypothesized that they would give different scores to the same health states. More specifically, it was hypothesized that the conditional utility function-
based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM would give higher scores than the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology. This was because the standard gamble scaling approach incorporates the notion of risk, while the other methods do not. Further, since the statistically inferred measurement strategy used as part of the DCCM presents each alternative health state as a composite, it incorporates interactions -- the "divide and conquer" principle of the conditional utility function-based measurement strategy would not allow this, and therefore the rating scale scaling method with this strategy would give lower values than those generated by any statistically inferred measurement strategy. Since the published literature did not provide much guidance in terms of the comparison between preference scores obtained using the CUF-SG methodology and the DCCM methodology, a non-directional research hypothesis was used in this particular case. This question was evaluated by a repeated measures one-way analysis of variance with a Scheffe test for multiple comparisons. Details of the analytical procedure are provided in Chapter V.

(iii) Correlations

The third aspect on which the different preference measurement methodologies were compared was in terms of the correlations between the preference scores of multiattribute health states obtained using the different methodologies. Since the different preference measurement methodologies are all intended to measure the same construct (i.e., health status
preferences), they should be positively correlated with one another. The correlation between the preference scores obtained using the three preference measurement methodologies was measured using two different types of correlation coefficients. Each of these correlation coefficients provided information about the degree of correlation between different aspects of the preference scores of multiattribute health states obtained using the three preference measurement methodologies.

The first correlation coefficient, Spearman's rho (or Spearman's ρ), provided information about the correlation between the preferential rank ordering of health states produced by the three preference measurement methodologies. The second correlation coefficient, Pearson's product moment correlation coefficient, provided information about the correlation between the actual preference scores obtained using the three preference measurement methodologies without adjustment for the fact that there would be systematic differences between these preference scores. The specific research questions, hypotheses and analysis procedures involved in each of these correlation coefficients are described in this section in turn.

**Research Question # 10(a):** Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?
Research Question # 10(b): Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

Research Question # 10(c): Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?

This set of research questions was addressed using Spearman's rho or ρ. Spearman's ρ is a measure of the correlation between two measures in terms of the order in which they rank a particular variable. i.e., it is a rank-order correlation coefficient between two measures. It ranges in value between -1 and +1, where -1 indicates perfect negative correlation between the rank ordering of the two measures, +1 indicates perfect positive correlation between the rank ordering of the two measures, and 0 indicates no relationship or association between the rank ordering of the two measures. Spearman's ρ was also be used to test the null hypothesis of no correlation or association among the rankings produced by the two measures. It was decided to conduct a one-tailed significance test for hypothesis testing purposes for the same reasons explained earlier in this section (in the discussion of Kendall's τ). These reasons will not be repeated here. The alternative hypothesis was that correlation or association exists
among the rankings produced by the preference measurement methodologies, taken two at a time. The analytical procedure followed to test this hypothesis involved computing Spearman’s ρ and determining its significance using the appropriate statistical tables. The familywise error was maintained at 0.05 using a Bonferroni adjustment. Details of the analytical procedure are provided in Chapter V.

**Research Question # 11(a):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

**Research Question # 11(b):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

**Research Question # 11(c):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?
This set of research questions was addressed using Pearson's product moment correlation coefficient or Pearson's \( r \). Pearson's \( r \) is a measure of the linear correlation between two measures in terms of the scores they give to a particular variable. It ranges in value between -1 and +1, where -1 indicates perfect negative correlation between the rank ordering of the two measures, +1 indicates perfect positive correlation between the rank ordering of the two measures, and 0 indicates no relationship or association between the rank ordering of the two measures.

Since Pearson's \( r \) is a descriptive measure of correlation or association between the scores obtained by two measures, it can also be used to test the null hypothesis of no correlation or association among the scores obtained by the two measures. As noted in the case of Spearman's \( \rho \) earlier in this section, it was decided to conduct a one-tailed significance test for hypothesis testing purposes. The alternative hypothesis was that correlation or association exists among the rankings produced by the preference measurement methodologies, taken two at a time. The analytical procedure followed to test this hypothesis involved computing Pearson's \( r \) and determining its significance using the appropriate statistical tables. The familywise error was maintained at 0.05 using a Bonferroni adjustment. Details of the analytical procedure are provided in Chapter V.

(iv) Concordances

The fourth aspect on which the different preference measurement methodologies were compared was in terms of the concordance among the preference scores of multiattribute health states obtained using the different
methodologies. As explained earlier in this chapter, it is important to maintain the distinction between correlation and concordance while comparing different measurement methodologies. This is especially important for the purpose of this study, since the different preference measurement methodologies were expected to give significantly different preference scores to the same health states. Concordance coefficients indicate the strength of agreement among preference scores obtained using different methodologies in recognition of the theoretically expected differences between the scores. Therefore, they allow the testing of the premise that although the different health status preference measurement methodologies ascribe significantly different preference scores to states of health, they still measure the same construct -- that is, if the concordance coefficients are found to be statistically significant.

The concordance coefficient used to measure the concordance among the different preference measurement methodologies investigated in this study was the intraclass correlation coefficient or ICC. The ICC assesses not only the strength of correlation between measures, but is also a test for the difference in means of the measures. As Kramer and Feinstein (1981) have noted, the ICC assesses both, the similarity in slopes (like the Pearson r) as well as the similarity in intercepts. Thus, the ICC provides an estimate of the correlation between two or more measures after adjusting for the fact that the measures might be significantly different from one another. In other words, it offers an estimate of the correlation corrected for systematic bias. The ICC is not restricted to the case of only two measures; it can be extended to several measures. Like the Pearson r, the values of the ICC range from -1
to +1, with -1 and +1 indicating perfect disagreement and agreement among the measures respectively and 0 indicating random agreement among the measures.

Two types of ICCs were calculated in order to address the issue of concordance among the preference scores obtained using the different preference measurement methodologies. The first considered all three health status preference measurement methodologies together, while the second considered the methodologies two at a time. It was decided to use both ICCs in order to give an overall indication of concordance among all three preference measurement methodologies as well as to provide more specific indications of concordance between each of the preference measurement methodologies taken two at a time. The specific research questions and hypotheses tested to address each particular research question will be presented first for the case of all three preference measurement methodologies taken together, and then for the preference measurement methodologies taken two at a time.

**Research Question # 12:** Is there significant concordance among the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM?
This particular research question was addressed using the overall ICC among the three different preference measurement methodologies. Since the ICC is a descriptive measure of concordance between the scores obtained by the three measures, it can also be used to test the null hypothesis of no concordance among the scores obtained by these three measures. Like the other hypotheses tested, it was decided to use a one-tailed hypothesis for hypothesis testing purposes. Therefore, the alternative hypothesis was that concordance exists among the preference scores produced by the preference measurement methodologies. The analytical procedure followed to test this hypothesis involved computing the overall ICC among the three preference measurement methodologies and determining its significance using the appropriate statistical tables. Details of the analytical procedure are provided in Chapter V.

**Research Question # 13(a):** Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

**Research Question # 13(b):** Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?
Research Question # 13(c): Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?

This set of research questions was addressed using the ICC between each of the two measurement methodologies in question. For the same reasons as described above, it was decided to use a one-tailed hypothesis in this case. The analytical procedure followed to test this hypothesis involved computing the ICC between the three preference measurement methodologies (taken two at a time) and determining its significance using the appropriate statistical tables. The familywise error was maintained at 0.05 using a Bonferroni adjustment. Details of the analytical procedure are provided in Chapter V.

(v) Functional Relationships

It had been noted earlier in this chapter, that although different preference measurement methodologies might ascribe significantly different preference scores to multiattribute health states, the preference functions determined using each methodology should still be related to one another by a functional relationship, defined as the expression of the preference scores obtained using one methodology as a mathematical function of the other. This is simply because if these methodologies do indeed measure the same construct (i.e., health status preferences), the scores obtained using any
one of them should in theory be a mathematical function of each of the others -- although not necessarily a linear function. The specific nature of the functional relationship would be dictated by theoretical considerations.

Three specific research questions were investigated in this study, pertaining to the functional relationship between each set of two preference measurement methodologies. The first two of these research questions were concerned the functional relationship between value and utility functions, while the third was concerned with the functional relationship between two value functions. The first two questions will be discussed together, followed by the third question.

**Research Question # 14(a):** What is the nature of the functional relationship between the preference functions for alternative health states obtained using the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

**Research Question # 14(b):** What is the nature of the functional relationship between the preference functions for alternative health states obtained using the discrete choice conjoint methodology or DCCM and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?
Theoretical considerations and past empirical research strongly support the existence of a power function relating measures of utility and value (Barron, von Winterfeldt, and Fischer, 1984; Torrance, Zhang, Feeny et al. 1992). Therefore, the relationship between value and utility functions -- i.e., CUF-RS and CUF-SG and also DCCM and CUF-SG methodologies -- was expected to be represented by a power function. Notationally, the expected relationships between the value and utility functions investigated in this study were:

\[ \text{CUF-SG} = \text{CUF-RS}^x \]  \hspace{1cm} (2.1)

and

\[ \text{CUF-SG} = \text{DCCM}^y \]  \hspace{1cm} (2.2)

where

- CUF-SG = conditional utility function-based procedure with the standard gamble scaling method
- CUF-RS = conditional utility function-based procedure with the rating scale scaling method
- DCCM = discrete choice conjoint methodology
- \( x \) and \( y \) are constants

The analytical procedure used to estimate the functional relationship between these functions was straight line regression on the natural logs of the preference scores. A Bonferroni adjustment was made to maintain the familywise error at 0.05 and the fit of the function was estimated using an
adjusted \( R^2 \) formula. Details of the analytical procedure which was followed are provided in Chapter V.

**Research Question # 14(c):** What is the nature of the functional relationship between the preference functions for alternative health states obtained using the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

Theoretical considerations and empirical evidence suggest that the relationship between two value functions -- i.e., the CUF-RS and DCCM methodologies -- may be represented by a simple linear function (Fischer, 1977). Notationally, the expected relationship between the two value functions investigated in this study was:

\[
DCCM = a + b \times (\text{CUF-RS})
\]

(2.3)

where

- CUF-RS = conditional utility function-based procedure with the rating scale scaling method
- DCCM = discrete choice conjoint methodology
- \( a \) and \( b \) are constants

This functional relationship was also estimated using straight line regression. Again, a Bonferroni adjustment was made and an adjusted \( R^2 \)
value calculated. Details of the analytical procedure which was followed are provided in Chapter V.

(c) Comparing Respondent Evaluations

The last category of research questions was concerned with comparing the respondent evaluations of the different preference measurement methodologies investigated in this study. As noted earlier in this chapter, respondent evaluations are an important aspect to be considered in comparing different preference measurement methodologies. As indicators of the degree of burden subjects face when performing the tasks necessary to implement different preference measurement methodologies and of the acceptability of the different preference measurement methodologies, respondent evaluations may be helpful in making the decision about which methodology to use in any given situation. However, a review of the published literature on respondent evaluations of preference measurement methodologies (see Chapter IV) revealed that there was a lack of a structure to such evaluations. In other words, there is little understanding about the particular dimensions which make up the overall construct of respondent evaluations. Therefore, a new instrument had to be developed for the purpose of measuring respondent evaluations. The rationale for the selection of the particular variables (i.e., interest, sensibility, difficulty in doing, usefulness, length, success, clarity, difficulty in understanding, and confidence) used to operationalize the respondent evaluations is provided in Chapter V.
Since the instrument used to measure respondent evaluations was a new one, its psychometric properties of the data collected with it needed to be investigated before these data could be interpreted. Therefore, it was necessary to address a few preliminary questions about the psychometric properties of this instrument before analyzing the data on respondent evaluations of the different preference measurement methodologies.

The analyses conducted in this regard were intended to estimate the reliability and construct validity of the data collected using this instrument. Reliability was estimated using Cronbach’s alpha (Cronbach, 1951), while construct validity was investigated using exploratory factor analysis. The specific research questions for these preliminary analyses on reliability were:

**Research Question # 15(a):** What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?

**Research Question # 15(b):** What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?
Research Question # 15(c): What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the discrete choice conjoint methodology or DCCM?

The specific research questions for these preliminary analyses on construct validity were:

Research Question # 16(a): What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?

Research Question # 16(b): What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Research Question # 16(c): What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the discrete choice conjoint methodology or DCCM?
After addressing the psychometric properties of the data collected on respondent evaluations of the different preference measurement methodologies, these data could be analyzed. The specific research question in this regard was:

**Research Question # 17:** Are there any significant differences among the respondent evaluations for the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM in terms of the following variables:

(a) interest
(b) sensibility
(c) difficulty in doing
(d) usefulness
(e) length
(f) success
(g) clarity
(h) difficulty in understanding
(i) confidence

Since the published literature does not offer much guidance in terms of the relative performance of the three preference measurement
methodologies investigated in this dissertation on the variables used to operationalize the respondent evaluations, non-directional research hypotheses were used in each case. The analytical procedure involved in testing each of these hypotheses involved a series of repeated measures one-way analyses of variance, with the Scheffe test for multiple comparisons. Details on the analytical procedure are provided in Chapter V.

Finally, given the existence of the phenomenon of preference reversal (as explained earlier in this chapter), it is possible that the preferential ordering of measurement methodologies inferred on the basis of respondent evaluations using judgmental ratings (as above) and choices may be different. Therefore, direct comparative respondent evaluations of the three preference measurement methodologies in the form of preferential rank ordering of the each of the variables noted above were also collected and compared to the preferential rank ordering inferred from the judgmental ratings provided in response to Research Question #17 above. The specific research question in this regard can be stated as:

**Research Question #18:** Is there a difference between the preferential rank ordering of the three health status preference measurement methodologies (the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) as determined by responses to questions using direct
comparison and judgmental ratings as the response modes in terms of the following variables:

(a) interest
(b) sensibility
(c) difficulty in doing
(d) usefulness
(e) length
(f) success
(g) clarity
(h) difficulty in understanding
(i) confidence

In order to provide a direct comparative assessment of the three preference measurement methodologies in terms of each of the variables measuring respondent evaluations, simple frequency counts and percentages were computed for each of three places in the rankings of the different attributes. These were then compared to the preferential rank ordering of the three methodologies as inferred from the judgmental ratings of the methodologies on each of the variables. On account of the small sample size for this comparison (only three methodologies), inferential statistics were not computed; only descriptive statistics (i.e., frequencies, percentages, and Kendall’s τ are reported for this research question).
Assumptions

The following assumptions underlying the conceptualization and operationalization of constructs in this study should be recognized. First, health status in this study is viewed as a multidimensional construct which can be split into separately measurable dimensions or attributes. As noted in the previous chapter, this view is the dominant one in health status research. Second, patients are the best judges of their own welfare. This is the reason why it was decided to select antihypertensive patients as the subjects for this study. Finally, a necessary assumption involved in using the particular discrete choice conjoint methodology used in this study (i.e., the CBC System) is that the MNL choice model describes the underlying decision making process of respondents.

Definitions

The definition of health that was adopted for this study was that of the World Health Organization (1947): “Health is a state of physical, mental, and social well-being, and not merely the absence of disease or infirmity.”

Health status of a patient is that patient’s health at a given point in time. It was operationalized in this study using the EuroQol, and therefore consisted of five dimensions -- mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Mobility was operationalized by three levels of functioning -- no problems in walking about, some problems in walking about, and confined to bed. Self-care was operationalized by three levels of functioning -- no problems with self-care, some problems with
self-care, and unable to wash or dress self. Usual activities was operationalized by three levels of functioning -- no problems with performing usual activities, some problems with performing usual activities, and unable to perform usual activities. Pain/discomfort was operationalized by three levels of functioning -- no pain or discomfort, moderate pain or discomfort, and extreme pain or discomfort. Anxiety/depression was operationalized by three levels of functioning -- not anxious or depressed, moderately anxious or depressed, and extremely anxious or depressed.

The definition of health-related quality of life that was adopted for this study was that of Pathak and MacKeigan (1992): “Health-related quality of life can be defined as a comparative judgment based on a point-in-time assessment of an individual’s present health state relative to that individual’s reference health state(s)” (p.37).

Preferences for health states, health status dimensions, and levels of health status dimensions as used in this study were operationally defined as the scores obtained using the CUF-RS, CUF-SG, and DCCM methodologies for health states, health status dimensions, and levels of health status dimensions respectively.

As determined by the CUF-RS methodology, preferences for a given health state were read off the 0-100 scale. As determined by the CUF-SG methodology, preferences for a given health state were defined as the indifference probability of being in that health state for sure and a gamble with chance $p$ of being in perfect health and a chance $(1-p)$ of immediate death. As determined by the DCCM, preferences for a given health state
were calculated using the parameter estimates yielded by multinomial logit modeling of the data collected using the pick-one scaling method.

**Respondent evaluations** of the health status preference measurement methodologies are the evaluations of the three preference measurement methodologies used in this study by the respondents of the study. Specifically, the variables evaluated were interest, sensibility, difficulty in doing, usefulness, length, success, clarity, difficulty in understanding, and confidence. The operational definition of the first eight of these attributes was the score the respondent provided on a seven-point semantic differential, with the descriptive labels “extremely”, “moderately”, and “slightly” on either side of the “neutral” label. The operational definition of the attribute of confidence was the score provided by respondents on a five-point Likert scale ranging from “very unconfident” to “very confident”, with the intermediate labels “somewhat unconfident”, “neither confident nor unconfident”, and “somewhat confident”.

Further, respondent evaluations were also measured using a direct comparison of the three methodologies (i.e., CUF-RS, CUF-SG, and DCCM) on each of the variables evaluated -- i.e., interest, sensibility, difficulty in doing, usefulness, length, success, clarity, difficulty in understanding, and confidence. In this regard, respondents were asked to rank the three methodologies on each of the variables evaluated. Respondents were also asked to provide an overall ranking of the three preference measurement methodologies.
Significance

Although the primary emphasis of this dissertation is methodological, it does have some significance for clinical practice. This section discusses the significance of this study with respect to health status measurement, pharmacoeconomic evaluations, theory development, and clinical practice.

Significance to Health Status Measurement

This study used the EuroQol health status measurement instrument to operationalize health status. There have been no published applications of the EuroQol in the US population. This study will provide such an application. Further, the issue of modeling the EuroQol data in terms of a model form for health status as measured by the EuroQol is still in its infancy (see Dolan and Kind, 1994; van Hout and McDonnell, 1992). By modeling the EuroQol health status preference function using three different preference measurement methodologies, this study will advance the development of the EuroQol from a methodological front.

Significance to Pharmacoeconomic Evaluations

Given the multitude of health status preference measurement methodologies, it is important to understand how and why the results obtained via different methodologies might vary. This is because there is a tendency to make a comparison between cost-utility ratios obtained using different measurement methodologies, sometimes drawn from separate studies. By estimating functional relationships between the methodologies investigated in this study, it will be possible to generate formulae to convert
scores obtained via one methodology to another. This will allow health status preferences to be compared on the basis of a common denominator, thus facilitating comparisons of denominators of cost-utility ratios.

Significance to Theory Development

The contributions this dissertation will make to theory development are twofold. First, the study will test the feasibility and validity of applying discrete choice conjoint methodology to the measurement of health status preferences. If the method does indeed prove feasible and valid, its theoretical linkage of preferences to choices would suggest that it supplant the usual judgment-based preference measurement methods used currently in health status preference measurement.

Second, the issue of respondent evaluations of preference measurement methodologies has dealt with in a relative ad-hoc manner in the published literature (as reviewed in Chapter IV). By developing and psychometrically testing an instrument for measuring respondent evaluations of preference measurement methodologies, this study will provide better insight into the underlying factor structure of the overall construct of respondent evaluations of preference measurement methodologies. Such information may be useful in selecting the appropriate preference measurement methodology for use in applied settings.

Significance to Clinical Practice

It was noted in the previous chapter that if pharmacists are expected to adopt and fulfill the mission of pharmaceutical care, it is necessary for
them to understand the role of health status evaluation and measurement. Further, in order to be effective counselors of drug therapy for patients, pharmacists need to understand patients' evaluation of their own health which, as noted earlier (Milio, 1983), is stated in terms of health status. This is especially true while dealing with patients with chronic conditions, in whom improvement of health status may be the primary goal of therapy.

This study will elicit measures of patients' preferences for health states using choice as the response measure. As explained earlier, such preference measures may help in providing true meaning to health status scores. This may facilitate the use of such scores in routine clinical decision making. This in turn will present patients with the opportunity to let their preferences be known to the clinicians making decisions on their behalf. As noted earlier, the published literature suggests that patients would like to let clinicians know what they want and discuss alternatives with clinicians but leave the final decision (or choice of therapy) to clinicians (Strull, Lo, and Charles, 1984). By letting clinicians know patients' preferences for health states, it is possible that the therapeutic decisions made by these clinicians would be more responsive to patient preferences. This in turn would lead to positive patient outcomes, such as better compliance with therapy, increased patient satisfaction, and improved health-related quality of life.

Finally, by providing meaning to health status scores, this study may facilitate their use in making "patient-centered" reimbursement decisions. As noted earlier, a major driver behind the development of health status scores was to let payors know what they were getting for the money they spent on health care. Preferences for health states will give the patient's
perspective on how much they value the money payors are spending on them. This information could in turn be used by payors in making "patient-centered" reimbursement decisions.

Outline of Remaining Chapters

The next chapter (i.e., Chapter III) presents a rather detailed review of the published literature on preference measurement in general and health status preference measurement in particular. This is followed by Chapter IV, which details the study design, data collection, and analysis procedures followed. Chapter V presents the results of the study. Finally, Chapter VI summarizes and discusses the results and makes conclusions and appropriate recommendations for further research in the area of health status preference measurement in general and discrete choice conjoint methodology in particular.
CHAPTER III

HEALTH STATUS PREFERENCE MEASUREMENT METHODOLOGIES

Chapter Overview

Given the methodological emphasis of this dissertation, it was decided to review the methodologies for health status preference measurement in some detail in this chapter. The purpose of this rather detailed review is to gain better insight into the relative strengths and weaknesses of the different methodologies for preference measurement, especially as they apply to the area of health status preference measurement in terms of the issues raised in the previous chapter. In order to set the stage, a framework of subjective measurement is provided, with the help of which respondents’ information processing strategies in response to different preference measurement methodologies can be understood. The chapter continues with a review of the basic measurement strategies for health status preference measurement, especially focusing on the strategies subsumed within conjoint methodology since the discrete choice strategy introduced as a health status preference measure in this dissertation falls within the purview of these strategies. The next section of this chapter reviews the different scaling approaches which can be used along with the different
measurement strategies and concludes with a list of points to consider while selecting the most appropriate measurement strategy/scaling approach combination for any given health status preference measurement study. Finally, the research by the three most influential groups of health status preference measurement researchers is reviewed.

Methodologies for Health Status Preference Measurement

Froberg and Kane (1989a, 1989b, 1989c, 1989d) have provided an excellent review of the methodologies for health status preference measurement. They conceptualized the general methodology of health status preference measurement as consisting of four basic steps -- defining a health status classification system, deciding upon a measurement strategy, choosing a scaling approach, and implementing the selected methodology (Froberg and Kane, 1989a). This review will focus on the second and third steps of this methodology -- i.e., deciding upon a measurement strategy and choosing a scaling approach -- since these are the steps which are of most relevance to this dissertation. The first and fourth steps -- i.e., defining a health status classification system and implementing the selected methodology -- will be addressed in Chapter V, specifically in terms of their usage in this dissertation.

Froberg and Kane (1989a) referred to the overall structure for posing questions to respondents (i.e., having respondents evaluate multiattribute or multidimensional health states versus evaluating each attribute or dimension separately) and the corresponding method of analyzing the data
(i.e., multiple linear regression analysis, analysis of variance, logit, or probit) as the measurement strategy. On the other hand, they considered the scaling approach to be the specific task required of the respondent to obtain scale scores for health states (Froberg and Kane, 1989a).

The measurement of preferences for health states is inherently subjective. In deciding upon a methodology for measuring such subjective constructs, health status researchers were heavily influenced by advances in the fields of psychophysics and decision theory. Over the decades, a number of strategies for measuring subjective preferences have been developed in these fields. These strategies differ primarily with respect to the research design used to generate the questions posed to respondents.

The research design dictates the kind of hypotheses that can be tested, the methods of analysis for testing these hypotheses, and the types of conclusions that can be appropriately drawn about the meaning of the observed responses. The strongest conclusions are those which specify why a particular observed response occurred and not some other response. Such conclusions are best drawn from experimental designs, which manipulate the levels of one or more independent variables and determine the effect of these manipulated levels on the dependent variable of interest in the study. When measurement strategies are used that appropriately allow such causal and subjective conclusions to be drawn, it becomes possible to know how to alter a given situation so as to alter the response (Veit and Ware, 1982).

Applied to health status preference measurement, if each health status dimension were considered to be an independent variable, experimental designs could manipulate the levels of these dimensions in generating health
status profiles to be evaluated by respondents. Depending on the evaluation given to a particular profile, an understanding of the respondents' preferences for the different health status dimensions could be obtained. Further, since experimental designs manipulate the levels of the independent variables (i.e., health status dimensions) and each respondent evaluates several health status profiles, use of such designs permits cause-and-effect conclusions to be drawn about respondents' preferences for various health status dimensions. Such conclusions are necessary to provide meaning to health status scores.

This focus on hypothesis testing and experimental research designs is particularly important in health status preference measurement. This is because subjective preferences for health states cannot be compared to any external standard of accuracy, i.e., there is no "gold standard" or criterion measure to act as a comparative standard. In the absence of an external standard for validating health status preferences, then, it is necessary to incorporate empirically testable hypotheses in the design of the measurement strategy, the results of which could provide some reassurance of the validity of the stated preferences for health states obtained in the study.

Measurement strategy considerations must therefore precede scaling approach considerations, because besides determining the types of questions that will be posed to the respondent, the measurement strategy also specifies the kinds of hypotheses that can be tested and therefore the types of conclusions which may be drawn from the data (Froberg and Kane, 1989a; Veit and Ware, 1982). The choice of measurement strategy also has a major
impact on the amount of information that can be given to respondents and the degree of respondent burden (Cadman and Goldsmith, 1986). Further, the particular scaling approach used will also depend on the constraints imposed by the measurement strategy. According to this view, then, greater emphasis is placed on considerations of measurement strategy than those of scaling approach. It should be noted that the importance of scaling remains unchanged, since the scale scores obtained are health state preferences and the objective of the study is to determine these health state preferences. However, by grounding scaling approach considerations in the measurement strategy, the interpretation of the scale scores is more meaningful.

A necessary step in understanding individuals' preferences for health states is understanding the process by which they integrate information about components of health status and process this information in giving an evaluation of a particular health state. This is particularly important in light of the role of interactions between health status dimensions in determining overall health status preference scores, as discussed in Chapter II. Therefore, before discussing and evaluating the different strategies and scaling approaches for measuring health status preferences, it is necessary to introduce a framework of subjective measurement which illustrates how respondents process the information presented to them and provide a response.
A Framework of Subjective Measurement

Subjective measurement refers to the measurement of preferences, events, processes, etc. that occur "in the head," i.e., they cannot be directly observed but have to be inferred from observed behavior. The observed behavior usually consists of responses that individuals make to questionnaire items. Measurement requires interpreting these responses and drawing conclusions about what they mean (Anderson 1970, 1971a, 1981, 1982; Veit and Ware, 1982).

Figure 3 is a framework of subjective measurement in the context of human information processing which is modified from Anderson's (1970, 1971a, 1981, 1982) research on functional measurement (which will be explained later in this chapter). The stimuli (i and j) and the observed response (R) are the only elements of the framework that the researcher can directly observe. The subjective parameters and processes (including scale scores and functions) that need to be measured are those that occur between the time an item or stimulus is perceived and a response is given by the respondent. Although Anderson's (1970, 1971a, 1981, 1982) functional measurement approach -- and therefore this particular framework -- was developed with respect to a particular type of observed response (i.e., at least ordinal in level of measurement) and does not strictly apply to discrete choice observed responses (i.e., nominal in level of measurement), the overall framework is essentially the same for all types of observed responses. It is in this regard that this framework is introduced here, as it is the overall framework is important in understanding the strengths and weaknesses of the different measurement strategies.
Legend:
i, j: stimulus information (observed)
$S_i, S_j$: stimulus scale scores (subjective)
$\psi_{ij}$: response scale score (subjective)
$R_{ij}$: overt response (observed)
H: psychophysical function or utility function
C: combination function or model form
J: judgment function or psychomotor function

Figure 3: A Framework of Subjective Measurement
[adapted from Anderson (1970)]
Anderson (1970) himself noted this point, when he drew attention to the basic similarity between functional measurement and signal detection theory (Green and Swets, 1966), which focuses on the judgmental processes individuals use in evaluating stimuli by means of discrete choice observed responses: "Signal detection theory applies primarily to discrete choice responses, whereas ... (functional measurement theory) ... is mainly concerned with continuous responses. This difference in response measures leads to considerable differences in theoretical approach and in method, but the parallel is still significant (Anderson, 1970 p.168)." Moreover, none of the published literature on health status preference measurement has used discrete choices as the response mode -- hence this framework will suffice for evaluating the published literature in this area.

If we think of the observed stimulus information (i and j) on the left as particular levels of two attributes or dimensions of a health state (e.g., mental and physical health), the framework operates as follows: first, the respondent transforms each piece of information (e.g., some problems in walking about, not anxious or depressed) contained in a health state into a subjective stimulus score (S_i, S_j) by the function H. In the language of psychometric theorists, this function H is referred to as the psychophysical function when the stimulus information consists of physical measures (such as time, probability, or willingness to pay or accept a monetary amount); in the language of decision theorists, this function H is the utility curve and the subjective stimulus scores (S_i, S_j) are the utiles. The respondent then uses a particular combination function (C) to transform these scale scores into a subjective response (ψ_ij). This combination function has been referred to by
various terms in the literature, including combination rule (Veit and Ware, 1982), psychological function (Anderson, 1970), integration function (Anderson, 1970), model form (Fischer, 1979), and algebraic model (Veit, Rose, and Ware, 1982; Veit and Ware, 1982). For the purpose of this dissertation, it will be referred to as a model form or combination function. Finally, the respondent transforms this subjective response into an observed response \( l \) using the judgment function \( f \) (also called a psychomotor function). It should be noted that although this framework was presented in terms of two stimuli, it can easily be extended to include more stimuli.

Measurement researchers are typically interested in measuring any one or all of the subjective scale scores and functions shown in Figure 3. However, in order to provide true and complete meaning to the observed response \( R \), it is necessary to make conclusions about all of the subjective scale scores and functions shown in Figure 3. Often, the researcher makes an explicit interpretation about only one of these events, and conclusions about the other events are implicit in the given interpretation but have to be deciphered by the reader. As Veit and Ware (1982) have noted, when these conclusions have been tested and verified, they are credible. The problem arises when conclusions are based on premises that are not tested, and further cannot be tested because the measurement method used does not provide the means to test them. In such cases, the observed as well as scale scores may not be credible and therefore could be considered suspect for use as inputs in public policy and clinical decision making.

Therefore, the methodology used for measuring health state preferences must allow for the inclusion and empirical testing of hypotheses
which specify all the subjective scale scores and functions in the subjective measurement framework. To test such hypotheses requires the use of appropriate experimental designs in the measurement strategy. In the absence of such experimental designs, the scale scores and functions of the framework can only be assumed to be correct, i.e., bases for conclusions about the scale scores and functions are definitional as opposed to empirical. As explained below, some of the measurement strategies used permit the use of such designs while others do not. Against the backdrop of this subjective measurement framework, the different measurement strategies and scaling approaches for measuring health status preferences can be evaluated.

**Measurement Strategies for Health Status Preference Measurement**

Building on the framework provided by Fischer (1979) and Veit, Rose, and Ware (1982), Froberg and Kane (1989a) have identified two basic types of strategies for measuring health status preferences -- holistic and decomposed preference measurement strategies. The major difference between these strategies is in the level of breakdown of scores obtained -- in the holistic strategy only an overall health status preference score can be obtained, while in the decomposed strategies preference scores can be obtained for each health status dimension, each level of each health status dimension, as well as an overall health status preference score. Apart from these two basic types of strategies, hybrid or mixed holistic-decomposed strategies may also be used. Figure 4 presents a classification of the health status preference measurement strategies that are discussed in this chapter.
Figure 4: A Classification of Preference Measurement Strategies
Holistic Preference Measurement Strategies

The holistic preference measurement strategy requires the respondent to evaluate every possible health state of interest in the study, where a given health state is a combination of one level of each dimension of health status decided upon in the health status classification system. This strategy is typically implemented in three steps: (a) developing a description of all possible health states under the health status classification system, (b) asking subjects to order these health states from most to least preferred, (c) measuring the preferences for being in each of the health states by an appropriate scaling approach. Most of the early examples of measuring health state preferences used this strategy (see, for e.g., Patrick, Bush, and Chen, 1973a, 1973b; Sackett and Torrance, 1978; Torrance, 1976a).

This strategy is characterized by the use of single-factor designs to generate questionnaire items, i.e., only one factor is manipulated along a particular continuum (Veit and Ware, 1982). The single factor is usually a multidimensional health state description, and contains information relevant to physical, mental, and social function and other health status dimensions deemed relevant for the study purpose. Such designs allow a test of the main effect of the manipulated factor on responses, under the assumption that the judgment function $f$ in Figure 3 is linear (i.e., the observed response is at least on an interval scale of measurement). Thus, cause-and-effect conclusions with respect to this relationship may be drawn.

However, the main effect of the manipulated factor is rarely of interest, since levels of each health status dimension are usually selected to be different enough so as to ensure the occurrence of such main effects.
Rather, the interest is typically on either determining the subjective scale scores associated with levels of health status dimensions or $S_i$ and $S_j$ in Figure 3 (i.e., stimulus scaling) and/or determining the subjective scale scores associated with the responses or overall health status scores or $\psi_{ij}$ in Figure 3 (i.e., response scaling). Further, the interest is also in ensuring that these subjective scores are measured at least at an interval-level of measurement since, as Kaplan (1982) has noted, one of the requirements for scaling health status is that preferences for health states be measured on an interval-level scale of measurement. Aggregating partial dysfunction into units of quality-adjusted life years rests on the axiomatic properties of an interval-level scale of measurement because in order to add cases that represent different degrees of dysfunction, it is necessary to have a scale where differences at different levels of the response scale are equal (Kaplan, 1982). Interval-level scales are also important for the application of some statistical analyses and in pharmacoeconomic analyses. Thus, obtaining interval-level data for health status preferences are essential for providing meaning to health status scores.

Veit and Ware (1982) have noted that researchers using this strategy typically conclude that the questionnaire responses represent at least an interval (i.e., interval or ratio) scale of either the stimulus or the response depending on whether their purpose is stimulus or response scaling. These conclusions require making the following assumptions: (a) the observed responses ($R$) are directly (i.e., linearly) related to the underlying subjective stimulus ($S_i, S_j$) or response ($\psi_{ij}$) scale scores for stimulus and response scaling respectively, (b) respondents' combination rule (i.e., $C$ in Figure 3)
corresponds to the task instructions in the scaling approach (typically instructions are to perform an interval or ratio task, e.g., respondents might be asked to report the interval between two health states or the ratio of two health states using a particular numerical scale), and (c) the scale properties (i.e., interval or ratio) of the observed scores are what might be expected if respondents follow the task instructions (Veit and Ware, 1982).

The problem is that none of the above assumptions is tested in this single-factor design. Indeed, these assumptions cannot be tested when a such a design is used because the design does not provide the constraints necessary for distinguishing between an infinite number of alternative explanations of a set of responses (see Anderson, 1974b; Birnbaum and Veit, 1974; Krantz and Tversky, 1971; Shepard, 1976; Veit, 1978). When only a single factor is used to construct questionnaire items, there is no way to test hypotheses about the relationship between observed responses and underlying subjective stimuli or response scale scores. Moreover, there is no way to test hypotheses about what combination rule respondents use to combine stimulus information. Finally, one cannot test the assumption of the interval-level of measurement. Therefore, in the holistic strategy, it is not possible to obtain knowledge about how the stimulus information affects responses (i.e., the combination function or how the respondent processes the information in the stimuli and comes up with a response) or what constitutes an appropriate set of subjective stimulus and/or response scale scores. Responses are merely defined to be scale scores having certain properties. However, there is no empirical basis for such conclusions.
Another major limitation of this strategy is that the separate effects of each dimension cannot be analyzed. This is because the single factor being manipulated is usually a multidimensional health state description. Consequently, the separate effects of these dimensions cannot be assessed since they are confounded in the single-factor design. Hence, this strategy does not provide information on the relevant importance of different health status dimensions. All the information that can be obtained from such a strategy is an overall health status score. Further, there is the issue of respondent burden -- the large number of judgments required of respondents may be a threat to the validity of the data due to respondent fatigue and restrict the applicability of this strategy.

If subjective scores obtained from health status measurement research are to be used as inputs for pharmacoeconomic analyses and for public policy and clinical decision making, it is necessary to verify the validity of the scores. Further, it is important to understand how scores associated with the different dimensions of health status interact to produce an overall judgment of health status. The holistic strategy of health status preference measurement can neither verify the validity of the scores nor incorporate interactions between health status dimensions. These limitations have prompted researchers to adopt decomposed strategies for health status preference measurement, which show promise in these areas where the holistic strategy is deficient. Indeed, as Froberg and Kane (1989a) have noted, most of the current studies of health status preference measurement adopt such strategies.
Decomposed Preference Measurement Strategies

Decomposed preference measurement strategies are based on the principles of Multi-Attribute Utility (MAU) theory (Fischer, 1979). Strictly speaking, in light of the distinction between value and utility described earlier, these strategies do not always measure utility. Therefore, in their usage in this dissertation, these strategies can be considered to be based on the principles of “Multi-Attribute Preference (MAP) theory” (NOTE: The terms “attribute” and “dimension” are used interchangeably in this dissertation). MAP theory is concerned with establishing an axiomatic basis for various MAP model forms. In other words, it seeks to explain the nature of the combination function C in Figure 3. There are three model forms which are most commonly encountered in practice -- additive, multiplicative (or quasi-additive), and multilinear. As noted previously, additive model forms do not allow interactions, multiplicative model forms allow limited interactions, while multilinear model forms allow all possible interactions between individual attributes or dimensions of the model. MAP theory also specifies the independence conditions which are necessary and sufficient for the existence of these model forms.

Strategies based on MAP theory enable the researcher to obtain scores for all health states without requiring the respondent to evaluate each and every health state. This is facilitated by expressing the overall preference for a given health state as a decomposed function of its dimensions. In doing so, these strategies greatly reduce the number of subjective evaluations required of respondents. Within the general category of decomposed preference
measurement strategies, two basic types of strategies can be distinguished -- statistically inferred strategies and explicitly decomposed strategies.

There are two major differences between these two types of decomposed preference measurement strategies -- in terms of their theoretical origins and consequently the underlying assumptions about respondents' abilities and in terms of the "procedural flow" adopted by each strategy. Regarding the first of these major differences, statistically inferred decomposed preference measurement strategies attempt to develop an algebraic model of the respondents' preferences (i.e., preferences for health states) from a series of evaluations of multiattribute or multidimensional alternatives (i.e., descriptions of health states) via inferential statistics. The development of these strategies was heavily influenced by advances in psychometric theory.

As opposed to psychometric theory, the development of explicitly decomposed preference measurement strategies was influenced by advances in decision theory. These strategies require relatively fewer (and in some cases no) overall holistic preference measurements. Instead, these strategies permit the respondent to break up the overall evaluation process into a set of simpler subtasks and evaluate each level of each dimension separately, assuming that all other dimensions are held constant (Fischer, 1979).

This difference between the two strategies is reflective of the assumptions made about respondents' abilities by each strategy. The statistically inferred strategies assume that respondents can make a relatively large number of holistic preference assessments, each of may include a random error component. On the other hand, the explicitly
decomposed strategies assume that the respondent can make a small number of extremely precise (i.e., error-free) preference assessments (Fischer, 1979).

The second major difference between these two types of strategies is in terms of their "procedural flow." In this regard, Adelman, Sticha, and Donnell (1984) casually (but quite pertinently) drew the distinction between "top-down" and "bottom-up" flows. Explicitly decomposed strategies adopt a "bottom-up" flow, where the individual attribute or dimension preference functions are estimated and subsequently aggregated into an overall preference function for the health state under evaluation. This is in contrast to the "top-down" flow adopted by the statistically inferred strategies, which estimate the individual attribute or dimension functions on the basis of the overall preference judgments for different health states.

It should be noted that this "bottom-up" versus "top-down" distinction has also been referred to as a "compositional" versus "decompositional" approach respectively in the marketing research literature (Green and Srinivasan, 1978). It can be seen that this terminology is somewhat confusing, since what Froberg and Kane (1989a) and Fischer (1979) consider explicitly decomposed strategies are referred to as compositional strategies by the marketing researchers. This difference is nomenclatural rather than substantive, since in either case the approach itself is the same (i.e., first estimate preference functions for the different dimensions and then aggregate these into an overall preference function over all dimensions). For the purpose of this dissertation, the Froberg and Kane (1989a) and Fischer (1979) terminology will be adopted. On occasion, however, the "top-down" versus "bottom-up" distinction of Adelman,
Sticha, and Donnell (1984) will also be used to differentiate between the two types of strategies. Each of these decomposed preference measurement strategies will be reviewed in turn.

Explicitly Decomposed Preference Measurement Strategies

Explicitly decomposed preference measurement strategies have been described as being based on the "divide and conquer" principle, which has been succinctly summarized by Howard Raiffa (1968): "The spirit of decision analysis is divide and conquer: Decompose a complex problem into simpler problems, get one's thinking straight on these simpler problems, paste these analyses together with a logical glue, and come out with a program of action for the complex problem" (p.271). While there are several different procedures that can be adopted as part of these explicitly decomposed strategies, most closely resemble one of two basic procedures -- the additive rating scale method and the conditional utility function-based procedure.

The Additive Rating Scale Method

The additive rating scale method or Simple Multi-Attribute Rating Technique (SMART) and its various derivatives (e.g., SMARTS and SMARTER) is one of the easiest procedures to use. Its simplicity is largely due to the fact that it provides no explicit test of the additivity assumption. Instead, researchers employing this procedure simply assume the additive model will provide a good approximation of the "true" MAP model form and proceed under the assumption that the additive model form is valid. Consequently, this procedure requires only two additional tasks -- assessing
single dimension preference functions and assessing additive scaling constants (i.e., model parameters). Therefore, this method is very similar to the expectancy-value modeling approach in psychology and marketing, where the overall preference for a multiattribute object is determined as a weighted sum (i.e., assuming an additive model form) of the object's perceived attribute levels and associated value ratings, as separately and explicitly judged by the respondent.

This procedure was initially proposed by Edwards (1971, 1977) as a simplified modification of axiomatic multiattribute utility theory for use in applied settings. Edwards (1977) has provided a ten-step process for implementation of this procedure: (a) identify the person or organization whose utilities are to be maximized, (b) identify the issue(s) to which the utilities needed are relevant, (c) identify the alternatives to be evaluated, (d) identify the relevant attributes or dimensions of the alternatives under evaluation, (e) rank the attributes or dimensions in order of importance, (f) rate the attributes or dimensions in terms of importance, making sure to preserve ratios, (g) sum the importance weights and divide each by the sum, (h) measure the location of each alternative under evaluation on each attribute or dimension, (i) calculate the preference scores for each alternative using the additive model form, and (j) make a decision as to which alternative to choose (Edwards, 1977).

This ten-step procedure was the initial description of the additive rating scale procedure as presented by Edwards (1977). As noted earlier, the development of this procedure was motivated by Edwards' belief that the indifference judgments between pairs of hypothetical alternatives required
by axiomatic multiattribute utility theory (as reviewed in the next section) seemed difficult and unstable. Edwards believed that more nearly direct assessments of the desired quantities would be easier and less likely to produce elicitation errors (Edwards, 1977). Accordingly, in step (f) above, Edwards (1977) recommended rating the attributes or dimensions in terms of importance, making sure to preserve ratios -- i.e., a more direct scaling approach.

However, Keeney pointed out that, as initially proposed, SMART suffered from one "fatal intellectual flaw" (see Edwards, 1994 p.303). The focus in this regard is on step (f), wherein weights are measured for each attribute or dimension used to describe the alternatives under evaluation. The weights which need to be generated by this step should reflect the range of the attribute as well as its importance. Ratios as weights only tap into the importance aspect of this condition -- they ignore the range aspect. In other words, SMART (as initially proposed) ignored the fact that attribute range as well as importance must be reflected in any attribute weight used in a multiattribute model (Edwards and Barron, 1994; von Winterfeldt and Edwards, 1986).

In order to rectify this shortcoming, von Winterfeldt and Edwards recommended using swing weighting (von Winterfeldt and Edwards, 1986). The word "swing" refers to the operation of changing the score of some object of evaluation on some dimension form one value to different one (usually from 0 to 100). Swing weight elicitation proceeds in two steps -- the first step yields the rank order of the weights, while the second step yields the weights themselves via magnitude estimation or indifference judgments.
Edwards dubbed the improved version of SMART as SMARTS, for SMART using Swings (Edwards and Barron, 1994).

Recently, Edwards and Barron (1994) have proposed another modification to SMART, once again in the weighting step, i.e., step (f) above. In this new modification, weights are obtained from only rank ordering the attributes or dimensions in order of importance. In other words, the second modified step of swing weighting is eliminated. Although the idea of using rank weights was first introduced by Stillwell, Seaver, and Edwards (1981), the process of converting rank orders into weights was formally justified by Barron and Barrett (see Edwards and Barron, 1994), who also demonstrated the quality of the resulting weighting system. Edwards and Barron (1994) dubbed this new modification as SMARTER, for SMART Exploiting Ranks.

These procedures (SMART, SMARTS and SMARTER) are indeed simple and easy to use and understand. However, the major drawback with these procedures is that they do not allow for the existence of any other model forms. Further, they cannot be extended to incorporate choice as the response mode. Given the importance of interactions between health status dimensions and the ramifications of not accounting for interactions while measuring health status preferences, the application of these procedures to health status measurement may be inappropriate. Indeed, there have been very few published applications of these procedures to health status preference measurement (see Hagart and Billington, 1982). The vast majority of applications of these procedures have been reported in the behavioral decision theoretic literature by Ward Edwards and his colleagues first at the University of Michigan and now at the University of Southern California.
It should be noted that, strictly speaking, this procedure is not truly a rating scale method, since the scaling method does not have to be confined to only rating scales. Indeed, as described, the original version of SMART uses magnitude estimation, SMARTS uses swing weighting, while SMARTER uses rank order weighting as scaling methods. The details about different scaling methods will be provided later in this chapter. Although rating scales were not included as an example to illustrate the additive rating scale method, there is nothing to preclude their use along with this procedure. Regardless of the specific scaling method used, the basic procedure remains the same. It was decided to adopt the “additive rating scale” terminology for this dissertation in order to be consistent with Fischer (1979) and Froberg and Kane (1989a). Further, along these lines, it should also be noted that the additive rating scale method has been referred to in the literature as self-explicated or subjective evaluation modeling (for e.g., see Green, Carmone, and Wind, 1972; Hoepfl and Huber, 1970; Huber, 1974a, 1974b; Huber, Daneshgar, and Ford, 1971; and Huber, Sahney, and Ford, 1969).

The Conditional Utility Function-based Procedure

The second basic procedure within the class of explicitly decomposed preference measurement strategies, the conditional utility function-based procedure, is based on the seminal work of Ralph Keeney and Howard Raiffa (Keeney, 1968, 1971, 1972, 1974, 1977; Keeney and Raiffa, 1976; Raiffa, 1968a, 1968b) who axiomatized and developed this procedure for multiattribute preference estimation in normative contexts. This procedure is considerably more complex than the additive rating scale method, and relies extensively
on holistic assessments both for testing independence assumptions and for obtaining scaling constants or model parameters.

Strictly speaking, in light of the distinction drawn between values and utilities in Chapter II, this measurement strategy is best referred to as the "conditional preference function-based procedure." However, in order to be consistent with the terminology adopted in the decision theoretic literature (see Fischer, 1979), it was decided to refer to this strategy as the conditional utility function-based procedure for the purpose of this dissertation.

The conditional utility function-based procedure involves three major subtasks: (a) checking independence assumptions to determine which, if any, of the decomposed model forms is appropriate, (b) assessing preference functions over each single dimension or attribute, and (c) measuring the preferences of selected multidimensional health states to determine scaling constants or parameters, thereby permitting aggregation of preferences over dimensions or attributes (Fischer, 1979). Each of these three subtasks will be discussed in order in this section.

(a) Checking Independence Assumptions

The first subtask (i.e., checking independence assumptions) refers to independence among the dimensions. This is necessary to determine which model form — additive, multiplicative, or multilinear — is appropriate. Three conditions must be satisfied in order to assume an additive model — utility independence, mutual utility independence, and additive utility independence. If only the first of these three conditions is satisfied the model
form is multilinear, while if only the first two of these conditions is satisfied
the model form is multiplicative (Fischer 1979; Keeney and Raiffa, 1976).

These three conditions form an hierarchy. Utility independence
requires that each dimensions is utility independent of all other dimensions.
This means that preferences for various levels of each dimension do not
depend upon the particular levels at which the other dimensions are fixed. A
model form satisfying only this condition is multilinear. Mutual utility
independence requires that every subset of dimensions is utility
independent of its complement (i.e., the set of remaining dimensions). This
means that preferences for the various levels of each subset of dimensions do
not depend on the particular levels at which the remaining dimensions are
fixed. A model form satisfying this condition and the utility independence
criterion is multiplicative. Finally, additive utility independence requires
that if we let the multidimensional health state with all dimensions at their
most preferred levels equal 1.0 and the multidimensional health state with
all dimensions at their least preferred levels equal 0.0, then if each
dimension takes on its most preferred score and at the same time all
remaining dimensions take on their least preferred scores, the sum of these
preferences across dimensions should equal 1.0. In other words, the whole
(i.e., overall preference for a given health state) is equal to the sum of its
parts (i.e., the sum of the preferences for each component level of the given
health state), and that the contribution of each dimension is independent of
the scores of the remaining dimensions. If this condition is satisfied in
addition to the first two, the model form is additive (Fischer, 1979; Keeney
and Raiffa, 1976). Keeney and Raiffa (1976) have shown that additive utility
independence implies mutual utility independence, but not vice versa. Anderson's graphical test of model forms (which will be explained later in this chapter) can also apply here -- if all conditional utility functions are parallel the model form is additive, while if the conditional utility functions have a fan shape the model form is either multiplicative or multilinear (see Anderson, 1974a).

A variety of methods are available for checking these independence assumptions (see Keeney and Raiffa, 1976). Unfortunately, since they all assume that the respondent's preference assessments are free from random response error, the researcher must decide how large a deviation from linearity to accept before rejecting the each independence assumption. This is because such random error will always be present to some extent. It should be noted that the functional measurement approach described earlier deals with this issue by using the inferential procedures of analysis of variance. However, it does so at the cost of requiring a larger number of holistic preference judgments. The conditional utility function-based procedure (actually the explicitly decomposed strategies in general), by contrast, requires substantially fewer holistic preference judgments than the functional measurement approach but provides no formal mechanism for dealing with random response error. A second difficulty associated with the establishment and verification of independence conditions is the fact that it is a tedious and intensive procedure, requiring considerable interviewer-respondent interaction (Keeney, 1977). Due to financial constraints, a priori establishment and verification of independence conditions would then be feasible only in studies with a very small sample size. Thus, in practice,
researchers often modify this step. For instance, Torrance and his colleagues at McMaster University have elected to assume the existence of mutual utility independence and tested this assumption later with respondents' holistic preference judgments of multidimensional health states (see Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992).

(b) Assessing Preference Functions over Single Attributes or Dimensions

After determining which model type -- additive, multiplicative, or multilinear -- is appropriate, the researcher proceeds to the subtask of assessing model parameters. If none of these simple model types is supported, more complex decomposed models may be used (see Keeney and Raiffa, 1976). The first step in assessing model parameters is constructing a preference function over each dimension, assuming that all other dimensions are held constant. Usually the most and least preferred levels within a particular dimension are arbitrarily assigned the scores 1.0 and 0.0 respectively, and preferences for the intermediate levels are measured using an appropriate scaling method (Fischer, 1979; Froberg and Kane, 1989a). Different scaling methods are reviewed later in this chapter.

It should be noted that since each single dimension preference function is arbitrarily scaled between 0.0 and 1.0, these functions are not strictly comparable. Scaling constants must be assessed to reduce the single-dimension functions to a common unit of measure. These scaling constants reflect the relative importance of each health status dimension. This is the essence of the third and final subtask in this procedure.
(c) Determining Scaling Constants

When either the additive or multiplicative MAP model form is appropriate, scaling constants can be assessed by merely having the respondent holistically assess preferences for the m outcomes of type where one health status dimension is at its most preferred level while the other dimensions are at their least preferred levels (and m is the number of health status dimensions). If these constants sum up to one, the model form is additive; if not, the model form is multiplicative. Further, depending on whether the sum of these constants is less than or greater than one (if the model form is multiplicative), insight can be gained into the respondent's attitude toward risk. The details behind this procedure will be provided in Chapter IV, when the published literature on value-based versus utility-based preferences is reviewed. At this point, however, it is sufficient to note that in this regard the interaction parameter is important. This interaction parameter is usually determined on the basis of an iterative search procedure outlined in Keeney (1974). Computer programs to help ease the computational burden imposed by these calculations are also available (Keeney and Sicherman, 1976; Mundt, 1989).

In either case, the preference weight of each dimension is the preference for the health state in which that particular dimension is fixed at its most preferred level while the other dimensions are fixed at their least preferred levels. Scaling constants for the general multilinear model form can be obtained by a similar process. Here, however, preferences must be holistically assessed for all "corner states," i.e., health states formed by taking the most preferred levels for some dimensions and the least preferred
levels for all other dimensions. In general, there will be \((2^m - 2)\) such health states, where \(m\) is the number of dimensions. (The other two health states are best and worst, arbitrarily assigned preferences of 1.0 and 0.0 respectively). Preferences for these corner states can then be translated into scaling constants of the multilinear model form (Fischer, 1979; Keeney, 1974; Keeney and Raiffa, 1976).

The conditional utility function-based procedure has been extensively used by the McMaster group of researchers in measuring preferences for health states (see, for e.g., Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al, 1992), who conclude that it is a relatively efficient procedure for measuring health state preferences that are defined by a multidimensional health status classification system. They have also noted that compared to the holistic strategy, it is efficient in that it requires fewer respondent judgments and permits an analysis of the separate effects of each health status dimension (see Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al, 1992).

The primary advantage of this strategy is its simplicity and thus the ability to use it even when the number of attributes or dimensions is very large. However, as Green and Srinivasan (1990) have pointed out, it is difficult for the respondent to provide ratings for all the levels of a dimension holding all else equal if there are substantial dependencies among dimensions. This last point is especially disturbing, given the importance of incorporating interactions between health status dimensions and that the McMaster studies (see Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al, 1992) and the studies by Rose (1980) and Veit,
Rose, and Ware (1982) on health status preference measurement have repeatedly supported a non-additive model form. Moreover, Veit and Ware (1982) have pointed out that, like the holistic strategy, the conditional utility function-based procedure "... does not provide any way to validate the weights, the utilities, or the model and thus any prescribed outcomes; nor is there any way beyond definition of knowing what the scale properties of the numbers are (p. 253)." Further, the lack of error theory is problematic in that no measurement error is assumed in the estimation of model parameters. In this regard, all that can be obtained is a point estimate of the model parameters. There is no way to construct a confidence interval around each point estimate. This means that there is no way of knowing how much faith one can have in the point estimate parameters of the model. This is a serious drawback, given the potential uses of health status preference measures for public policy and clinical decision making.

Statistically Inferred Decomposed Preference Measurement Strategies

As noted previously, statistically inferred decomposed preference measurement strategies attempt to develop an algebraic model of the respondents' preferences (i.e., preferences for health states) from a series of evaluations of multiattribute or multidimensional alternatives (i.e., descriptions of health states) via inferential statistics. The development of these strategies was heavily influenced by advances in psychometric theory. These strategies have been referred to by various names in the published literature -- for instance, in the early psychometric literature they were
referred to as *bootstrapping* techniques, while in the marketing literature Green and Srinivasan (1978, 1990) have broadly referred to these strategies as *conjoint analyses*.

Green and Srinivasan (1978) have used the term conjoint analysis to refer to "... any decompositional method that estimates the structure of a consumer's preferences (e.g., part worths, importance weights, ideal points) given his/her overall evaluations of a set of alternatives that are prespecified in terms of levels of different attributes (emphasis in original, p.104)." Louviere (1988a) offered a similar view when he noted, "The term 'conjoint analysis' means decomposition into part-worth utilities or values of a set of individual evaluations of, or discrete choices from, a designed set of multiattribute alternatives (p.93)." Excellent reviews of conjoint analysis have been provided by Green and Srinivasan (1978, 1990) and Louviere (1988a). These review papers are complementary in the sense that the Green and Srinivasan (1978, 1990) reviews emphasize the commonalities between the different conjoint analysis strategies, while the review by Louviere (1988a) focuses on the differences between these strategies.

Conjoint analysis has been referred to by several other names in different disciplines. For instance, in the psychometric literature, it has been referred to as "external analysis" (see Carroll, 1972), while in transportation research conjoint analysis has been variously referred to as "direct utility assessment" (see Lerman and Louviere, 1978), "functional analysis" (see Benjamin and Sen, 1983), and "stated preference analysis" (Ministry of Transport and Public Works, The Netherlands, 1985). In order to avoid semantic confusion, it is necessary to decide on the nomenclature for the
statistically inferred strategies for the measurement of health state preferences. For the purpose of this dissertation, the adopted terminology for this set of strategies will be *conjoint methodology*. This is not in keeping with Louviere's (1988a) recommendation and with the terminology used in marketing and other social science disciplines, which advocate the use of the term 'conjoint analysis.' It was decided to adopt the 'conjoint methodology' terminology for this dissertation in order to include both conjoint analysis and conjoint measurement as part of these set of strategies. The distinction between conjoint analysis and conjoint measurement is explained later in this section. Since the new methodology for measuring health status preferences introduced in this dissertation falls within the purview of conjoint methodology, these strategies will be discussed in more detail than some of the other health status preference measurement strategies discussed in this chapter.

While its foundations go back to at least the 1920s, it is generally agreed that 1964 marks the beginning of conjoint methodology, with the publication of the seminal paper by Robert Duncan Luce, a mathematical psychologist, and John Tukey, a statistician (Luce and Tukey, 1964). Shortly thereafter, a number of theoretical (see, for e.g., Krantz, 1964; Krantz and Tversky, 1971a; Tversky, 1967b) and algorithmic (see, for e.g., Carroll, 1969; Kruskal, 1965; Young, 1969) contributions were published which further developed conjoint methodology.

Conjoint methodology is concerned with the joint effect of two or more independent variables or stimuli on the ordering, rating, or choice of a
dependent variable. The term conjoint refers to the simultaneous scaling of two or more stimulus variables (Anderson, 1974a). The term arose out of an attempt to apply extensive measurement to preference judgments. Extensive measurement refers to a method to build a scale by comparing relative lengths (or extensions) of objects (Huber, 1987). For example, by comparing the lengths of different rods put end to end, one can form a scale on which it is appropriate to perform mathematical operations such as addition and subtraction (i.e., an interval-level scale).

In applying extensive measurement to preference judgments, the developers of conjoint methodology reasoned that certain types of preference judgments had to be based on respondents' preferences for compound or conjoint objects. The types of judgments used initially in conjoint methodology involved selecting one out of a pair of multidimensional comparisons (i.e., the method of paired comparisons). It was shown that by putting together a number of such judgments, it was possible to derive intervally-scaled additive measures of preference. Further, a number of tests were specified to determine if such an interval scale was justified, given the preference orderings (Huber, 1987). Hence, conjoint methodology provided a way of creating an interval-level scale from ordinal judgments on compound or conjoint objects.

The initial contributions to conjoint methodology focused on finding sets of axioms and/or conditions required to uncover the latent intervally-scaled preferences which were assumed to exist in respondents. However, when these axiomatic models were applied to actual human behavior, it was found that they were violated in relatively minor but systematic ways.
(Huber, 1987). Still, the initial contributions provided a useful framework which was subsequently taken up by market researchers, who focused on the parameters obtained from an assumed model form as opposed to testing the model form itself.

This distinction led Green and Srinivasan (1978) to differentiate between the terms conjoint measurement and conjoint analysis, where the original theoretically motivated research was referred to as conjoint measurement while the applied focus was referred to as conjoint analysis. Green and Srinivasan (1978) noted, "Conjoint measurement, as practiced by mathematical psychologists, has primarily been concerned with the conditions under which there exist measurement scales for both the dependent and independent variables, given the order of the joint effects of the independent variables and a prespecified composition rule. ... However, applications by psychometricians and consumer researchers have emphasized the scaling aspects -- finding specific numerical scale values, assuming that a particular composition rule applies, possibly with some error (p.103, emphasis in original)."

As initially developed, the theory of conjoint methodology was rather limited in scope -- it only allowed for the existence of an additive model form and could only deal with ordinally-scaled responses. Since then, conjoint methodology has been substantially extended in scope and presently allows rank order, rating scale, or choice-based responses, the existence of correlations between attributes or dimensions and the existence of non-additive model forms such as multiplicative and multilinear model forms.
The development of conjoint methodology was highly influenced by advances in psychometric theory. Like the holistic strategy discussed earlier in this chapter, conjoint methodology requires the respondent to judge multidimensional health states. However, the important difference between conjoint methodology and the holistic strategy is that conjoint methodology does not require that all possible health states be evaluated. Further, from an experimental standpoint, an important methodological feature of conjoint methodology is the use of stimulus combinations via experimental designs. Each informational stimulus is systematically placed in several different combinations. This allows the effects of all stimuli to be factored out and scaled. Applied to health status preference measurement, this means that the different dimensions that comprise a given health state can be evaluated separately, thus permitting the relative importance of each dimension to be measured. This latter feature is important since it provides information about how respondents combine the dimensions and arrive at an overall evaluation -- thus one can verify the existence of interactions between the health status dimensions.

Researchers adopting conjoint methodology strategies usually assume that respondents know their own preferences and can readily provide holistic evaluations. These strategies generally consist of three stages. In the first stage, the respondent is asked to directly assess holistic preferences for a relatively large number of health states (although not as large a number as for the holistic strategy). In the second stage, the appropriateness of the different model forms considered (i.e., additive, multiplicative, or multilinear) is evaluated using a variety of goodness of fit and/or
hypothesis testing criteria. It should be noted that this stage is not always conducted since some researchers assume, rather than test, the existence of a particular model form. In the final stage, the parameters of the underlying algebraic model are estimated, typically by applying regression or analysis of variance procedures (if the dependent variable is continuous) or probit or logit analysis procedures (if the dependent variable is discrete) to the holistic evaluations.

Within this general framework, a number of different strategies are available to researchers for implementing conjoint methodology. All these strategies share the use of experimental or quasi-experimental designs to construct sets of multiattribute or multidimensional alternatives. Despite a common reliance on such designs, the various conjoint methodology strategies differ with respect to (a) response modes used to obtain information from subjects, (b) methods of analysis, and (c) inferences that can be made about judgment or choice behavior (Louviere, 1988a). On the basis of these differences, three major conjoint methodology strategies can be identified -- rank order judgment strategies, rating scale judgment strategies, and discrete choice conjoint strategies.

**Rank Order Judgment Strategies**

Rank order judgment strategies were the first conjoint strategies to be developed. Indeed, the earliest developments and publications about conjoint methodology pertained to rank order judgment strategies (see Carroll, 1969; Krantz, 1964; Krantz and Tversky, 1971a; Kruskal, 1965; Luce and Tukey, 1964; Tversky, 1967b). Further, it was these strategies that were
popularized by Green and Rao (1971) and Green and Wind (1973) in their introduction of conjoint methodology to marketing research in the early 1970s. As the name indicates, these strategies require respondents to rank order the various health states in order of preference. Ties may or may not be permitted, depending on the specifics of the research design. The axiomatic theory of rank order judgment conjoint strategies is called 'conjoint measurement' (Krantz and Tversky, 1971a). However, the theory and practice of rank order judgment strategies are rather unrelated, since the methods of analysis are not strictly based on the theory of conjoint measurement, do not have a statistical error theory, and therefore cannot be used to test the adequacy of the theory. In other words, the practical applications assume the axiomatic theory holds and do not (actually cannot) test the theory.

The theory of conjoint measurement requires real ranking data to satisfy a large number of ordinal conditions before one can conclude that a particular overall preference model form is appropriate for scaling (i.e., estimating) preferences for each level of each dimension under consideration (also called part-worths) from a respondent’s judgmental data. Most individuals are not perfectly consistent in their rankings; therefore there is error in their data. However, conjoint measurement has no error theory on which to base statistical tests of part-worth parameters or competing preference model forms.

Thus, most researchers who use these strategies assume an additive model form of the unknown part-worth parameters. These unknown parameters are usually estimated by algorithms such as PREFMAP (Carroll,
1972), MONANOVA (Kruskal, 1965; Kruskal and Carmone, 1969) LINMAP (Peelman and Sen, 1974; Srinivasan and Shocker, 1973a, 1973b), or Johnson's nonmetric tradeoff procedure (Johnson, 1973; Nels, Seaman, and Montgomery, 1976). These algorithms find a monotone transformation of the dependent variable (i.e., the observed rankings) so as to optimize the fit between the observed and predicted rankings assuming that an additive model form is correct. In terms of the subjective measurement framework presented in Figure 3, the judgment function $j$ is assumed to be only monotonic to subjective preferences (i.e., observed responses are assumed to contain only ordinal information). However, these algorithms differ in terms of their operational definitions of the poorness-of-fit or "stress" index (which measures the degree of departure between the observed rankings and rankings predicted by the additive model form) and the optimization method used to determine parameter estimates to achieve the minimum poorness of fit.

Unfortunately, these "stress" measures are closely related to the quantity $(1-R^2)$. It has been shown that $R^2$ or any monotone transformation of $R^2$ is an unreliable measure of the adequacy of conjoint as well as other models (Louviere, 1988a). Dawes and Corrigan (1974), Wainer (1976), and Anderson and Shanteau (1977) are among those who have demonstrated that: (a) conjoint methodology experiments ensure high goodness-of-fit or low poorness-of-fit measures, (b) many possible model forms can produce approximately equivalent fit measures for the same dataset, and (c) "wrong" model forms can produce better fit measures than "correct" model forms in real, fallible data (Louviere, 1988a).
Further, Louviere (1988a) has noted that factorial-type experiments guarantee that main effects will account for most of the variance in judgment data, even when wrong, i.e., when an additive model form is not correct. This is because "true" preference functions are conditionally monotone in each attribute or dimension, and the joint combination rule can be well approximated by functions that predict "higher overall preference corresponds to more high part-worths" and "lower overall preference corresponds to more low part-worths." Conjoint models mimic these conditions very well (Louviere, 1988a).

Hence, use of these strategies may support an additive model form when in fact a more complex model form is the "correct" model form. Given the importance of accounting for interactions between health status dimensions discussed earlier, the use of such strategies for measuring health status preferences may be questionable. Further, the lack of an error theory does not allow testing different model forms. Therefore, rank order judgment strategies do not allow the underlying information processing framework to be tested empirically. Hence, not only can the "wrong" assumed model form give excellent fit measures, but also there is no way to test the "correctness" of the assumed model form. As noted earlier, this is a serious drawback in terms of the lack of meaning provided to the resulting health status scores generated from the use of such strategies.

**Rating Scale Judgment Strategies**

Rating scale judgment strategies are most closely related to Information Integration Theory (Anderson, 1970, 1971a, 1981, 1982) and
Social Judgment Theory (Adelman, 1974; Brunswick 1956; Hammond, Stewart, Brehmer, et al., 1975). Accordingly, there are two different approaches to the measurement of preferences under rating scale judgment strategies: (a) functional measurement which is associated with Information Integration Theory and (b) social/clinical judgment analysis which is associated with Social Judgment Theory.

(a) Functional Measurement

The functional measurement approach to the measurement of subjective constructs using rating scale judgment strategies is mainly associated with the research of Norman Anderson (see Anderson, 1961, 1962, 1970, 1971a, 1972, 1974a, 1974b, 1981, 1982; Anderson and Zalinski, 1988). Although this approach has been classified in this dissertation as a part of conjoint methodology, attention should be drawn to the early (and sometimes acrimonious) literature which tried to differentiate between functional measurement and conjoint measurement (see Anderson, 1971b; Krantz and Tversky, 1971b). Strictly speaking, since conjoint measurement is the axiomatic theory of rank order judgment strategies and functional measurement is more concerned with the application (rather than axioms) of rating scale judgment strategies, there is a difference between functional and conjoint measurement.

Further, the early theory of conjoint measurement only applied to additive model forms, while functional measurement was more flexible in this regard since it allowed the existence of non-additive model forms as well as additive model forms. As Anderson (1974a) noted, "Conjoint
measurement has been primarily concerned with abstract mathematical analysis of certain axiom systems; functional measurement is primarily concerned with substantive, empirical theory. ... In conjoint measurement, order relations are taken as basic, and the axioms allow only for rank-order properties of the data. ... Functional measurement, in contrast, places first emphasis on numerical response measures (p.286-287).”

However, this distinction has blurred in recent times, mainly because of the extension of conjoint methodology beyond the scope of the initial theory of conjoint measurement to allow for non-rank order judgments and non-additive model forms. In this regard, functional measurement can be considered to be one of the approaches under the rating scale judgment strategies of conjoint methodology. This is the view adopted in this dissertation.

The functional measurement approach is based on the assertion that subjective constructs can be validly measured only in the context of valid theories. The emphasis is on testing the theory (i.e., the combination function) as opposed to developing measurement scales. Indeed, Anderson (1970) has noted, “A guiding idea of functional measurement is that measurement scales are derivative from substantive theory. ... The present view is that the development of the theory and of the scale are integrally and intimately related and that the final validation of the scale depends on establishing the empirical validity of the theory. On this view, the first order of business is the development of an experimental base for a quantitative behavioral law. ... Conceptually, therefore, scaling takes its proper place as a part of behavior theory (p.153-154).” It should, however, be noted that the
notion that measures derive from substantive theory was not originally proposed by Anderson. Indeed, this notion was introduced to the field of psychometrics in the 1920s by Thurstone (1927a, 1927b).

The substantive theory in the context of which functional measurement was developed is Information Integration Theory. The use of Information Integration Theory models to estimate subjective scale scores from ratings judgments in conjoint tasks is called functional measurement. The term "functional" is derived from the combination function, which constitutes the base and frame for measurement. Anderson (1976) has noted that in a subsidiary sense, the term functional is also appropriate because the stimulus scores so obtained are those that are directly functional in the behavior. The major strength of Information Integration Theory is that it has a theory of errors for rating judgments -- therefore, it has both, measurement models as well as the means to falsify them. The property of falsification is essential for theory development and hypothesis testing (Hunt, 1987). This is because it allows various model forms to be tested; hence, the actual process by which individuals process information contained in multiple stimuli and make an overall evaluation can be tested. Such testing validates the combination function or model form, and thus provides meaning to the responses.

Applied to health status preference measurement, this philosophy asserts that preferences for health states can be validly measured only in the context of a valid multidimensional preference model or combination function (i.e., additive, multiplicative, or multilinear). Thus, the process of testing models or combination functions and measuring preferences occurs
simultaneously (Anderson, 1970, 1971a, 1981, 1982). Measurement thus involves three simultaneous problems: (a) measuring the subjective stimulus scores, (b) measuring the subjective response score, and (c) finding the psychological law that relates the subjective scores of stimuli and response. In the functional measurement approach, all these three problems are solved simultaneously (Anderson, 1970, 1971a, 1981, 1982). In order to solve these problems simultaneously, a factorial design is used. Such a design permits a test of (c) above, the law that relates the subjective scores of stimuli and response (i.e., the combination rule \( C \) in Figure 3). If the functional measurement axioms are satisfied, analysis of variance can be used as error theory, thus making it possible to test alternative model forms.

If the data support the model form, subjective stimuli \( (S_i \text{ and } S_j) \) and responses \( (\psi_{ij}) \) can be derived from the model (Veit and Ware, 1982). The data produced by factorial designs are analyzed using analysis of variance procedures. If the data generated by respondents' evaluations of each multidimensional health state obey the conditions of the model, the model is supported as an appropriate description of the combination process and the stimulus and response scale scores are separately derived from the model. They are the least squared estimates of the stimuli and response scores under the model. These subjective scale scores have substantive meaning with respect to the model that explains the responses. The additive model is supported if no interactions are present. In practice, the additive model is supported even if some interactions are present, as long as they are not statistically significant. If, however, statistically significant interactions are present, procedures are available for determining whether these interactions
can be described by a multiplicative or multilinear model form. If so, it is possible to derive stimulus and response scores (Anderson, 1970, 1971a, 1981, 1982). On the other hand, if the predictions of the model are not supported by the data, the subjective scale scores remain unknown.

Thus, the idea is that scaling is tied to the substantive theory that explains how information affects responses (Veit and Ware, 1982). This emphasis on the model form as opposed to the scale scores themselves is especially appropriate for the measurement of health status preferences, where there is no external standard of accuracy for health status preferences. As mentioned earlier, in the absence of such a standard, validation of the scores obtained rests on testing the process by which information is integrated by individuals in order to provide a response. In other words, the scale scores obtained are considered valid because the process by which they were generated is shown to be valid.

An important scaling feature of functional measurement is that knowledge of the appropriate combination function (C) allows separation of stimulus (Sj and Si) from response (vij) scaling. These two different subjective scales are typically confounded with other measurement strategies. Further, when enough ordinal constraints are built into the design (i.e., use of enough dimensions with dimension levels that are close together in value), scale scores derived from a model form are known to an interval scale in the ordinal uniqueness sense, i.e., only a linear transformation of the stimulus scales when applied to the model will reproduce the order of the data in case of additive model forms (Krantz and Tversky, 1971a). Scale scores obtained from non-additive model forms, however, yield scores
unique to only a log-interval scale (Krantz and Tversky, 1971a). Without knowledge of the model form or the design constraints, the meaningfulness of both the scale scores and their properties are relegated to definition.

The functional measurement approach is noteworthy in that it provides a way for simultaneously validating the combination rule and the scale scores. Of all the preference measurement strategies, it is the only approach which permits testing and subsequent conclusions about the level of measurement (i.e., nominal, ordinal, interval, or ratio) of scaled health states. This corresponds to the function / is Figure 3.

Anderson (1976) has explained how functional measurement can yield validated interval level scales of subjective constructs. Consider the case of two independent variables (e.g., physical function and mental function) combined in an ordinary Row x Column factorial design, where several levels of physical function constitute the rows of the design and several levels of mental function constitute the columns of the design. Each cell of the design Thus corresponds to a pair of stimuli. Respondents are instructed to judge the health status profile (i.e., combination of physical and mental function) on overall desirability by assigning a number to the stimulus combination in each cell of the design. These numbers constitute the data to be analyzed (Anderson, 1976). Depending on the type of model form or combination function hypothesized, the data analysis rests on different theorems.

If an additive model form is hypothesized, the data analysis rests on the parallelism theorem, i.e., if the additive model form is "correct" and if the response measure is on an interval scale, then the ordinary two-way
graph of the data will appear as a set of parallel lines. Given this parallelism, Anderson (1974a) has explained how the absence of interactions among stimuli validates an interval-level scale: "A priori, there is no great reason to think that ordinary ratings constitute an interval scale of response. However, if the overt response were a nonlinear function of the underlying response, then the data would not plot as parallel lines even if the model were true. Parallelism thus provides a joint validation of the psychological law, and of the response scale (p.221)." Finally, it can also be shown that the row means of the design are interval scale estimates of the subjective scores of the row stimuli (i.e., physical function in this example) and that the column means are interval scale estimates of the subjective scores of the column stimuli (i.e., mental health in this example). Moreover, the cell entries are interval scale estimates of the subjective scores of responses (i.e., health states). In this way, the observed parallelism accomplishes three goals simultaneously: (a) it supports the additive model, (b) it indicates that the response is on an interval scale, and (c) it yields interval scales of the stimulus variables (Anderson, 1976).

However, it is logically possible that parallelism may be obtained even though the additive model form is invalid and the response is not on an interval scale. If just one of these two premises is incorrect, then parallelism would not be obtained; however, if both are incorrect then parallelism would be obtained if the non-linearities in the model form effectively cancel the non-linearities in the response measure. Although this may not seem very likely to occur in practice, it is still a logical possibility (Anderson, 1976). Moreover, as Anderson (1976) has noted, the parallelism theorem
applies strictly only to the ideal case of error-free data. Since actual data usually contains some variability which would cause deviations from exact parallelism, mere visual inspection of the factorial plot may sometimes be inadequate to assess the model form. Therefore, in general, a formal statistical analysis is necessary. Ordinary analysis of variance provides a rigorous test of parallelism that takes account of the prevailing unreliability (i.e., it can serve as an error theory). Exact parallelism is equivalent to zero interaction. In practice, therefore, the F ratio for the interaction should be non-significant if the hypothesized additive model form is correct. Significant interaction would cast doubt on the hypothesized model form and/or response scale (Anderson, 1976).

If a multiplicative model form is hypothesized, the data analysis rests on the linear fan theorem. Suppose for the moment that the subjective scale scores of one of the stimuli (e.g., mental function, whose levels constitute the columns of the Row x Column design) are known and that the corresponding mental function stimuli are spaced on the horizontal axis at these subjective scores. Then, if the multiplicative model is "correct" and if the response measure is on an interval scale, the data in each row (i.e., level of physical function) should plot as a straight line function of the subjective scale scores of the column stimuli (i.e., mental function) with slope equal to the subjective scale scores of the row (i.e., physical function level) in question. The complete set of curves should form a diverging fan of straight lines (Anderson, 1976). A quick version of this graphical test can be obtained by plotting each row as a function of the other -- each such plot should be a straight line (Anderson, 1974a).
However, one vital element is lacking in this analysis -- the subjective scores of one of the stimuli (mental function in the example above) will not ordinarily be known. But, if the multiplicative model form is "correct" and the response measure is on an interval scale, then the column means of the design are interval scale estimates of the subjective scores of the column stimuli. In practice, therefore, the observed column means may be used as provisional scores of the column stimuli to test the linear fan prediction. In this way, the observed linear fan accomplishes three goals simultaneously: (a) it supports the multiplicative model, (b) it indicates that the response is on an interval scale, and (c) it yields interval scales of the stimulus variables (Anderson, 1976). Further, a rigorous statistical test is also available, using analysis of variance techniques along the lines of the test of the additive model form described above. The essential idea is that since the curves are theoretically non-parallel, the interaction term is expected to be significant. However, if the hypothesized model form is "correct," this interaction should be concentrated in the bilinear component of the interaction. The residual interaction should be nonsignificant (Anderson, 1974a).

Stimulus scaling for multilinear model forms is essentially the same as for the additive and multiplicative model forms. The marginal means of the data tables estimate the subjective stimulus scores on interval scales and the cell entries estimate the subjective response scores on interval scales. However, the validity of these scale scores depends on the "correctness" of the hypothesized multilinear model, which can be tested by analysis of variance procedures. In general, each multilinear model form implies a particular pattern of interactions, and that provides a basis for testing the
model form. This pattern of interactions can diagnose which, if any model form, is the "correct" one (Anderson, 1974a).

The major drawback of the functional measurement approach (especially in terms of this dissertation) is its reliance on judgments as the response mode. The distinction between judgment and choice had been introduced in Chapter II. A more detailed discussion on the judgment versus choice issue will be deferred until after the social/clinical judgment approach to rating scale judgment strategies is discussed, since the drawback of using judgments instead of choices as the response mode is common to both approaches.

The other drawbacks of the functional measurement approach are primarily logistical in nature. First, when there are many dimensions and levels within dimensions, the number of multidimensional judgments required of respondents in order to achieve a complete factorial design may be prohibitive. In this regard, Fischer (1979) has noted that when the number of dimensions exceeds three, the functional measurement approach may be infeasible. However, in such cases, fractional factorial designs may be employed which produce the necessary information from a smaller number of multidimensional judgments. Second, application of this approach requires technical expertise, especially in the area of experimental design and analysis of variance -- especially if multiplicative or multilinear model forms are involved (Fischer, 1979; Froberg and Kane, 1989a).
(b) Social/Clinical Judgment Analysis

In spite of the advantages of the functional measurement approach, it has seldom been used in health status preference measurement (however, see Cadman and Goldsmith, 1986; Llewellyn-Thomas, Sutherland, Ciampi, et al., 1984; Rose, 1980). Most attempts to statistically infer the parameters of MAP models have relied on less sophisticated assessment techniques, such as social/clinical judgment analysis, which are associated with Social Judgment Theory. Social Judgment Theory is based on Brunswick's Lens Model (Brunswick, 1956), which is a descriptive theory of the relationship between real stimuli, perceived stimuli, and individuals' response to each. The estimation of respondents' scale scores using the Lens Model framework is referred to as "Judgment Policy Capturing."

There are two major differences between Information Integration Theory and Social Judgment Theory -- conceptual and methodological. Conceptually, the two theories have different purposes. Information Integration Theory is intended to describe individuals' idiosyncratic methods of combining and weighting information contained in stimuli by developing mathematical equations representing their combinatorial processes or combination functions (Hoffman, 1960). The focus of this theory is therefore on the individual. Social Judgment Theory, on the other hand, is intended to describe not only individuals' methods of processing information contained in stimuli but also the manner by which individuals learn the characteristics of their environment and how the environment affects their evaluation. The focus of this theory is therefore not on the individual per se, but on the adaptive interplay between the individual and
his or her environment. Accordingly, the lens model represents the probabilistic interrelations between individualistic and environmental components of the evaluation situation (Brunswick, 1952, 1956).

This conceptual difference dictates the second point of distinction (i.e., methodological) between Information Integration Theory and Social Judgment Theory. The methodological differences between these theories are twofold. In terms of design, social judgment theorists are advocates of “representative” design, which in essence means that individuals should be studied in realistic settings, in experiments that are representative of their usual ecology. This emphasis follows from the need to incorporate the role of the environment in the evaluation process, which as noted earlier is one of the cornerstones of Social Judgment Theory. The lens model provides a means for appropriately specifying the structure of the situational variables in such an experiment. Information integration theorists, on the other hand, often make use of research designs which are not very representative of reality, for example orthogonal designs. Indeed, orthogonal designs are of limited use for health status preference measurement, given the importance of incorporating interactions between health status dimensions. It should, however, be pointed out that Information Integration Theory is not restricted to using orthogonal designs, just that several applications of this theory have used such designs. The second methodological difference between these two theories is in terms of error theory -- Information Integration Theory has an error theory while Social Judgment Theory does not. As a result of this lack of error theory, the social judgment theorist must assume (as opposed to test) the validity of the scale scores obtained as a result of social/clinical
judgment analysis. The implications of the lack of error theory in Social Judgment Theory will be discussed below in the course of the discussion on the analytical procedures used in social/clinical judgment analysis.

Hammond, Stewart, Brehmer, et al. (1975) have commented on the implications of these differences on the application of Social Judgment Theory, "... the social judgment theorist places less emphasis on mathematical precision in cognitive modeling and more emphasis on empirically demonstrating the usefulness of a given model with regard to a given problem ... (p.284)." Such a comment calls into question the usefulness of laboratory studies on information integration with regard to a given problem since, as noted earlier, these studies control for environmental influences rather than build them into the design. One of the important issues in this regard is the generalizability of laboratory results to natural settings. Several prominent researchers have raised such concerns about experimental designs (see, for e.g., Brunswick, 1956; Neisser, 1976; Rosenthal and Rosnow, 1969) and many alternatives to such designs have been proposed (see, for e.g., Campbell and Stanley, 1963; Ebbesen and Konecni, 1980; Webb, Campbell, Schwartz, et al., 1966) based on the premise that the external validity of experimental designs is questionable.

However, Levin, Louviere, Schepanski, et al. (1983) have provided several examples which provide direct and indirect tests of the external validity of such designs. Levin, Louviere, Schepanski, et al. (1983) defined direct tests as those in which laboratory-derived models successfully predict or account for actual choice behavior and indirect tests as those in which the parameters of the laboratory-derived models are shown to be logically and
predictably related to factors external to the task. On the basis of these examples, it seems appropriate to note that properly designed experimental studies can be generalized to natural settings. Given this observation, the major distinction between functional measurement and social/clinical judgment analysis lies in not in their conceptual differences but their methodological differences, particularly the method of analysis.

Under social/clinical judgment analysis, correlation and regression procedures have mainly been used as the method of analysis to obtain some understanding of the combination rule (C in Figure 3). Respondents are usually asked to evaluate a set of multidimensional health states on a continuous response scale and the subjective weights and scale scores of a simple multidimensional preference model (usually an additive model) are estimated using regression procedures. A common formulation of a subjective model is (Veit and Ware, 1982):

$$\Psi_{1\ldots i\ldots n} = \sum w_i s_i + c; \ i = 1 \ldots \ n,$$

(3.1)

where

$$\Psi_{1\ldots i\ldots n} = \text{subjective response to the } n \text{ stimuli presented to the respondent for judgment}$$

$$w_i = \text{subjective weights associated with the } i^{th} \text{ stimulus}$$

$$s_i = \text{subjective score associated with the } i^{th} \text{ stimulus}$$

$$c = \text{additive constant.}$$
Different combination rules can be obtained by raising the $s$ values to powers, multiplying various combinations of the scale scores together (i.e., adding interaction terms), etc. The standardized partial regression coefficients (beta weights) are often interpreted as measures of the relative importance of predictors ($s_i$) in estimating the dependent variable ($y_{ij...i...j}$). Conclusions regarding the appropriateness of the multidimensional preference model are usually based on an index of goodness of fit, such as the magnitude of the multiple correlation coefficient ($R^2$). If $R^2$ is "acceptably high" (e.g., 0.7 or above is a cutoff recommended by Fischer, 1979), it is usually concluded that there is a good degree of correspondence between the model-generated preferences and the respondents' multidimensional judgments and the model is deemed appropriate.

Regression analysis typically proceeds under two assumptions, that the scale scores are known and that the observed response is on an interval scale. These assumptions are especially problematic since Social Judgment Theory has no error theory -- hence, one must assume that regression models are adequate descriptors of rating scale judgments of multidimensional alternatives. The problems associated with these assumptions are best illustrated in the case of a linear regression model not fitting the data well. In such a case, nonlinearity may be nothing more than an inappropriate stimulus scale, for if the ostensible stimulus scale is not an interval scale then a linear function of the true subjective scores will appear to be nonlinear. Similarly, apparent nonlinearity may not represent true interaction but only an inappropriate response scale (Anderson, 1961, 1962; Bogartz and Wackwitz, 1970). Without an error theory to support these scaling
assumptions, interpretations in terms of nonlinearity rest on uncertain ground.

Moreover, this approach does not test the validity of the scale scores. The model form and scale scores shown in the common formulation above are subjective in nature and are therefore unknown. Since these scale scores are unknown, researchers using multiple regression analysis typically employ direct scaling approaches (which will be discussed in the section of this chapter which deals with scaling approaches) to derive proxy scores for these subjective scale scores and use them in the testing of the model form. However, the relationship between the scores used in the model (i.e., $i$ and $j$ in Figure 3) and their subjective counterparts (i.e., $S_i$ and $S_j$ in Figure 3) are simply defined to be the same. In other words, $H$ in Figure 3 is assumed to be an identity function. Thus, at the outset, scores with unknown validity are used as inputs into the model. This technique does not provide a way to determine the validity of these input scores. Whereas the functional measurement approach incorporates scaling as an integral part of testing the underlying information-processing framework, the social/clinical judgment approaches do not provide a way to determine the validity of the scale scores. Instead, these approaches simply assume these scores are valid and represent interval measures and use them to test the combination rule.

Further, $R^2$ is not an adequate test of model form, since it can be high even when deviations from model predictions are significant and systematic (Veit and Ware, 1982). It has been demonstrated that important interactions can exist with an $R^2$ as high as 0.98 for an additive model (Anderson and Shanteau, 1977). Thus, it is possible that the researcher may "confirm" an
incorrect model form (and scale scores) using such indices of goodness of fit. In this regard, the criticisms of the multiple correlation coefficient discussed under rank order judgment strategies also apply to social/clinical judgment analysis.

The technique used by some researchers for comparing goodness of fit indices across model forms (i.e., different forms of the multiple regression model) presents additional problems (Veit and Ware, 1982). First, the number of possible models that could be entertained is infinite. There is no way of knowing if the correct model form has been included as one of the forms being tested. Moreover, even if it was included, it may be rejected because of the above problem of "confirming" an incorrect model form. Birnbaum has demonstrated how $R^2$ values can be higher for an incorrect as opposed to a correct model form when the input scores for the correct model form are not the actual scores of the model form (Birnbaum, 1973, 1974). Such comparisons are especially problematic if the competing model forms are compared in terms of their correlations with the data, because estimation properties may not be comparable across model forms. For instance, an incorrect model form that required three estimated parameters could yield a higher correlation than a correct model form that required only two estimated parameters (Anderson, 1970). Moreover, correlations depend on the range and reliability of the predictor variables. Unless these are equal for the different model forms being compared, the correlations are not comparable. An incorrect model form could yield a higher correlation than the correct model form simply because of having greater range and/or greater reliability in the predictor variables (Anderson and Shanteau, 1977).
Finally, the fact that attributes or dimensions are often correlated and interact with each other further obscures the meaning of the multiple correlation coefficient as well as any interpretations of regression coefficients indicating the relative importance of the dimensions (Froberg and Kane, 1989a; Veit and Ware, 1982). As Darlington (1968) has noted, interpreting regression coefficients as measures of importance is dangerous when dimensions are correlated because their magnitude and even their rank order in magnitude can change with changes in the alternatives under evaluation (and thus the dimension intercorrelations). Using statistical controls like partial correlations will not help in this case because such statistical controls cannot unconfound variances that are empirically confounded (Veit and Ware, 1982). Further, given the importance of incorporating interactions for health status preference measurement (as discussed earlier), the use of orthogonal designs is of little use in such cases.

These criticisms should not detract from the usefulness of regression procedures for the purpose of prediction. As Yntema and Torgerson (1961) have noted, systematic discrepancies from a simple additive model form may not be serious in certain instances. Indeed, simple additive models have been shown to be fairly accurate as predictors of a variety of types of decisions (Slovic and Lichtenstein, 1971). However, good prediction does not imply good explanation (Hunt, 1987). When the main concern is to understand the psychological processes that underlie evaluations, any systematic discrepancy may be meaningful and important. As noted, in the absence of an external standard of accuracy for health status preference measurement, the process by which health status preferences are arrived at becomes
important in providing meaning to health status scores. Hence, the focus in such cases should be explanation of the underlying decision making process, not prediction.

It should also be noted that it is possible to apply regression analysis in all instances where analysis of variance can be applied. This is because analysis of variance and multiple regression are both applications of the general linear hypothesis of mathematical statistics and have considerable similarity (Cohen, 1968). However, the typical regression design corresponds to a highly confounded factorial design -- therefore discrepancies from prediction are generally difficult to localize and interpret. Moreover, nonindependence of the regression coefficients can complicate stimulus scaling (Darlington, 1968). On the other hand, regression designs can be very efficient, since they require far fewer observations than complete factorial designs. However, the use of fractional factorial designs somewhat nullifies this advantage of regression analysis over analysis of variance procedures.

In summary, the rating scale judgment strategies -- especially the functional measurement approach -- offer significant improvements over their rank order counterparts. Unlike the rank order judgments, these strategies permit powerful tests for model form and adequacy, provided that the rating data are measured on an interval or ratio scale. The advantage of the functional measurement approach over the social/clinical judgment approach is that it actually allows a test of the underlying decision making process and level of measurement of the ratings. Thus functional
measurement is an extremely powerful approach to simultaneously validate the model form and the scale scores derived from the model form. Indeed, rating scale judgment strategies are more popular than rank order judgment strategies among commercial users of conjoint methodology (Wittink and Cattin, 1989). Having noted this, it should be pointed out that most marketing research applications of conjoint methodology focus on predictive validity (i.e., prediction of responses to a holdout sample or prediction of responses to a new stimulus and compare this predicted response to actual behavior) as opposed to explaining the underlying decision making process followed by responses. For this purpose, the two strategies have produced very similar results at both the individual as well as the aggregate level in published studies which have compared the two strategies (Green and Srinivasan, 1978).

However, both rank order and rating scale judgment strategies have limited usefulness in explaining and/or predicting choice from rankings or ratings data respectively. The issue of judgment versus choice is particularly important to this dissertation, and was introduced in Chapter II. As noted earlier, Louviere (1988a) has drawn a distinction between judgmental data and choice data. According to him, judgment data are evaluative rankings or ratings of a set of multiattribute or multidimensional alternatives obtained from individuals or groups of individuals, which are assumed to be at least ordinal in measurement level. On the other hand, choice data are responses that identify one and only one of a set of alternatives as the “highest,” “best,” etc. or responses which involve the allocation of fixed sets of resources to a set of alternatives so that the individual allocations sum to the
total fixed amount (Louviere, 1988a). Choice data need only be measured at a nominal level of measurement.

Louviere (1988a) has noted that the distinction between judgmental and choice data is important because judgmental data may not contain information about choice behavior and may not satisfy various assumptions necessary to explain and/or predict choice behavior. Further, it should be noted that there is no adequate theoretical basis for relating respondents’ judgments to choices. Although the literature on the effects of response mode (and the judgment versus choice distinction) on preference measurement will be reviewed in Chapter IV, at this point it is necessary to note that preference measurement researchers realized the implications of the judgment versus choice distinction and made some attempts to account for it in their analyses. These attempts mainly involved the use of choice simulation methods.

**Choice Simulation Methods**

Choice simulation methods were developed as an attempt by some researchers to avoid abandoning rank order and rating scale judgment strategies altogether in light of the distinction between judgment and choice. These researchers essentially used simulation methods as an adjunct to judgments in order to predict individual choices on the basis of the observed judgments collected from respondents. It should be noted that the word “simulation” is used here to mean prediction of individual choices under hypothetical scenarios and subsequent aggregation of choices -- it does not
necessarily mean Monte-Carlo simulation. These choice simulation methods are generally of two types -- deterministic and probabilistic.

**Deterministic Simulation Methods**

Deterministic simulation methods operate on the "first choice rule," which involves estimating respondents' part-worths for each alternative to identify the most preferred alternative in the choice set (i.e., the alternative which has the highest estimated worth) and assuming that respondents would choose this most preferred alternative if asked to make a choice from the choice set in question (Curry, Louviere, and Augustine, 1981, 1983; Green, Carroll, and Goldberg, 1981; Green DeSarbo and Kedia, 1980). This first choice rule has been used in several conjoint choice simulations (see, for e.g., Finkbeiner, 1988; Green, DeSarbo, and Kedia, 1980; Urban and Hauser, 1980) as well as in several algorithms for optimal new product identification (see, for e.g., Albers, 1979; Gavish, Horsky, and Srikanth, 1983; Green, Carroll, and Goldberg, 1981; Sudharshan, May and Shocker, 1987).

Although the assumption that respondents will always choose the alternative with the highest estimated worth may be intuitively appealing, Louviere has noted that it is incorrect in two respects (Louviere, 1988a). These two respects may be thought of as two sources of error -- error in linking estimated preferences from judgmental data to choices (i.e., choice error) and error in estimation of preferences (i.e., estimation error). Regarding choice error, the conjoint attributes and part-worths do not fully account for the preferences governing choice behavior. Several studies have demonstrated that individuals do not always choose their most preferred
alternative in a choice set (see, for e.g., Bass, Pessemier, and Lehmann, 1972; McAllister and Pessemier, 1982; Morrison, 1979; Reibstein, 1978; Silk and Urban, 1978). The discrepancies between the preferences governing choice and the conjoint predicted part-worths will be larger as the difference between the actual choice task and alternatives differs from the conjoint task and hypothetical profiles. Therefore, an error term is required to capture these discrepancies. However, the first choice rule assumes no error term is needed.

Regarding estimation error, the first choice rule ignores the fact that there is error in estimation of respondents' part-worths. It assumes that the alternative with the highest predicted worth is the one with the highest actual worth. However, since the part-worths are estimated with error, the alternative with the highest predicted worth may not always be the one with the highest actual worth. Using the first choice rule then may lead to erroneous conclusions about choices. Indeed, Elrod and Krishnakumar (1989) have shown that using the first choice rule in choice simulation methods may lead to biased predictions of choice.

Using estimated part-worths to simulate expected choices of a sample of individuals is also problematic because of the assumptions involved in the procedure. Specifically, one must assume: (a) individuals are perfectly transitive and consistent in their rankings or ratings, (b) there is no bias in the estimated part-worths, (c) individuals are aware of all the alternatives and possess perfect information about their attributes, and (d) there are no income, time, or other situational constraints to cause individuals to choose an alternative other than that with the highest predicted preference
(Louviere, 1988a). Further, as Louviere (1988a) has noted, it is logically inconsistent that stochastic models are used to derive estimates of part-worths (e.g., linear regression models) but deterministic rules are used to apply the parameters to predict choice.

This logical inconsistency may be avoided by using the second (i.e., probabilistic) simulation method or by making assumptions about the statistical properties of the estimated part-worths. Probabilistic simulation methods are discussed below. Regarding making assumptions about the statistical properties of the estimated part-worths, if respondents' estimated part-worths satisfy the scale properties of a choice model like the MNL model and the standard error of the regression model used to estimate these part-worths satisfies the variance assumptions in the error distributions used to derive the MNL model, one can use the predicted preference scores to forecast choice probabilities for different alternatives (Louviere, 1988a). Louviere (1988c) has compared the parameters of MNL models estimated from simulated choice data using the first choice rule and the assumption that the predicted part-worths satisfy the scale properties of the MNL model. He found little difference in the MNL parameters so estimated (Louviere, 1988c).

However, making such assumptions about the statistical properties of the estimated part-worths may involve making contradictory assumptions in the estimation of part-worths and the prediction of choices from these estimated part-worths. Consider the usual case where ordinary least squares (OLS) regression is used to estimate part-worths and it is assumed that these estimated part-worths satisfy the properties of the MNL model. The problem
is that the assumptions required to use OLS regression are contradictory to those of the MNL model. Specifically, in OLS regression, the relationship between the dependent and independent variables is assumed to be linear and the impact of an independent variable on the dependent variable is seen as constant over the total range of values. However, in the MNL model, the underlying relationship between the dependent and independent variables is assumed to be a nonlinear slanted S-shaped function or a sigmoid curve. This implies that the independent variables have the greatest impact on the dependent variable at some midpoint, where the slope of the curve is the greatest and have the least impact at the ends of the range, where the slope of the sigmoid curve is less (Morrow-Howell and Proctor, 1992). Given these contradictions, making assumptions about the statistical properties of the estimated part-worths may have weak theoretical foundations. These drawbacks have prompted interest in the second type of choice simulation methods -- probabilistic simulation methods.

Probabilistic Simulation Methods

Probabilistic simulation methods attempt to develop choice models and choice forecasts from ranking or rating data (Chapman, 1984; Chapman and Staelin, 1982; Hensher and Louviere, 1983). Although several probabilistic choice rules can be applied at the individual and aggregate level, all recognize that respondents will probably choose the alternative having the highest predicted worth. These choice rules differ in the exact method by which they estimate this probability.
Although these methods address the logical inconsistency problems of deterministic methods, they still require the following assumptions to be satisfied by individuals' rankings or ratings: (a) the chosen choice model (usually the multinomial logit choice or MNL model) is a good approximation of the unobserved choices implied by the rankings or ratings, (b) the individual is perfectly transitive in the unobserved choice sets implied by the rankings or ratings, (c) the individual is perfectly consistent in his or her ranking or rating behavior in the unobserved choice sets implied by the rankings or ratings, and (d) individuals' preferences as indicated by their ranking or rating data will be reflected in their unobserved choices (Louviere, 1988a).

The major problem with these assumptions in choice simulation methods is that they cannot be tested, since real choices are not observed in the studies using these methods. Instead, models are estimated from simulated choices based on an explosion of the rankings or ratings (Chapman, 1984; Chapman and Staelin, 1982) or based on deterministic expressions from the marginal frequencies of choices of each alternative -- the frequencies which result from applying an individual's ranking or rating to all possible choice sets (Hensher and Louviere, 1983).

Although these assumptions cannot be tested in the studies since actual choices are not observed, empirical tests of these assumptions in other studies have provided less than encouraging results. As Tversky (1972) has demonstrated, individuals frequently violate the assumptions required to translate preferences as indicated by ranking and rating data into choices. In particular, individuals can (and do) use elimination and nesting strategies
which violate the independence of irrelevant alternatives (IIA) assumption of the MNL model. (The IIA assumption holds that individuals' preferences for a particular alternative are unaffected by preferences for competing alternatives in the choice set.) Further, as Louviere (1988a) has noted, it is unrealistic to assume that individuals' rankings or ratings will perfectly correspond to the choices that they would make if faced with subsets of the ranked or rated alternatives varying in size and composition. Individuals' rankings or ratings of a single set of alternatives are thus unreliable estimates of their choices in subsets of the full choice set. In light of these findings, Louviere (1998a) has therefore recommended that caution be exercised before applying these simulation methods to estimate the parameters of the choice models under consideration.

Finkbeiner (1988) has reported a comparison of the predictive validity of deterministic and probabilistic simulation methods in terms of their predicting choices from holdout samples in conjoint studies. He found that the first choice rule yielded the best estimates of choice for both of two holdout samples, as compared to five probabilistic rules. This result is in agreement with other studies, which have found that the first choice rule has acceptable predictive validity in practice (Braun and Srinivasan, 1975; Huber and Moore, 1979; Parker and Srinivasan, 1976; Sudharshan, May, and Shocker, 1987). Elrod and Krishnakumar (1989) have argued that although this result may be true in certain instances, it is not always so. More specifically, they have shown that under certain conditions the first choice rule may yield substantially biased results, while under other conditions it
may yield unbiased ones. They have also shown that the same argument holds for probabilistic simulation methods. They have described these conditions with respect to the two sources of error introduced earlier -- choice error and estimation error (Elrod and Krishnakumar, 1989).

According to Elrod and Krishnakumar (1989), both choice simulation methods yield nearly unbiased results when there is little estimation and choice error. However, as they themselves note, the conditions which need to be met for having less choice and estimation error are not only difficult to satisfy in practice but are also difficult to satisfy simultaneously. Regarding the first choice rule specifically, it yields nearly unbiased results when there are approximately equal amounts of choice and estimation error. However when the magnitude of one source of error is greater than the other, the first choice rule yields biased results. In such cases, probabilistic simulation methods can reduce the bias if there is little or no estimation error and a substantial amount of choice error. However, if there is much estimation error and little or no choice error, the probabilistic simulation methods will yield more biased results than the deterministic methods (Elrod and Krishnakumar, 1989).

However, these conclusions are only appropriate for the purposes of prediction, not explanation of the decision making processes that underlie respondents' choices. In this regard, neither of the choice simulation methods represent this decision making process, and therefore are of little use in explaining the choices made by respondents. Given the distinction between explanation and prediction discussed earlier and the importance of explanation over prediction for the purpose of providing meaning to health
status preference measures, this constitutes a serious drawback of choice simulation methods for the purpose of health status measurement.

It should also be noted that both the choice simulation methods -- deterministic as well as probabilistic -- proceed on the assumption that respondents will always make a choice. However, in real markets individuals often decide not to choose any alternative or decide to delay their choice. In this regard, these choice simulation methods may be unrealistic in that they do not reflect such actual behavior in the market. Moreover, awareness of alternatives influences the formation of a choice set, and it is not easy to incorporate measures of future awareness into choice models if one has no information except rankings or ratings. Thus the researcher must assume that all individuals have identical choice sets; otherwise the researcher must develop a way to forecast what a choice set will include.

Therefore, if a major objective of research conducted using conjoint methodology is to explain and/or predict choices among competing alternatives, rank order and rating scale judgment strategies may be of limited use. On the other hand, choice data by definition contain information about individuals' choice behavior, but one must make assumptions about which of several possible decision making processes underlie the data (Louviere, 1988a). If such assumptions can be made, conjoint strategies using choice data may be more useful than rank order or rating scale judgment strategies in predicting choices between alternatives. It is in this regard that discrete choice conjoint strategies are helpful.
Discrete Choice Conjoint Strategies

Discrete choice conjoint strategies were developed by Jordan Louviere and his colleagues in the late-1970s to early-1980s for the purpose of overcoming the drawbacks of rank order and rating scale judgment strategies in dealing with choice as the dependent variable in conjoint methodology (see Louviere, 1983, 1984; Louviere and Hensher, 1982, 1983a, 1983b; Louviere and Meyer, 1981; Louviere and Woodworth, 1983). This section will begin by introducing discrete choice conjoint strategies. This will be followed by a discussion of the theoretical rationale behind these strategies, after which issues relevant to the design and analysis of such strategies will be discussed. The section will conclude with an evaluation of the strengths and weaknesses of these strategies.

Introduction

Discrete choice conjoint strategies represent the culmination of developments in psychometric and econometric theory, in that they integrate the strengths of functional measurement (developed in the field of psychometrics) with those of discrete choice modeling (developed in the field of econometrics).

Since the principles of functional measurement have already been discussed, they will not be repeated here. In discrete choice modeling, attention centers upon the choices made by one or more individuals between competing alternatives. These choices are assumed to be driven by a preference function which may be characterized by a systematic or estimable component and a random or unobservable component. Applied econometric
choice models typically assume that all individuals (or a relatively homogenous subset of individuals) share a common or representative preference function but have idiosyncratic tastes and/or unobserved influences on choice. Such “disturbances” are assumed to be independent and drawn from a particular type of distribution (Louviere, 1983). As Amemiya (1981) has noted, “… what kind of QR (Qualitative Response) model one gets is equivalent to what distribution one assumes for (the differences in the errors).” For example, assuming a double exponential distribution leads to the Luce or MNL model, while assuming a normal distribution leads to the Thurstone Comparative Judgment or probit model (see Manski and McFadden, 1981).

There are two general types of discrete choice models -- binary (or binomial) and multinomial. Binary discrete choice models focus on the choices made by individuals between only two competing alternatives. The method of paired comparisons and its various derivatives often serve as the frameworks for collecting choice data in such cases. Multinomial discrete choice models, on the other hand, focus on the choices made by individuals between more than two competing alternatives. Choice data may be collected by several different methods in such cases, as explained later on in this section. For the purpose of this dissertation, the attention is on multinomial as opposed to binary discrete choice models. This focus reflects the situation commonly seen in practice -- there are usually more than two alternatives among which to choose. Specifically with respect to the topic of this dissertation, when choosing among antihypertensive agents on the basis
of their effects on health status, there are usually more than two agents among which to choose.

Although popular and relatively easy to implement, paired comparison choice experiments have limited use for modeling and understanding multiple choice behavior. Their primary limitation is theoretical -- it is difficult to generalize multiple choices from binary choice data unless the choice process has a fairly simple structure. Sometimes paired comparison experiments are augmented with triples or quadruples, but such additions often are ad hoc, and may not satisfy the statistical assumptions of multiple choice models and further possess unknown and uncertain efficiency properties. Hence, as Louviere (1992) has noted, such experiments may (and often do) lack the power to test or diagnose differences that make a difference in model comparisons. For example, Louviere and Woodworth (1983) have demonstrated that paired comparison experiments are significantly inferior to experiments based on fractions of $2^N$ factorials (which are discussed later) even in case of simple choice models.

Theory

Discrete choice conjoint strategies achieve the integration of functional measurement and discrete choice theory through the design of interlocking fractional factorial experiments (Louviere and Hensher, 1983a). These experiments are used to generate multiattribute alternatives (as in functional measurement) and to place these alternatives into choice sets from which respondents make choices (as in discrete choice modeling). In other words, discrete choice conjoint strategies use experimental designs which consist of
a series of choice alternatives and a series of choice sets in which these alternatives appear. Since there are two separate designs involved (i.e., choice sets and choice alternatives), the overall experimental design is referred to as a Double Conditional Design, because the choice responses are conditional on both, the functional measurement treatment combinations as well as the choice set generating design. It is generally necessary to design at least two interlocking experiments to manipulate both the multiattribute alternatives as well as the choice sets in which they appear. This permits the researcher to design choice experiments sufficient to estimate the parameters of the choice model assumed to underlie the decision making process (e.g., MNL, probit, etc.). Hence, one can develop models commonly used in conjoint methodology from a discrete choice task, and these models can be used to explain and/or predict respondents choices rather than judgments, i.e., rankings or ratings (Louviere and Hensher, 1983a).

Discrete choice conjoint strategies operate on the following principle - if one assumes that the choices of a reasonably homogenous behavioral segment of individuals can be closely approximated by some form of choice model (e.g. the MNL choice model), it is possible to estimate the unknown part-worths in a statistically efficient manner from experimental designs. In this way, the choice model (i.e., the MNL choice model in the above example) provides the statistical techniques for analyzing the choice data and thus serves as an error theory to diagnose or test alternative model forms if the choice experiment is designed in such a way so as to accommodate the required tests (Louviere, 1988a).
The first step in using discrete choice conjoint strategies is deciding which choice model to adopt as representative of respondents' decision making process. Excellent reviews of the different probabilistic models of individual choice are provided by Corstjens and Gautschi (1983), McFadden (1980, 1981, 1986), and Meyer and Kahn (1991). The overview provided in this section of these probabilistic choice models is rather brief, since the only purpose of the overview is to briefly introduce the other probabilistic choice models that one could use as part of a discrete choice conjoint strategy. At the end of the overview, one particular probabilistic choice model will be used as an exemplar for the rest of this section. This will be the same probabilistic choice model that was assumed to represent this study's respondents' decision making processes -- the multinomial logit (MNL) choice model.

Meyer and Kahn (1991) have broadly divided probabilistic choice models into two categories -- utility maximizing models and heuristic elimination models. The guiding premise underlying this categorization is their view of the field as consisting of two intellectual traditions of models, each reflecting different views of the ability of individuals to process information. At one extreme are the utility maximizing models, which hold the view that individuals make choices by considering all relevant information available to them at the time of choice and that individuals choose that alternative which maximizes some utility (actually preference, according to the clarification in terminology provided in Chapter II) function defined across this information set. Central to the basic forms of such models is an assumption that preferences for items can be defined independently from
the set of options under consideration, and that any errors in the measurement of preferences are also independent of the consideration set. Termed simply scaleable choice models, they are best illustrated by the Luce (1959, 1977) model and the Multinomial Logit or MNL (McFadden, 1974) model (Meyer and Kahn, 1991).

At the other extreme are the heuristic elimination models, which are based on the view that individuals are inherently limited in their ability to process information and therefore make choices through simplified heuristics, which in turn do not make use of all the information available to them at the time of choice. Central to these models is the assumption that preferences are inherently context dependent -- specifically, these models assume that the likelihood that an option will be chosen is a function of both the attractiveness of its features as well as the extent to which these features are shared by other alternatives under consideration. These models are best illustrated by the attribute elimination models of Restle (1959) and Tversky's Elimination By Aspects or EBA (Tversky, 1972) model (Meyer and Kahn, 1991).

The tradeoff between these two broad categories of models is one of analytical tractability versus descriptive validity -- the utility maximizing models are easier to estimate empirically but make some assumptions which are usually violated in practice, while the heuristic elimination models are rich in explanatory power but are very difficult to estimate empirically. Still, it is interesting to note that the dichotomization between the two broad categories of probabilistic choice models seems to be blurring, since both types of models seem to be moving toward a middle ground -- specifically,
research on utility maximization models is focused on attempts to make them more sensitive to the effects of choice context, while research on heuristic elimination models is focused on making them more analytically tractable (Meyer and Kahn, 1991).

Within the broad category of utility maximization models, two general approaches have been taken to relax the assumption of context independence. The first approach retains the assumption that preferences can be defined as independent of context, but relaxes the assumption of independence among errors in the measurement of preferences. Examples of models adopting such an approach include the Generalized Extreme Value or GEV (see McFadden, 1980) model and the Covariance Probit (Hausmann and Wise, 1978) model. The second approach relaxes the assumption that preferences can be defined as independent of context, but retains the assumption of independence among errors in the measurement of preferences. Examples of models adopting such an approach include the differential cross-effects choice models of Batsell and Polking (1985) and Cooper and Nakanishi (1988). Another avenue of extension of the utility maximization models is the treatment of temporal effects -- dynamics have been handled both through the addition of lagged effects in preferences of alternatives (while assuming that errors are not time-dependent) and by allowing time dependencies in measurement errors (Meyer and Kahn, 1991).

Research on heuristic elimination models has focused on ways to approximate a general sequential choice model (i.e., EBA) by simpler forms which are more easy to estimate empirically. Considerably less work has
been done on adding greater generality to EBA models, e.g., allowing them to account for learning effects (Meyer and Kahn, 1991).

In the following discussion of discrete choice conjoint strategies, the MNL choice model is assumed to represent the underlying decision making process of respondents. It was decided to discuss discrete choice conjoint strategies with respect to the MNL choice model because this is the model which will be used in this study. Further, although some of the assumptions of the MNL choice model may not be satisfied in practice, Louviere and Woodworth (1983) have noted that the MNL choice model is yet more tractable than other choice models (e.g., multinomial probit) and requires considerably less a priori information to make estimation of model parameters manageable than other choice models (e.g., nested logit). Adding to the appeal of the MNL choice model is the fact that it is linked to a number of other well-known models for discrete data analysis, such as the Bradley and Terry (1952) models in statistics and the Luce (1959, 1977) model in psychology for the scaling of paired comparison data, and the models used for the analysis of market share data in marketing (e.g., Nakanishi and Cooper, 1974). Moreover, it is the most commonly used of all the discrete choice models, with published applications in several disciplines including consumer behavior, transportation, psychology, and econometrics (see, for e.g., Batsell, 1980; Batsell and Lodish, 1981; Hensher and Johnson, 1981; Louviere and Woodworth, 1983; Louviere and Hensher, 1983a, 1983b; Reibstein, 1978). In all these applications it has been shown to provide a good account of the relationship between the attributes or dimensions of sets of alternatives and the choices made from the sets of these alternatives.
The MNL choice model is derived from random utility theory in econometrics and psychology (Amemiya, 1981; Manski and McFadden, 1981; McFadden, 1974, 1980, 1981; Theil, 1969), although its roots can be traced to Thurstone's (1927a) Law of Comparative Judgment (see Yellot, 1977). Random utility models are concerned with choices between two or more alternatives. As noted previously, individuals have an unknown "true" preference or worth function which may be characterized by a mean value (i.e., the systematic or estimable component which can be inferred from a series of observations of choice) and a random or unobservable component. Therefore,

\[ U_i = V_i + e_i \]  

where

- \( U_i \) = unknown preference or worth of interest for choice alternative \( i \)
- \( V_i \) = systematic, observable component or mean value of choice alternative \( i \) and
- \( e_i \) = random error component associated with choice alternative \( i \)

Following Lancaster (1966, 1971) and Knight and Menchik (1976), each alternative is considered to be a "bundle" of attributes or dimensions. Thus, the systematic, observable component of the preference function may
be written as a function of the levels of the different dimensions of the attributes:

\[ V_i = f[v(x_{ik})] \]  \hspace{1cm} (3.3)

where

- \( V_i \) = systematic, observable component or mean value of choice alternative \( i \)
- \( v(x_{ik}) \) = scale score or part-worth of the \( k \)-th attribute \( x \) for choice alternative \( i \) and
- \( f \) = function or combination rule defined over the \( v(x_{ik}) \) that map the scale scores into the mean value \( V_i \)

The probability expression of interest is that of the maximum worth to be found in some set of alternatives,

\[ P(i/A) = P[V(i) + e(i) > V(a) + e(a) > \ldots > V(N) + e(N)], \]  \hspace{1cm} (3.4)

for all \( j \) in \( A \).

The above equation states that the probability of selecting a particular alternative \( i \) from a given choice set \( A \) of which \( i \) is a member is equal to the probability that the mean value of \( i \) plus its random error is greater than the mean values and associated errors of all the other \( j \) alternatives in set \( A \). It should be noted that this equation is different from the first choice rule (which was discussed in relation to the deterministic choice simulation
methods) on account of the random error terms, which allow for estimation error and also the fact that it is stated in probabilistic as opposed to deterministic terms, which allow for choice error.

As noted earlier, the particular choice model selected as representative of the underlying decision making process depends on the assumptions made about the error distribution of the \( e \)'s in the above equation. The assumption used to derive the MNL choice model is that the errors are independent and identically distributed extreme value type I random variates with mean zero and constant variance. This type of distribution is known as an extreme value type I distribution because one is interested in the maximum of a series of random variables (i.e., the extreme value). This distribution is also known as the double exponential distribution on account of the ensuing choice model form, as shown below.

On the basis of this assumption (and after some algebra), the MNL choice model may be represented by the following equation:

\[
P(i/A) = \frac{e^{V(i)}}{\sum e^{V(j)}},
\]

for all \( j \) in \( A \)

where

\[
e = \text{natural constant, the base of natural logarithms} = 2.7183
\]

and all other terms are as previously defined

This equation states that the probability of choosing a particular alternative \( i \) from a given choice set \( A \) (i.e., the relative frequency with which
the alternative $i$ is chosen by a group of individuals offered the choice set $A$ is proportional to the ratio of the exponential preference or worth of $i$ relative to the sum of all the exponential preferences or worths in the choice set $A$. Louviere (1983) has demonstrated how this equation can be related to the choice behavior of respondents and thus deduce the expected form of the behavioral equation, which expresses the expected functional relationship between overall preferences for an alternative and choice behavior associated with that alternative (controlling for choice context).

In order to make the model complete, it is necessary to specify a functional form of the preference function [$f$ in equation (3.3)], i.e., additive, multiplicative, or multilinear. This is necessary to draw inferences about the effects of various attributes of interest. Louviere and Woodworth (1983) have recommended a “linear in the parameters and additive” model form, on account of the fact that this model form is not restrictive because one can define nonlinear marginal relationships and also nonadditivities or interactions. However, such a model form is not the only one which can be estimated. The MNL choice model has a closed form and can be estimated using a variety of estimation techniques (as explained later in this section) depending on whether the choice data are discrete or aggregated frequencies of choices. Further, the estimates derived from most estimation techniques are unique. In this way, MNL serves as an error theory and one can make statistical inferences about the parameters of the preference function, thereby allowing empirical testing of alternative model forms.

For example, if a strictly additive model form is assumed, a simple test is to design the choice alternatives so that the various interaction effects
can be tested. If the strictly additive model is approximately true, these interaction effects should be nonsignificant. Hence, the presence of statistically significant interactions would permit one to reject the hypothesis of additivity. Such choice experiments permit the researcher to make inferences about the parameters of preference models, and in so doing reject incorrect model forms. In the spirit of functional measurement, then, the parameter estimates gain meaning as a consequence of the validation of the model form.

Design

Louviere and his colleagues have formalized the necessary and sufficient conditions which experimental designs must meet to satisfy the statistical requirements of MNL choice models. As noted previously, the MNL model involves the IIA assumption, which essentially postulates that individuals' preferences for a particular alternative are unaffected by preferences for competing alternatives in the choice set. Conjoint methodology offers a useful way to experimentally test whether this assumption is valid or not. Indeed, McFadden (1986) -- who is credited with developing the MNL model -- has observed, "In a conjoint study it is possible to construct powerful direct tests of IIA by observing choices from set C and subsets A, ... (p.293)." Louviere and his colleagues have suggested experimental designs for MNL models that satisfy the IIA assumption as well as designs to generalize MNL models to account for violations of the IIA assumption (see Louviere, 1983, 1984; Louviere and Hensher, 1982, 1983a; Louviere and Meyer, 1981; Louviere and Woodworth, 1983).
All that is needed to fully specify MNL choice models that satisfy the IIA assumption is the choice probabilities of each alternative. This can be seen in the double exponential equation (given above), in which the V's are completely defined by the choice probabilities for any alternative. Thus, a necessary and sufficient condition for efficiently estimating MNL models is to be able to estimate these marginal probabilities independently. Louviere and Woodworth (1983) have shown that many factorial and fractional factorial designs commonly used in rank order and rating scale judgment strategies can be used to construct a series of choice sets that guarantee independence of marginal probabilities by design. They have also demonstrated that a $2^N$ (where $N$ is the total number of choice alternatives) factorial design can be used to put $N$ choice alternatives into choice sets such that the parameters of the MNL choice models -- estimated from choices made in response to the design -- will be near-optimal in statistical efficiency (Louviere and Woodworth, 1983). While other types of factorial designs may also be used to construct series of choice sets that satisfy the independence of marginal probabilities condition, the efficiency properties of these designs have not been studied (Louviere, 1988a).

Discrete choice conjoint experiments using $2^N$ designs are usually constructed in two steps: (a) one first creates a set of multiattribute conjoint alternatives using a factorial or other design to produce combinations of levels of attributes (i.e., creation of the alternative set) and (b) one next treats each of the designed multiattribute alternatives as a two-level factor -- present or absent -- in a choice set. A $2^N$ fractional factorial design is then used to generate combinations of “present/absent” (i.e., creation of choice
set). Orthogonal $2^N$ designs have the property that the presence/absence of each alternative in/from a choice set is independent of the presence/absence of other alternatives. Therefore, constructing choice sets by using $2^N$ designs to place the $N$ designed alternatives into choice sets according to which of the $N$ are "present" satisfies the independence of marginal probabilities condition needed for the efficient estimation and testing of parameters of MNL choice models.

The MNL choice model, if true, requires only that a main effects plan of choice sets be drawn from the full factorial of choice sets. This is because the MNL choice model only estimates marginal probabilities; all joint probabilities are assumed to be constant. This is a consequence of the IIA assumption, which in turn gives rise to the constant cross-elasticity assumption of the MNL choice model. Thus, all effects of interest in the MNL choice model occur within an alternative -- there are no cross effects of one alternative on another. If there are similarity or nesting or non-constant cross elasticity effects, these will usually be manifest as interactions between the alternatives. Therefore, by testing for the presence of significant interactions between alternatives, one can test the validity of the IIA assumption of MNL choice models in discrete choice conjoint strategies. It should be noted that the interactions tested in this case are between alternatives, and not within alternatives. Interactions within alternatives are just the interactions between dimensions or attributes of alternatives, which are a function of the functional form of the preference function of respondents -- which can also be tested empirically by such strategies.
However, these latter tests are not related to the IIA assumption of MNL choice models.

Respondents are usually asked to choose one and only one of a given choice set of multiattribute alternatives. Louviere and Woodworth (1983) have recommended the use of a constant "base" alternative in each choice set in order to set the origin of the preference scale. They recommended that this base alternative be scaled to be the zero point on the preference scale and the remaining scale scores can then be interpretable relative to this alternative. Technically, this base alternative is not needed for the analysis, but if it is formulated as a "do-nothing" alternative it would make the choice task more realistic in that it allows respondents the option of not choosing any alternative or delaying their choice.

Analysis

There are several estimators available to analyze these data statistically, depending upon whether one wants to aggregate the observed choices over all respondents or a subset of respondents or one want to analyze the discrete responses for each respondent individually. The decision on the level of aggregation required for the statistical analysis usually depends on requirements of the research, hypotheses to be tested, availability of computer programs, a priori information, etc. (Louviere, 1983). The drawbacks of using individual-level data is that a large number of choice tasks may be required of each respondent [Amemiya (1981) has recommended a sample size of 30 observations per cell of the factorial design]. This is a serious issue, since it increases respondent burden and
therefore may compromise the validity of the results if respondents do not
give meaningful responses on account of fatigue. Further, MNL estimators
are biased and for small samples there is a finite probability that they will be
infinite (Bunch and Batsell, 1989). These problems of estimation disappear in
aggregate-level analysis. On the other hand, if one aggregates the data one
must be careful about standard errors estimated from the aggregated data
because the repeated observations are not truly independent. In other words,
one must be concerned about variations in model parameters across
respondents (Louviere, 1988a). As Elrod, Louviere, and Davey (1992) have
noted, aggregate-level analysis of consumers who are typically
heterogeneous in their choice behavior has several shortcomings -- the
model can predict individual difference in choice behavior only if they can
be accounted for by including individual characteristics in the model, more
complex models are needed to characterize aggregate behavior, and
interpretation of aggregate models is more difficult because they confound
the choice process operating at the individual level with heterogeneity in
that process across individual respondents. Most applications, however, use
aggregate-level data in the analysis.

Most of the early studies by Louviere and his colleagues aggregated
the choices across respondents and implemented Generalized Least Squares
(GLS) using weighted multiple linear regression -- also known as Weighted
Least Squares or WLS -- to estimate the parameters of interest. In this case,
the dependent variable is the natural logarithm of the frequencies with
which a particular alternative is selected. The weight for each observation is
the observed frequency of choices. The independent variables are: (a) N-1
dummy variables for all alternatives except the base alternative. Each observation is coded "1" if the observation pertains to a particular alternative; "0" otherwise and (b) I-1 dummy variables for all choice sets except one (where I is the total number of choice sets). The dummy is coded "1" if the observation pertains to a particular choice set which it represents; "0" otherwise. These choice set dummy variables are needed to avoid biasing the V-estimates -- however, their coefficients are of no interest. Louviere and Woodworth (1983) have noted that the estimates derived by using this WLS approach are minimum modified chi-square solutions, which have been demonstrated to be asymptotically efficient in large samples (see Bhapkar, 1966; Grizzle, Starmer, and Koch, 1969).

This approach is essentially derived from the estimation procedure introduced to the marketing literature by Nakanishi and Cooper (1974), who demonstrated that a simple log transformation on compensatory multiattribute choice models rendered the model linear in the parameters; in that case, a Taylor expansion leads to estimators that can be calculated by ordinary or generalized least squares (Nakanishi and Cooper, 1974). In fact, as Bunch and Batsell (1989) have noted, this approach applied to the MNL choice model is a generalization of the minimum logit chi-square estimator (MLCS) proposed by Berkson (1944) for the binary logit model.

It should be noted that this particular estimation procedure is referred to by various names in the published literature. Following Berkson (1955), Bunch and Batsell (1989) refer to it as minimum logit chi-square or simply minimum chi-square, reflecting the fact that the quantity being minimized by GLS is asymptotically chi-square. As noted before, the estimator has also
been referred to as WLS, because weighted least squares is the GLS estimator for the binary case. Finally, it is also known as the “Berkson” or “Berkson-Theil” estimator in the econometrics literature (see McFadden, 1974). For the purpose of this dissertation, the nomenclature adopted by Bunch and Batsell (1989) shall be followed -- hence it shall be referred to as minimum logit chi-square or MLCS.

Apart from MLCS, other possible estimation procedures for discrete choice conjoint strategies include ordinary least squares regression (OLS), nonlinear least squares estimation (NLLS), minimum Pearson chi-square estimation (MPCS), and maximum likelihood estimation (ML). Bunch and Batsell (1989) have noted that maximum likelihood and minimum logit chi-square estimators have received the most attention in the literature. They attribute this to the following reasons: (a) MPCS requires solution of a nonlinear optimization problem though having the same asymptotic properties as MLCS -- in other words, MPCS is computationally more challenging but still does not yield any greater benefit as compared to MLCS, (b) NLLS and OLS, though consistent, are not asymptotically maximum likelihood, and (c) NLLS requires nonlinear optimization (Bunch and Batsell, 1989).

Given this multitude of estimation procedures, it is important to know how they compare to one another so that an informed decision can be made when selecting which one to use for use in a particular study. In an important simulation study, Bunch and Batsell (1989) conducted a Monte Carlo comparison of five estimation procedures for the MNL choice model -- maximum likelihood (ML), minimum logit chi-square (MLCS), minimum
Pearson chi-square (MPCS), nonlinear least squares (NLLS), and ordinary least squares (OLS). Their bases for choosing an estimator were that, if all other things were equal: (a) the closer the parameter estimates are to the true parameters, the better, (b) the more accurate the predicted choice probabilities are, the better, and (c) the faster test statistics' sample distributions converge to their asymptotic distributions, the better (Bunch and Batsell, 1989).

Using a variety of point estimation and predictive accuracy measures, Bunch and Batsell (1989) demonstrated the superiority of ML estimators over the other estimation procedures tested in the simulation. Specifically, their simulation found the following rank ordering of the five estimators: ML > NLLS > MPCS > MLCS > OLS. They noted that this rank ordering was especially relevant when the data collection process was based on a fractional factorial design -- which, as noted earlier, is the type of design most commonly used in discrete choice conjoint strategies. They further noted that ML was preferred also in terms of statistical inference, i.e., whenever individual test statistics are used to determine whether certain explanatory variables are determinants of choice (Bunch and Batsell, 1989).

It is important to note that the results of the Bunch and Batsell (1989) simulation are contradictory to most other comparisons of these simulators, as seen in the biostatistics literature. In their comparison between ML and MLCS, Smith, Savin, and Robertson (1984 p.471) noted that the published numerical examples (until that time) were "... unanimously in support of the ... (MLCS) ... estimator" on the mean square error criterion, a commonly used point estimation accuracy measure for individual parameter estimates.
However, in the case of statistical inference, even Smith, Savin, and Robertson agreed that ML was superior to MLCS in terms of inference and asymptotics (Smith, Savin, and Robertson, 1984).

The divergent results of Smith, Savin, and Robertson (1984) and Bunch and Batsell (1989) may be explained by the nature of the context of the studies on which they based their respective conclusions. In this regard, the Smith, Savin, and Robertson (1984) article was primarily concerned with the application of logit analysis in a biostatistics framework -- where the choice model assumed is usually binary logit. Further, from a design point of view, studies in the biostatistics literature rarely use fractional factorial designs to manipulate their independent variables and consequently have a small number of "design points." The Bunch and Batsell (1989) article, on the other hand, was primarily concerned with the application of logit analysis in a marketing research framework -- where the choice model assumed is usually multinomial in nature, such as multinomial logit. Moreover, from a design point of view, these studies make extensive use of fractional factorial designs on account of the large number of alternatives studied -- consequently, these studies have numerous multiple subsets in the data collection process and can therefore be characterized as having a large number of "design points."

Given this line of reasoning, the next question that arises is which framework do studies on health status preference measurement follow -- the biostatistical or marketing research framework? As noted previously, health status is a multidimensional construct and needs to be operationalized by several dimensions or attributes. Further, each of these dimensions or
attributes can be described by several levels of functioning. Therefore, there are a significant number of potential combinations of each dimension/level. In order to get a preference function for the health status classification system, the preferences for being in each of these combinations needs to be measured. A few examples may help in appreciating the enormity of possible combinations in published health status measurement instruments. For example, in the Mark III version of the McMaster Health Utility Index, there are a total of 972,000 possible states of health an individual can experience. Although the revised version of the EuroQol instrument has only 243 possible states of health, this is still a large number of alternatives to be evaluated holistically. Therefore, it is essential to use fractional factorial designs in order to measure health status preferences. Further -- and specifically for the purpose of this dissertation -- in the case of antihypertensive drugs, there are several drugs available. Therefore, assuming a binary choice model is not reflective of clinical practice. To summarize, in measuring health status preferences for hypertensive patients, it is necessary to adopt a multinomial choice model and use fractional factorial designs. These conditions fall within the purview of the marketing research framework, and hence support the use of ML as the method of estimation.

Evaluation of Discrete Choice Conjoint Strategies

Compared to the other conjoint methodology strategies discussed, discrete choice conjoint strategies offer several advantages. From a feasibility point of view, they are much easier to implement because all a respondent
has to do is select one of several alternatives. In this regard, the scaling task corresponds closely to the behavior observed in the market. Further, the researcher need only assume a nominal level of measurement.

A more important advantage of these strategies is that one can estimate choice models directly from choice data, and thus avoid the potentially unrealistic assumptions about choice behavior that would be implied by any choice simulation method. Moreover, it is possible to study how choices vary as a function of both size and composition of choice sets (Louviere, 1988a). Therefore, unlike rank order and rating scale judgment strategies, discrete choice conjoint strategies enable the researcher to study and model -- and therefore explain -- choice processes of respondents. This ability to explain responses is especially important for the measurement of health status preferences, as explained earlier.

Louviere and Woodworth (1983) have summarized the advantages of discrete choice conjoint strategies: "The approach ... enables a researcher to study simultaneously the choice process and the attribute tradeoff or evaluation process in choice. The researcher can infer the scale values of conjoint-like models from choice data, and can apply the models derived to the prediction of choice. Furthermore, by appropriate choice of experimental design, a researcher can test competing choice models and/or reject as false the deductions from an assumed model" (p.351).

However, discrete choice conjoint strategies are not without their own drawbacks. Most practical applications will not yield enough data to enable the statistical analysis to be conducted at the individual level. This may preclude their use in segmentation studies. Moreover, since these strategies
require two separate experiments to be combined, the design of these strategies is more difficult as compared to rank order or rating scale judgment strategies. This is because both experimental designs must satisfy certain statistical properties to enable the estimation of parameters and the testing of hypotheses (Louviere and Woodworth, 1983).

Other drawbacks pertain to the specific type of choice model assumed to represent the decision making process. For instance, if the MNL choice model is used, it has been shown that some assumptions of this model are not always satisfied in practice. Foremost among these assumptions is the IIA assumption. Although alternative choice models have been proposed which do not involve this assumption [e.g., nested logit (see McFadden, 1980), dogit (Gaudry and Dagenais, 1979; Caudry and Wills, 1979), these models have provided difficulties in estimation. Therefore, at this stage in the development of choice models, the particular choice model chosen represents a tradeoff between the veracity of its assumptions and its ease in estimation.

It had been noted earlier in this chapter that rank order and rating scale judgment strategies yield similar results. Although there have been no published comparisons between rank order judgment and discrete choice conjoint strategies, the three published empirical comparisons between rating scale judgment and discrete choice conjoint strategies have not found any significant differences between these strategies in terms of the evaluative criteria used in each study (Louviere and Gaeth, 1988; Elrod, Louviere and Davey, 1992; Oliphant, Eagle, Louviere, et al., 1992).
The first empirical comparison between rating scale judgment and discrete choice conjoint strategies was by Louviere and Gaeth (1988), who reported the results of three studies in which both types of strategies were used in such a way that the parameters of models based on these response modes could be compared. Louviere and Gaeth's (1988) comparison focused on the parameter estimates of the scale scores of attributes or dimensions obtained via inferential statistics on the rating and choice data collected as a result of implementation of these strategies. Using a variety of types of alternatives (airline tickets, hypothetical careers, and city parks) and modifications in experimental design (response mode as a between- as well as a within-subjects factor, different position order of rating scale and choice tasks), Louviere and Gaeth (1988) found a high degree of linear correlation between the parameters estimated via the two strategies (Pearson correlation coefficients of 0.94, 0.73, and 0.98 for the three sets of data). In some instances, there were no significant differences between the parameter estimates from the different strategies (however, concordance coefficients were not reported). Further, the data supported the existence of a linear functional relationship between the two sets of scores, as shown by a simple linear regression with the rating scale parameters as the independent variable and the choice-based parameters as the dependent variable. Louviere and Gaeth (1988) concluded the two strategies were proportional to one another and that the researcher can draw similar conclusion about the magnitude of the effects of the attributes or dimensions of the alternatives under evaluation from both strategies (Louviere and Gaeth, 1988).
These data are encouraging, since they demonstrate that discrete choice conjoint methodology can produce valid data -- if one assumes rating scale conjoint strategies yield valid data (an assumption which is supported by empirical evidence, see, for example, Green and Srinivasan, 1990) and then applies the principles of convergent validation, as discussed in Chapter II. However, in spite of the high linear correlations between the two sets of measures, there were instances in which application of the two strategies yielded different preferential ordering of the attributes or dimensions. On a related note, Kendall's tau coefficients were not reported for any of the sets of data. It should be noted that the different ordering of attributes or dimensions is not a necessary and sufficient condition to guarantee the phenomenon of preference reversal, since the two strategies might have prompted respondents to combine the subjective worth or preference scores of the attributes or dimensions in different manners. This is analogous to different forms of the function C in Figure 3 and is consistent with change of process theory (Mellers, Ordonez, and Birnbaum, 1992), which is one of the proposed theories forwarded to account for response mode effects, and will be reviewed in Chapter IV.

Finally, the point should be made that neither of the three studies reported in Louviere and Gaeth (1988) were designed for the purpose of empirically comparing and contrasting rating scale and discrete choice conjoint strategies. Indeed, as Louviere and Gaeth (1988) themselves noted in the introduction to the article, " ... there is little theory or empirical evidence about this issue [i.e., comparisons between rating scale and discrete choice conjoint strategies], and as we have access to several sets of data that
permit us to make comparisons, we believe that we can shed some light on this issue" (p.60). While this does not necessarily detract from the significance of the contribution of Louviere and Gaeth (1988), it does serve to emphasize the point that the conclusions drawn by Louviere and Gaeth (1988) are -- at best -- tentative and await empirical corroboration via experiments specifically designed to compare the two measurement methodologies.

In apparent recognition of this drawback of the Louviere and Gaeth (1988) article, Elrod, Louviere, and Davey (1992) conducted an experimental investigation of this very issue. Specifically, Elrod, Louviere, and Davey (1992) empirically compared rating scale judgment and discrete choice conjoint strategies in terms of their abilities to predict shares in a holdout choice task. It is noteworthy that they designed the holdout choice task in such a way that it contained alternative attributes which were maximally negatively equicorrelated, since previous research had suggested that this was the most demanding prediction task for conjoint models (Johnson, Meyer, and Ghose, 1989).

In comparing the predictive ability of the two strategies, Elrod, Louviere, and Davey (1992) used two measures of predictive fit. The primary measure was the proportion reduction in mean square error of prediction (PRMSEP). Elrod, Louviere, and Davey (1992) noted that PRMSEP was a much more stringent measure of predictive fit than $R^2$ since $R^2$ only assesses goodness of fit of an arbitrary linear (as opposed an equality) relationship between the actual and predicted scores. Both strategies performed equally well with respect to the PRMSEP. The
secondary measure of predictive fit used by these researchers was the mean absolute deviation (MAD) of prediction. Once again, the two strategies fared equally well with respect to this criterion of predictive fit. Therefore, Elrod, Louviere, and Davey (1992) concluded that both rating scale judgment and discrete choice conjoint strategies were equal in performance in terms of predictive ability.

The two studies discussed above compared rating scale judgment and discrete choice conjoint strategies in terms of two different criteria on different datasets. Oliphant, Eagle, Louviere, et al., (1992) compared rating scale judgment and discrete choice conjoint strategies in terms of these two criteria (among other criteria) on the same dataset. In their comparison of the parameters of the two types of models, Oliphant, Eagle, Louviere, et al., (1992) adopted a different approach than that of Louviere and Gaeth (1988). To compare the aggregate preference functions generated by the two strategies, Oliphant, Eagle, Louviere, et al., (1992) first factor analyzed the vectors of parameters to obtain the eigenvalues and then plotted the parameters against the derived factor scores from the first eigenvector. Their results indicated that the preference vectors were approximately proportional to one another. They interpreted this to mean that the strategies measured the same utilities at the aggregate level -- in other words, both strategies measured the same construct. At this point, it should be noted that the proportionality condition they invoked derives from the fact that the scale of the preference scores in an MNL model is inversely proportional to the variance. The scale cannot be identified for any particular dataset, but
the ratios of scales can be identified for pairs of datasets (Swait and Louviere, 1993).

The second criterion for comparison of rating scale judgment and discrete choice conjoint strategies used by Oliphant, Eagle, Louviere, et al., (1992) was predictive ability. These researchers used three measures of the predictive ability of the strategies under evaluation: (a) the total chi-square (sum) calculated with respect to the observed and predicted choice frequencies, (b) the mean square calculated with respect to observed and predicted choice frequencies, and (c) the slope and intercept obtained by regressing the natural log of the observed choice frequencies on the natural log of the predicted choice frequencies. In terms of the first two measures of predictive ability, the two strategies yielded equivalent results. In order to assess whether the chi-square and mean square results were due to exceptional fits in some sets and not others, as well as whether there were scale differences between the estimation and holdout choices, Oliphant, Eagle, Louviere, et al., (1992) used the third measure of predictive fit. As noted by these researchers, such regressions are appropriate because differences in scale scores give rise to differences in slopes, although intercepts will differ significantly from zero unless the two scales are the same in both datasets. In particular, if the scale is the same in both observed and predicted datasets, the slope should not differ significantly from one, and the intercept should not differ significantly from zero (Oliphant, Eagle, Louviere, et al., 1992). The results indicated that both strategies produced approximately the same levels of fit. Therefore, in terms of all measures of predictive ability, there were no differences between rating scale judgment
and discrete choice conjoint strategies. This finding therefore replicated that of Elrod, Louviere, and Davey (1992).

Finally, Oliphant, Eagle, Louviere, et al., (1992) conducted an additional test to determine whether there were differences between the preference functions obtained via the two strategies. The rationale for this test was based on the fact that the preference scores derived in MNL models are affected by a scaling constant, which is inversely proportional to the error variance in the dataset. While the scaling constant for any one dataset cannot be uniquely identified, the ratio of scales between the two datasets can be determined and tests can accordingly be conducted. The idea behind these tests is to determine whether the preference functions differ only by a scaling constant or because their parameters are actually different (Oliphant, Eagle, Louviere, et al., 1992). Using a test developed by Swait and Louviere (1993), Oliphant, Eagle, Louviere, et al. (1992) found that the difference between the estimated preference functions was due solely to a scaling constant for choice sets which forced the respondents to make a choice (i.e., choice sets which did not include a 'no-choice' alternative). However, for choice sets which included a 'no-choice' alternative, there were significant differences between the preference functions obtained via the rating scale judgment and discrete choice conjoint strategies (Oliphant, Eagle, Louviere, et al., 1992). This result is interesting, in that it suggests that the two strategies may differ in the way they handle the measurement of the 'no-choice' alternative in the decision making process -- even though both strategies were similar in terms of predictive ability in choice sets with and without the 'no-choice' alternative. Therefore, in terms of an overall
evaluation in terms of all three criteria used in this study, there seem to be little differences between rating scale judgment and discrete choice conjoint strategies in terms of construct validity and predictive ability.

Given these similarities between rating scale judgment and discrete choice conjoint strategies, it is necessary to reiterate the major strength of discrete choice strategies over rating scale judgment strategies -- the fact that they have a theoretical base to explain respondents' choices and do not have to rely on untestable assumptions to link respondents' stated preferences to choices. Although the functional measurement approach can explain how respondents integrate information and make a judgment, it falls short of explaining respondents' choices between several alternatives. Discrete choice conjoint strategies, on the other hand, can explain both these processes.

Still, there are some instances where the rating scale judgment strategies may be preferred over their discrete choice counterparts. In this regard, Elrod, Louviere, and Davey (1992) have observed, "The choice of approach may depend more on intended use. Choice-based models, which are fit at the aggregate level, have several advantages. The values and statistical significance of all parameters are easily reported, share predictions for new brands are easily produced, and the researcher has the added assurance that choice data were used to calibrate the model. However, aggregate choice models hinder segmentation studies of a market. In contrast, the ratings-based models are well-suited to segmentation studies, but the estimation results are difficult to summarize, tests of statistical significance of attributes in aggregate are cumbersome, and simulation of
choice shares is awkward. Which of these factors are paramount can be decided only by the individual researcher” (p.375).

Louviere (1988a) has summarized the pros and cons of discrete choice conjoint strategies in comparison to rank order and rating scale judgment strategies: “Thus, with traditional conjoint rating or ranking methods one trades simplicity of design for unreality of task, strong response measurement level, and choice simulation assumptions. With discrete choice experiments one trades simplicity of response, better understanding of choice process, and minimal assumptions for increased complexity of design (p.100).”

In summary, conjoint methodology provides a very powerful tool for measuring health status preferences. Of all the different strategies within conjoint methodology, the discrete choice conjoint strategies are especially noteworthy because they combine the strengths of the functional measurement approach (i.e., testing the combination function and providing meaning to scale scores as a consequence of a valid combination function) with the theory of relating preferences to actual choices. Although these strategies are more difficult to design, the recent availability of computer programs (e.g., Choice-Based Conjoint System by Sawtooth Software Corporation) may help researchers overcome some of the current disadvantages of discrete choice conjoint strategies.
Hybrid or Mixed Preference Measurement Strategies

Before moving on to the scaling approaches which can be used in health status preference measurement, hybrid or mixed preference measurement strategies need to be briefly discussed. In this regard, three different strategies need to be discussed -- hybrid conjoint modeling, Huber-hybrid modeling, and the analytic hierarchy process. On account of the similarities between hybrid conjoint modeling and Huber-hybrid modeling, these two strategies will be discussed together, while the analytic hierarchy process will be discussed separately.

Hybrid Conjoint Modeling and Huber-hybrid Modeling

From the classification of preference measurement strategies presented in Figure 4, it can be seen that both hybrid conjoint as well as Huber-hybrid modeling fall under the class of decomposed preference measurement strategies. This is because in both cases, it is possible to obtain preference scores for each health status dimension, each level of each health status dimension, as well as an overall health status preference score.

Hybrid conjoint modeling was developed by Paul Green and his colleagues a way to reduce the complexity of data collection in regular rank order and rating scale conjoint strategies (Green, Goldberg, and Montemayor, 1981). Hybrid conjoint modeling is usually implemented in two stages. The first stage entails some type of self-explicated preference measurement procedure, where respondents evaluate the levels of each attribute or dimension (one at a time) on some type of desirability scale. This is followed by an evaluation of the attributes or dimensions themselves on
an importance scale. In the second stage of the hybrid procedure, the respondent evaluates a small set (usually four to nine) of multiattribute or multidimensional alternatives by ranking or rating them on some type of scale. Each of these alternatives is drawn (usually in a balanced way) from a much larger master design, normally constructed as a fractional factorial design. Hybrid models combine these two types of information to estimate a preference function that combines some parameters at the individual level and some at the aggregate level (Green, Goldberg, and Wiley, 1983).

The multiattribute evaluations are treated as the dependent variable in a regression equation, with the self-explicated preference scores as the independent variable. If the multiattribute evaluations are completely predicted by the self-explicated preference scores (up to a linear transformation), no other parameters need to be estimated. Usually, however, this is not the case (Green, 1984). Therefore, the multiattribute or multidimensional evaluations are analyzed statistically in order to estimate the main effects and selected interactions at the aggregate level. These estimated aggregate-level parameters are then introduced into the regression equation as independent variables in order to better predict the multiattribute or multidimensional evaluations. This procedure is quite flexible in that it allows for the existence of a variety of model forms (see Green and Kreiger, 1994). In terms of predictive validity, these hybrid models have been shown to strongly outperform self-explicated models and compare favorably with rating scale conjoint judgment strategies (Green, Goldberg, and Wiley, 1983).
The similarities between hybrid conjoint modeling and the conditional utility function-based procedure need to be noted. Both strategies involve a combination of multi- and single-attribute evaluations by respondents. Further, in both strategies, the role of the multiattribute or multidimensional evaluations is to obtain information relevant to interactions among attributes or dimensions. However, the differences between the two strategies should also be noted. The primary difference between the two strategies is that unlike hybrid conjoint modeling, the conditional utility function-based procedure does not have any error theory. Moreover, from an origin point of view, the conditional utility function-based procedure was developed within the field of decision theory, while hybrid conjoint modeling has its roots in psychometric theory.

Finally, the distinction between hybrid conjoint modeling and Huber-hybrid modeling needs to be noted. Both, hybrid conjoint as well as Huber-hybrid modeling share a similar methodology, i.e., respondents provide attribute- or dimension-level preference scores and a few holistic alternative preference scores. However, there are two major differences between these two modeling techniques. First, hybrid conjoint modeling requires respondents to provide importance scores for each attribute or dimension, usually by some type of rating or ranking scaling method. Huber-hybrid modeling, on the other hand, estimates these importance scores via inferential statistics, usually multiple regression.

The second major difference between these two modeling techniques lies in the selection of the holistic alternatives to be evaluated. Hybrid conjoint modeling, in the spirit of functional measurement selects these
holistic alternatives on the basis of some type of fractional factorial design -- the emphasis here is on experimental manipulation of levels of attributes or dimensions. Huber-hybrid modeling, on the other hand, operates in the spirit of social/clinical judgment analysis and emphasizes representative design (as opposed to experimental design) in the construction of holistic alternatives. Therefore, the discussion on the relevant merits and demerits of functional measurement versus social/clinical judgment analysis is also relevant in the discussion of hybrid-conjoint versus Huber-hybrid modeling. Since this discussion was already presented earlier in this chapter, it shall not be repeated here.

A final point of distinction between hybrid conjoint and Huber-hybrid modeling is more in terms of motivation for development. In this regard, the Huber-hybrid models appear to be been developed in response to the concern that purely additive subjective evaluation models may not adequately describe individual decision making (Huber, Sahney, and Ford, 1969). On account of the then-existing controversy between importance weights obtained directly from respondents (i.e., "subjective" weights) versus those derived from statistical models (i.e., "objective" weights) which questioned the validity of "subjective" weighting methods (for e.g., see Goldberg, 1968; Hoffman, 1960; and Slovic and Lichtenstein, 1971) as well as problems in estimating scaling constants for parameters of multiplicative models [which were subsequently dealt with by Keeney and Raiffa (1976)], the Huber-hybrid models used holistic evaluation to determine importance weights. Thus, the "motivation flow" for Huber-hybrid models can be described one from the bottom to the top, in the sense that they built upon
the premise of improving subjective evaluation or self-explicated models in general. In order to understand this description, the "top-down" versus "bottom-up" distinction drawn between strategies introduced earlier is important, since the Huber-hybrid models began at the bottom (i.e., the levels of attributes or dimensions) and required "top-level" evaluations (i.e., the holistic evaluations of alternatives) for operationalization.

Hybrid conjoint models, on the other hand, were developed for the purpose of reducing complexity in data collection requirements and consequent respondent burden in regular conjoint studies. Therefore, the "motivation flow" for these models can be described as one from the top to the bottom, in the sense that they built on the premise of improving conjoint models in general. In other words, they began at the top (i.e., the holistic evaluations of alternatives) and required "bottom-level" evaluations (i.e., the levels of attributes or dimensions) for operationalization.

The Analytic Hierarchy Process

Another strategy which may be used to measure health status preferences is the Analytic Hierarchy Process (AHP). Developed by Saaty (1980), the AHP was designed for use with individuals or groups for decision-making and planning in complex situations where multiple considerations -- both subjective and objective -- are important. Therefore, like the additive rating scale method and the conditional utility function-based procedure, the AHP has its roots in the field of decision theory. Unlike these methods, however, the AHP can be implemented as both a holistic as well as an explicitly decomposed health status preference measurement
strategy depending on the particular hierarchy chosen (as explained below). This is the reason why the AHP is presented as a separate type of preference measurement strategy (i.e., a mixed holistic-decomposed strategy) in this discussion.

The AHP can be divided into four basic tasks: (a) structuring the problem as a hierarchy, (b) conducting pairwise comparisons between the elements of adjacent hierarchy steps, (c) combining the pairwise comparisons to derive weights for hierarchical elements, and (d) synthesizing the element weights into a set of scores for each decision alternative (Dolan, Isselhardt, and Cappuccio, 1989; Zahedi, 1986). Each of these tasks will be discussed in relation to measuring health status preferences.

Structuring the problem as a hierarchy involves breaking the problem down into its component parts and arranging them into a hierarchy according to a set of specific guidelines. In most cases, the hierarchy consists of three steps -- an overall goal, a set of criteria, and a set of decision alternatives. The goal consists of a concise statement of the objective of the strategy, e.g., to measure preferences for health states. The set of criteria consists of all considerations that need to be satisfied in order to meet the goal, e.g., physical function, mental function, and social function. More general criteria can be divided into progressively more focused subcriteria if deemed necessary. As originally formulated for the context of decision-making, the set of decision alternatives consists of the different actions that can be taken to achieve the goal. Applied to health status preference measurement, the way the set of alternatives is defined dictates whether the
AHP is implemented as a holistic or explicitly decomposed preference measurement strategy.

For instance, if the set of alternatives consists of descriptions of health states, the AHP would be implemented as a holistic preference measurement strategy because the scale scores obtained will be for the overall health state and will not be able to be decomposed into part-worths. On the other hand, if the set of alternatives consists of levels of different health status dimensions (i.e., each alternative is one level of one health status dimension), the AHP would be implemented as an explicitly decomposed preference measurement strategy because the scale scores obtained will be for the individual part-worths or levels of each health status dimension.

In the second task of the AHP, paired comparisons are made between the decision elements identified in each step of the hierarchy. This is done by comparing the elements of every hierarchy step relative to each element on the next higher step. For instance, comparisons can be made between physical function and mental function relative to the goal of measuring preferences for health states. Similarly, comparisons can be made between two different levels of one health status dimension relative to that or another dimension. These comparisons can be made in terms of preference, likelihood, or importance. The comparisons are made based on an assessment of the magnitude of the difference between the two elements with respect to the referent criterion (Dolan, Isselhardt, and Cappuccio, 1989). Saaty has recommended a 1-9 point comparison scale for making these comparisons, with 1=“the two elements are equal” to 9=“one element is absolutely favored” (Saaty, 1980).
Each set of comparisons is entered into a matrix. By convention, the matrix is constructed so that the judgment in each cell of the matrix represents the weight of the row element relative to the column element. The comparison between each pair of elements is made only once; the reciprocal is used to complete the matrix (Dolan, Isselhardt, and Cappuccio, 1989).

In the third task of AHP, the paired comparisons made during the previous task are transformed into relative weights for the elements. This is done by calculating the right principal eigenvector of the comparison matrix. The components of this vector, when normalized, represent the weights assigned to the elements (Dolan, Isselhardt, and Cappuccio, 1989). Saaty (1980) has proposed four different ways to calculate an approximate eigenvector: (i) sum the elements in each row of the comparison matrix and normalize the result so that all element scores add up to one, (ii) sum the elements in each column of the matrix, take the reciprocals, and then normalize the reciprocals so that they add up to one, (iii) normalize the columns of the matrix, then sum the elements of each row of the matrix and normalize the results, and (iv) take the geometric mean of each row in the matrix and normalize the results. Saaty (1980) has shown that the last two methods are usually more accurate than the first two. The exact eigenvector can be closely approximated by raising the comparison matrix to successive powers, \( k \), until the absolute value of the difference between the normalized row sums of the matrix \( A^1 \) and \( A^k \) is smaller than a prespecified number such as 0.0001 (Gass, 1985). At that point the vector composed of the normalized row sums of \( A^k \) very closely approximates the eigenvector. It should be noted that regardless of the computational method used, the
resulting weights are always normalized so that they sum up to one. Thus, the weights represent a ratio comparison among the paired elements in terms of the referent criterion.

The final task in the AHP consists of synthesizing the element weights generated in the previous task into a composite weight for each alternative under consideration (i.e., either individual part-worths or holistic health states, depending on the implementation of the strategy). The calculation for the holistic implementation of the AHP consists of multiplying the relative weight of each alternative by the relative weight(s) of each criterion (or family of related criteria) and adding the results (Dolan, Isselhardt, and Cappuccio, 1989). This implies an additive model form. For the explicitly decomposed implementation of the AHP, the results of the previous task are all that is needed, since they give the part-worths and the relative importance of different health status dimensions.

The major strengths of the AHP are the generation of a ratio scale of preference measurement and the means to test some of the assumptions underlying the analysis. A ratio scale of measurement retains all the advantages of an interval scale discussed earlier; further, it permits one to say by what magnitude one health state is preferred to another (e.g., twice, five times, etc.). One of the assumptions of the AHP is judgmental consistency or transitivity, i.e., if alternative A is twice as preferable as alternative B, and B is four times more preferable than C, then, to be consistent, A must be eight times more preferable than C. Saaty (1980) has shown how the consistency of the responses can be checked using a measure called the consistency ratio.
However, the AHP suffers from some major drawbacks. Foremost among these is respondent burden. In this regard, the number of paired comparisons each subject is required to respond to is prohibitively large. For instance, with a three step hierarchy with five criteria and only three alternatives, the total number of paired comparisons is 25. It should be noted that there will be much more than three alternatives in the case of health status preference measurement, so the number of paired comparisons will be much more. This is especially true for the explicitly decomposed implementation of the AHP. The holistic implementation of the AHP also shares the drawbacks of the regular holistic preference measurement strategies discussed earlier. Further, as explained by Saaty (1980), the AHP only allows the existence of an additive model form. Given the importance of incorporating interactions between health status dimensions in any measure of health status preference (as discussed earlier), this is another major drawback of the AHP. Eckman (1989) has voiced concern that the AHP "... is not rigorous enough to allow formal conclusions to be derived from the results of an analysis." Indeed, there have been no published applications of the AHP to the measurement of health status preferences.

In summary, although the holistic preference measurement strategy was used in initial attempts to measure preferences for health states, the decomposed preference measurement strategies are now more popularly used by health status researchers. The principal advantage of these decomposed strategies is that they require fewer respondent judgments. Of all the approaches within the measurement strategies discussed, only the
functional measurement approach and discrete choice conjoint strategies actually permit a test of the underlying information processing framework and the scale level of the measure used. Moreover, only discrete choice conjoint strategies provide a theoretical rationale for linking stated preferences to actual choices of respondents.

Measurement strategies are one of the steps in the methodology of health status preference measurement. Another important step in this methodology is the choice of a scaling approach. Within each measurement strategy, several different scaling approaches can be used to measure health state preferences. These different scaling approaches are discussed in the next section.

**Scaling Approaches for Health Status Preference Measurement**

In discussing the various scaling approaches for health status preference measurement, the foci of scaling will first be discussed, followed by an introduction to the two basic scaling approaches used in health status preference measurement. As explained later, within these basic types of scaling approaches, several different types of scaling methods may be employed. These scaling methods are derived from two major theoretical disciplines, which shall be briefly summarized, after which the specific scaling methods themselves will be described in some detail. Finally, some points to consider in choosing a particular scaling method will be discussed.
Scaling People Versus Scaling Stimuli

Before discussing the various scaling approaches for health status preference measurement, it is necessary to distinguish between two different foci of scaling approaches -- scaling people versus scaling stimuli (Froberg and Kane, 1989b). Bush (1984) has referred to this distinction as one between scaling relative frequency of events versus relative preferences for these same events. The distinction between these two foci is best illustrated by revisiting the purpose of the scaling exercise, i.e., are we interested in comparing people by identifying their location along a continuum or are we interested in comparing something else along a continuum, namely health states? (Froberg and Kane, 1989b).

As Froberg and Kane (1989b) have observed, the distinction between scaling people and scaling stimuli is important for two reasons: (a) it has implications for the selection of appropriate scaling method. In this regard, methods for scaling people are usually drawn from the psychometric theory literature of mental test theory, while methods for scaling stimuli are usually drawn from the fields of psychophysics and decision theory, e.g. Likert or Summated Scaling methods are generally used for scaling people, decision theoretic methods such as the standard gamble are generally used for scaling stimuli, while Guttman or Cumulative Scaling methods may be used for both purposes, and (b) it has implications for the way in which variability in the preferences is handled, e.g., the objective in scaling people is to discriminate between them by spreading them out along a continuum while the objective in scaling stimuli is to obtain consensus among judges regarding the scale score of each stimulus (Froberg and Kane, 1989b).
In the field of health status measurement, health status profiles are usually developed for the purpose of scaling people while health status indices are usually developed for scaling stimuli. The focus of this dissertation is on determining scale scores for stimuli, i.e., health states. Therefore, the following discussion will emphasize scaling approaches for scaling stimuli, not people.

Direct Versus Indirect Scaling Approaches

Approaches for stimulus scaling are of two general types -- direct and indirect (Froberg and Kane, 1989b). The designation of "direct" versus "indirect" depends on whether subjects make numerical or non-numerical responses. In direct scaling, respondents are instructed to make evaluations at a certain level of measurement and the resulting data are treated as such. For instance, respondents may be asked to perform a ratio-level scaling task such as assigning a number which represents the magnitude to which a certain health state under evaluation is preferred to the health state of death. On the other hand, in indirect scaling approaches, respondents are instructed to make evaluations at a certain level of measurement and the data are later converted to a different level of measurement by the researcher(s). For instance, in the method of paired comparisons (which is a scaling method used as part of a rank order judgment conjoint strategy), all possible pairs of health states are presented to respondents, who indicate which of the two states they prefer, i.e., an ordinal judgment. In order to convert these ordinal-level judgment data into interval-level or ratio-level measures of preference, it is necessary to apply a set of theoretical
assumptions based on the variability of the subjects' responses. One such set of assumptions is Thurstone's Law of Comparative Judgment, which is based on the idea that stimulus differences which are detected equally often are subjectively equal (Thurstone, 1927a).

The major difference between the direct and indirect scaling approaches is in their underlying assumptions. Direct scaling approaches have two salient assumptions: (a) subjects are capable of directly generating an interval-level or ratio-level scale of measurement via interval-level or ratio-level responses and (b) although there is some error in evaluations made by one respondent on one occasion, error may be reduced by averaging evaluations over respondents, i.e., respondents are replicates of one another (Froberg and Kane, 1989b). The indirect scaling approaches do not make these assumptions.

Patrick and Erickson (1993) have argued that the designation of "direct" versus "indirect" may be inappropriate since there is no evidence that numerical responses are more "direct" than non-numerical ones. Nevertheless, a large number of "subjective numerical estimation" methods are in use in health status preference measurement. Since the designation of "direct" versus "indirect" scaling approaches seems to be well-founded in the health status preference measurement literature, it was decided to adopt this terminology for this dissertation.

Within each general class of approaches for scaling stimuli (i.e., direct and indirect), specific scaling methods may be used. These scaling methods have been developed in two major fields -- psychophysics and decision
theory. Before discussing each scaling approach, these two major fields will be introduced.

Psychophysics

Psychophysics is the study of the relationship between physical stimuli and the psychological sensation or perception of these physical stimuli. Gustav Theodor Fechner was the founding father of the field of psychophysics, with the publication of his book *Elemente der Psychophysik* in 1860 (Fechner, 1860). Most commonly, psychophysics is concerned with the correlation between physical measurements of different stimuli and reactions to these stimuli, e.g., decibels of sound intensity and perceptions of loudness. With nonphysical stimuli such as health states, psychophysics is concerned with the most valid and reliable representation of psychological perceptions of different dimensions or attributes of these stimuli. Therefore, the task of psychophysicalists is to find the numerical laws that relate psychological dimensions to physical dimensions. Using the notation of Figure 3, these laws take the general form,

\[ S = H(s) \]  

(3.6)

where

- \( S \) = the psychological dimension,
- \( s \) = the physical dimension, and
- \( H \) = numerical function that expresses the relation between physical and physical dimensions (i.e. the psychophysical function).
The psychophysical approach to solving this problem is to establish special psychological procedures for measuring sensations. Fechner’s (1860) procedure for measuring psychological sensations was based on the work by Weber on difference thresholds, or “confusability” between stimuli. According to Weber, confusability is proportional to the stimulus magnitude at which it is measured. In other words, the ratio between two physical stimulus magnitudes that are correctly discriminated by a subject is constant. Using this principle, Fechner (1860) conjectured that if two stimuli were just noticeably different, then regardless of changes in physical intensity of the stimuli, the psychological difference must be equal. Employing the just noticeable difference as his unit of psychological magnitude, Fechner (1860) established his psychophysical law which stated that physical stimuli and the psychological sensations in response to these physical stimuli were related by a logarithmic function, or

\[ S = k \log(s) \]  

(3.7)

where

- \( k \) = constant that varies according to the phenomena being measured
- \( S \) = the psychological dimension
- \( s \) = the physical dimension

While both Weber’s and Fechner’s laws required the existence of some physical continuum (s) against which sensory evaluations were to be made, Thurstone’s extension of Fechnerian methods -- which is known as
Thurstone’s Law of Comparative Judgment -- allowed psychophysicists to work with nonphysical stimuli (Thurstone, 1927a). Thurstone’s extension thus allowed researchers to develop psychological scales for various qualitative or subjective dimensions like attitudes, values, and interests. Like Weber and Fechner, Thurstone assumed some reference continuum regarding a stimulus attribute of interest. For each stimulus, the subject is assumed to possess a discriminal (psychological) process that is not directly observable. The discriminal process is assumed to have a modal (i.e., most frequently occurring) value on the subject’s psychological continuum and some dispersion around the mode. Thus, a given stimulus $s_j$ does not always produce the modal response $S_j$. If the discriminal dispersions of two different stimuli have average scores which are close together and their variances are sufficiently large, these dispersions will overlap. If repeated measures are taken and the shapes and variances of the dispersions are known, a measure of the psychological distance can be calculated.

Therefore, like Weber’s Law, Thurstone’s Law relies on subjective “confusions” regarding a subject’s evaluation of which of a pair of stimuli possesses more of some desired (or undesired) attribute or dimension, e.g., which health state is “better” in terms of physical function. The proportion of times that one health state is evaluated as better than another is assumed to be functionally related to how far separated the modes of the two discriminal processes are, the dispersions of the discriminal processes, and their correlations. By making various assumptions about a subject’s discriminal processes, Thurstone (1927a) was able to scale qualitative attributes such as handwriting esthetics from observed proportions of the
cases in which each member of a stimulus pair was chosen has having subjectively more of the designated property. Thurstone suggested five different versions of the Law of Comparative Judgment on the basis of the assumptions made about subjects' discriminative processes, of which the most commonly used is the Case V assumption where it is assumed that the dispersions are normal, of equal variance, and uncorrelated (Thurstone, 1927a).

The Fechnerian tradition stood for nearly a century in the field of psychophysics, until Stevens put forth an alternative psychophysical law in the late 1950s on the basis of his experiments (Stevens 1957, 1960, 1961, 1971). Stevens argued that Fechner had derived the wrong psychophysical law because he had used a faulty scaling method. Any mathematical relationship between two variables requires that one measure the underlying stimulus and response scale on an interval-level or ratio-level scale. Stevens contended that the Fechnerian methods did not produce linear or interval-level response scales. As a result, Stevens argued that Fechner's psychophysical relationships were meaningless (Stevens 1957, 1960, 1961, 1971). When Stevens substituted a subjective ratio method of scaling for Fechner's just noticeable difference method, he found that the psychophysical law was more appropriately specified as a power function, or

\[ S = s^n \]  

(3.8)
where

\[ n = \text{a unique exponent that varies with the property being measured} \]
\[ S = \text{the psychological dimension} \]
\[ s = \text{the physical dimension} \]

The simple statement of this law is that psychological magnitude increases in geometric proportion. The significance of the Fechner versus Stevens controversy lies in terms of the "correct" form of the psychophysical law or the relative merit of Fechner or Stevens. Rather, as Patrick and Erickson (1993) have noted, the importance of the controversy lies in the disparate results obtained by using the specific scaling methods developed under either tradition. Since these specific scaling methods yield different results, the question arises as to which method is the "correct" one. Unfortunately, this issue has not yet been resolved. Therefore, it is necessary to discuss scaling methods developed under each of these traditions. Scaling methods developed under the Fechnerian tradition include the method of category scaling or the rating scale method. The scaling method of paired comparisons was developed as a result of Thurstone's Law of Comparative Judgment. Finally, the scaling method of magnitude estimation was developed under the Stevensonian tradition of psychophysical measurement. Each of these scaling methods will be discussed later in this chapter.
Decision Theory

The other field responsible for the development of scaling methods is decision theory. Decision theory is concerned with the normative study of decision-making, i.e., how decisions should be made. It has its roots in the disciplines of economics, psychology, and mathematical logic. A critical component of decision theory is the concept of utility or the measure of worth associated with a given construct, object, or action. The classical theory of utility was derived from the eighteenth century work of Jeremy Bentham, who formulated the utility principle that all individuals and society, as an aggregate of individuals, are directed toward a single goal — to increase pleasure and/or decrease pain (Bentham, 1789). These pleasure-and pain-giving properties of any construct, object or action were termed as the utility associated with the construct, object or action in question. Bentham (1789) essentially argued that since the goal was to seek the maximum utility, people choose between alternatives so as to fulfill this goal. According to this utilitarian philosophy, preferences can be measured on a cardinal (i.e., interval or ratio) scale. This concept was subsequently incorporated into Alfred Marshall's *Principles of Economics* (Marshall, 1890) and was useful in developing the theories of general economic equilibrium and interpersonal comparisons and the maximization of welfare as the sum of cardinal utilities (see Jevons, 1871; Menger, 1871; Walras, 1874).

However, the initial attempts at measuring utility on a cardinal scale met with limited success. Therefore, economists in the early twentieth century developed the notion of ordinal utility, wherein people report how much of one commodity is equal to standard amounts of a second
commodity. Edgeworth initially provided a major departure from the classical theory of utility by showing that preferences between collections of goods can be represented by indifference curves without relying on the assumption of cardinal utilities for individual choices based on the assumption of independence (Edgeworth, 1881). Pareto built on this work, by showing that indifference curves could be inferred from an ordinal utility function from preference rankings based on certain assumptions regarding the consistency of choices (Pareto, 1927). Although there were certain inconsistencies in Pareto's formulation, these were resolved by Hicks and Allen (1934) who showed that indifference curves together with given price ratios were sufficient to establish an individual's point of maximum utility. This notion of ordinal utility differs from classical utility in the point of origin and the arbitrary units of scale. However, it has been shown to obtain many of the same results as the classical theory of utility in applications (Blaug, 1983). An excellent review of the development of the classical and ordinal approaches to utility has been provided by Stigler (1950a, 1950b). It should be noted that both these approaches to utility (i.e., classical and ordinal) assume that the outcomes are known with certainty. However, since these approaches have not been used for the purpose of developing measures of health status preferences, they will not be further discussed. For the purpose of health status preference measurement the notion of expected utility is of importance, as discussed below.

In 1944, John von Neumann and Oskar Morgenstern developed the notion of expected utility as a way of incorporating uncertainty into preference measurement (von Neumann and Morgenstern, 1944). As these
researchers showed, if certain outcomes or consequences or alternatives (e.g.,
health states) are presented as gambles and if certain axioms are satisfied for
expressions of preference, then such preferences can represent the
underlying mental structure as a cardinal utility function. Game theory, as
von Neumann and Morgenstern named their theory, centers on the
assumption that the utility of two objects or alternatives is reflected by the
relative expectations for the occurrence or acquisition of the two alternatives.
By constructing various gambles or game situations, researchers can
determine the expected utility associated with each alternative. It should be
noted that the expected utility so determined will have up to interval-level
properties and will incorporate the notion of risk or uncertainty.

von Neumann and Morgenstern proposed that if people follow
certain axioms, then the person faced with different lotteries (i.e., gambles or
bets) will choose the lottery with the highest expected utility. Three axioms
of rational behavior govern game or expected utility theory:
(a) transitive ordering of consequences -- a decision maker is able to state his
or her preferences between consequences of a lottery so that these
preferences are transitive, i.e., if consequence A is preferred to consequence
B and consequence B is preferred to consequence A, then consequence A is
preferred to consequence C (or if A > B and B > C, then A > C),
(b) independence or substitutability -- any compound (two-stage) lottery is
equivalent to a simple lottery with the same final consequences and with
probabilities computed from the component probabilities according to the
ordinary probability calculus (i.e., in any lottery L, the lottery L_{new} can
replace the consequence A -- provided the two are indifferent to the decision maker -- thus making the original simple lottery a compound lottery), and (c) continuity -- if the decision maker's order of preferences for consequences is A>B>C, then there is some probability p between 0 and 1 such that he or she is indifferent between consequence B for sure and a lottery giving consequence A with probability p and consequence C with probability (1-p).

It has been shown that if an individual's preferences satisfy these three axioms, he or she can encode his or her preferences in a utility function that assigns a utility number to every consequence in the lottery such that the utility of the lottery is equal to the expected utility of its component consequences (this property is referred to as the expected utility hypothesis) and that the rational action would be to choose the lottery with the highest expected utility (Baumol, 1972). A utility function derived from these axioms is linear in probability and unique up to a positive linear transformation, i.e., the results are unaffected if all utility scores u(Oj) are replaced by the quantity \(a + b[u(Oj)]\), where a is any constant and b is any positive constant. In other words, the utility function is measured on an interval-level scale of measurement.

In the von Neumann-Morgenstern formulation, uncertainty was treated as objective, i.e., it was assumed that the probability distribution over consequences was well-specified. Savage (1954) generalized the von Neumann-Morgenstern formulation to include subjective uncertainty. Expected utility theory forms the basis for the gamble methods for
measuring health status preferences, which are discussed later in this section.

A different contribution to the field of decision theory has come from welfare economics. In 1844, Dupuit proposed the notion of consumer surplus as a measure of social benefit obtained from the development of collective goods, i.e., goods which are used collectively by a society, such as roads and bridges (Dupuit, 1934). Dupuit (1934) observed that the price people were willing to pay for such goods was, in most cases, greater than the price actually paid. He termed this differential in price as the consumer surplus. Alfred Marshall (1890) later refined Dupuit’s original work, linking it with the notion of utility. The notion of consumer surplus is used along with the concept of Pareto optimality (i.e., no one can be made better off without making someone else worse off) in societal decision-making. Welfare economists often use the contingent valuation method to determine the value people place on non-market goods or services contingent on a hypothetical market existing in which to trade this good or service. Since there have been some applications of the contingent valuation method to the measurement of health status preferences, this method will also be discussed later in this section. Apart from the scaling methods based on the principles of psychophysical and decision theory, there have also been some ad-hoc scaling methods used to measure health status preferences. These ad-hoc methods will also be briefly reviewed in this section.
In discussing the different scaling methods, the methods adopting the direct scaling approach will first be discussed, followed by the methods following the indirect approach. Each scaling method will be discussed with respect to five criteria: (a) character of the response expected from the subject, i.e., quantitative judgments (more specifically, judgmental ratings or pricing judgments), indifference judgments, or choices, (b) number of attributes or dimensions (i.e., stimuli) simultaneously involved in the scaling method, (c) whether the stimuli can be described as discrete or continuous, (d) extent to which the scaling method incorporates the notion of risk, and (e) level of measurement of the subject’s response, i.e., nominal, ordinal, interval, or ratio.

**Direct Scaling Approaches**

Direct scaling approaches can be considered to be methods in which the step between the raw data and the final scale is as short as possible, because when respondents provide interval-level or ratio-level data the investigator can derive scale scores via relatively simple computational procedures such as averaging across respondents (Torgerson, 1958). Although there were initial concerns about whether respondents would be able to quantify subjective feelings (Kneppreth, Gustafson, Leifer, et al., 1974), these concerns appear to be unfounded. A practice session before the actual evaluation session has been shown to help respondents understand the task required of them and therefore help them in quantifying their subjective feelings. The vast majority of scaling approaches used in the
health status preference literature have used these approaches. Under the class of direct scaling approaches, there are six scaling methods which have been commonly used in health status preference measurement -- rating scale (including visual analogues) methods, the magnitude estimation method, gamble methods, the time trade-off method, contingent valuation methods, and the equivalence method (also called the person trade-off method). Each of these methods will be discussed in turn.

**Rating Scale Methods**

Rating scale methods are also known as category scaling methods. They were developed in the field of psychophysics as part of the Fechnerian tradition of psychophysics. These are considered by some researchers to be the simplest methods for measuring health status preferences. The category scale is a visual analogue scale where each step is considered to be an equal interval, i.e., the distance from 0 to 5 is the same as the distance from 20 to 25, 75 to 80, etc. The range of the scale is defined by the researcher or the respondent, i.e., the researcher or respondent specifies the endpoints of the scale. Hence, these methods have also been referred to as “double-anchored estimation methods” (Kneppreth, Gustafson, Leifer, et al., 1974).

It should be noted that not all psychophysicists consider the rating scale method as a direct scaling approach. Indeed, Torgerson (1958) has argued these methods lie in between direct and indirect scaling approaches, because the data collection proceeds along similar lines as direct scaling approaches while the data analysis is often based on the Fechnerian (1860) tradition and Thurstone’s (1927a) Law of Comparative Judgment. However,
as Jones (1974a, 1974b) has noted, "If we should assume that numbers are but instances of categories, then the distinction between direct and categorical methods becomes tenuous indeed. Furthermore, data treatment can, and does, take the form of either the direct or indirect methods. That is, one may either take category placements at their face value, calculate mean category placements of stimuli, and use the resulting numerals as though they were on a scale of measurement, or one may rely on the Law of Comparative Judgment, calculate category boundaries, and use the resulting distances (from boundary to boundary) as measures (p.347)."

Indeed, both these approaches to data analysis or treatment have been used in the psychophysical literature. For instance, Anderson (1974a) and Parducci and Perrett (1971) have used the former (i.e., direct) approach, while Bock and Jones (1968) have used the indirect approach. For the purpose of this dissertation, it was decided to include these methods as part of the direct scaling approaches because most of the published applications of rating scale methods to health status preference measurement have adopted the direct scaling approach to data analysis.

The rating scale method can be used as part of any of the different measurement strategies discussed earlier, i.e., holistic, conjoint, or explicitly decomposed preference measurement strategies -- with the exception of the discrete choice conjoint strategies. This method usually involves the following steps. At first, the respondent ranks the different stimuli (i.e., health states or attribute levels, or for continuous attributes specifies the most preferred and least preferred levels). The next step involves anchoring the scale at the extreme levels. When evaluating health states, the anchors
are usually labeled "death" and "perfect health," although this is not always necessary, especially given the observation that respondents often consider some states of health less preferable to death (see, for e.g., Kind and Rosser, 1979, Rosser and Kind, 1978; Torrance, 1984). The respondent then rates the desirability of each stimulus by placing it on the line between the anchors. To achieve an interval scale, respondents must be instructed to place the stimuli on the line such that the intervals between the placements reflect the differences they perceive between the stimuli.

A commonly used variation of the rating scale method is the method of equal-appearing intervals or category ratings. In this method, respondents sort the stimuli into a specified number of categories assuming equal changes in preferences between adjacent categories. Visual aids may be used with either form of rating scale methods. For instance, the McMaster group of researchers have employed a “feeling thermometer,” which is essentially a rating scale shaped like a thermometer along which respondents place foam sticks labeled with the stimuli. In the equal-appearing intervals method, respondents may actually sort cards labeled with the stimuli into piles, or they may simply assign a category number to each stimulus.

Most category judgments are converted into a 0-1 scale with the least preferable anchor stimulus at 0 and the most preferable anchor stimulus at 1. For instance, if the stimuli were health states (i.e., the rating scale method was used as part of a holistic preference measurement strategy), death is usually set at 0 and perfect health is usually set at 1. In this case, the preference value for each of the other health states is simply the scale score associated with its placement. However, if death is not considered to be the
least preferable state (i.e., there are some health states considered worse than death) and is placed at some intermediate point on the scale -- say \( d \) -- the preference values for the states are derived from the formula \( \frac{x - d}{1 - d} \), where \( x \) is the scale placement of the health state under evaluation (Torrance, 1986, 1987).

In terms of the five criteria identified earlier, the rating scale method: (a) depends on quantitative judgments -- specifically, judgmental ratings -- from respondents, (b) involves three stimuli at a time (i.e., the two anchors and the stimulus under evaluation), (c) works with either discrete or continuous stimuli, (d) does not incorporate the notion of risk (i.e., measures values, not utilities), and (e) provides responses on either an ordinal- or interval-level scale. This last point is deserves further comment. The level of measurement of the response obtained from rating scale methods has been a source of controversy in the field of psychophysics for many decades. This issue is related to the direct approach versus indirect approach controversy noted earlier, for those who believe that rating scales give interval-level responses adopt a data analysis plan in accordance with direct scaling approaches, while those who believe that rating scales give ordinal-level responses adopt a data analysis plan in accordance with indirect scaling approaches. It should, however, be noted that experiments using functional measurement as a means of testing for equal intervals have repeatedly shown that rating scale methods have interval-level properties (see, for e.g., Anderson, 1976; Weiss, 1975). This supports the view that rating scale methods can be considered under direct scaling approaches which, as noted earlier, is the view adopted in this dissertation. With respect to the health
status preference measurement strategies with which it can be used, the rating scale method can be used with all the strategies discussed except the rank order judgment strategies (if the responses are considered to have interval-level properties), discrete choice conjoint strategies, and the analytic hierarchy process.

The rating scale method is one of the most frequently used methods of scaling health state preferences, mainly on account of its simplicity in use and interpretation. Further, it can provide a great deal of information in a relatively short period of time. Moreover, it is a very flexible method and can be easily modified to fit specific situations. Indeed, it has been used for scaling acute and chronic health states as well as health states worse than death. In health care, it is the scaling method of choice of the San Diego group of researchers, who have extensively used it to measure health status preferences as part of the Quality of Well Being Scale (see Kaplan and Anderson, 1990). It has also been extensively used by the members of the EuroQol group in both versions of the EuroQol (see The EuroQol Group, 1990). It has also been used by the McMaster group of researchers for the generation of value functions for health states (see Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992). Apart from these public policy applications, the method has been extensively used in several studies focusing on clinical decision making in a variety of patients and practice settings (see, for e.g., Christensen-Szalanski, 1984; Llewellyn-Thomas, Sutherland, Ciampi, et al., 1984; Llewellyn-Thomas, Sutherland, and Theil, 1993; Llewellyn-Thomas, Sutherland, Tibshirani, et al., 1984; O'Connor, Boyd, Warde, et al., 1987).
Rating scale methods have been criticized as being prone to the problem of distribution and contextual effects. The distribution effect is the tendency to distribute responses across the scale and use all categories equally often. Stevens and Galanter (1957) have argued that the rating scale method is biased because subjects attempt to use each category equally often, spreading out the ratings when the stimuli are actually close together and pushing them together when the true scores are far apart.

Contextual effects occur when mean ratings for individual items differ when they are presented in different contexts. Evidence for contextual effects has been presented by Parducci, who asked subjects to rate the seriousness of certain offenses using a rating scale with five categories (Parducci, 1968). Subjects were randomly assigned to two groups. Half the subjects received a list of relatively minor offenses while the other half received a list of offenses which "could be counted on to evoke strong disapproval." Each list contained six items which were common to both lists. Parducci found that the six common offenses were judged more leniently when they were included in the list consisting mainly of severe offenses (Parducci, 1968).

Parducci explained these distribution and contextual effects in terms of a range-frequency model (Parducci, 1963, 1965, 1968, 1974). According to this model, subjects first estimate the upper and lower boundaries of the stimulus range. Then they divide the stimulus range into a set of equal ranges, one for each category. If the stimuli are not presented in random order, some categories will be used more frequently than others. When this happens, subjects adjust the boundaries of the categories so that each
category will be used equally often. In other words, the ranges of categories that are used less often will be increased and the ranges of categories that are used more often will be decreased (Parducci, 1963, 1965, 1968, 1974).

Kaplan and Ernst (1983) investigated the veracity of these criticisms of distribution and contextual effects of rating scales as applied to the measurement of health status preferences. Through a series of experiments, they showed that with careful administration, health status preference measurements can be obtained from rating scale methods without risk of response distribution and contextual effects (Kaplan and Ernst, 1983). It is important to note the role of careful administration of rating scale methods in decreasing the incidence of distribution and contextual effects. For instance, according to Parducci's explanation, contextual effects may occur because subjects develop and adapt their frame of reference as they proceed. This implies that contextual effects may be limited to situations in which the subjects are not already familiar with the end points of the stimulus condition. Kaplan and Ernst (1983) showed that by anchoring the rating scale at the most and least preferable stimuli and choosing these anchor stimuli so that they were familiar to respondents (e.g., anchoring the scale with death and perfect health), it was possible to prevent the existence of contextual effects.

Further, Kaplan and Ernst (1983) showed that distribution effects are most likely to occur when the stimuli are not along some well-defined value continuum, the end points of the response continuum are not clearly defined, and all the stimuli can be inspected before they are rated. This suggests that clearly defining the end points (as for the contextual effects
above) of the scale and not allowing subjects to inspect the stimuli before rating them, it is possible to prevent the occurrence of distribution effects in rating scale methods.

Several studies have estimated the reliability of the rating scale method. With respect to the scaling of health status preferences, reliability may be assessed in three different ways -- intra-rater reliability (i.e., a single rater's consistency when a stimulus is presented more than once), test-retest reliability (i.e., the stability of scale scores over short periods of time), and inter-rater reliability (i.e., consistency between raters regarding scale scores). It should be noted that internal consistency reliability, or the consistency in response from item to item within a scaling task, is not applicable to the scaling of health states. This is because there is no reason to expect high intercorrelations between the stimuli -- indeed, this would be undesirable. Internal consistency is important in situations where a series of items are used to scale people -- not stimuli -- along a particular dimension (Froberg and Kane, 1989b).

Rating scale methods have been shown to have acceptable reliability, as estimated by these indicators. For instance, estimates of the intra-rater reliability range from 0.70 to 0.94 (Cadman and Goldsmith, 1986; Patrick, Bush, and Chen, 1973b; Torrance, 1982b), one-week test-retest reliability has been estimated as 0.77 (O'Connor, Boyd, and Till, 1985), and the inter-rater reliability has been estimated to be between 0.75 and 0.77 (Patrick, Bush, and Chen, 1973b). It should be noted that one study has determined the one-year test-retest reliability to be rather low -- 0.49 (Torrance, 1976a). However, it is unclear whether the low reliability coefficient reflects measurement error or
a true change in preferences. Further, a different interviewer was used in the second interview in this study, which may have affected the reliability of the observed results.

These results, coupled with the positive aspects of rating scale methods (e.g., ease in implementation and interpretation), suggest that rating scale methods are a viable alternative for measuring health status preferences in that they are capable of providing valid and reliable data. However, rating scale methods do not permit the incorporation of risk, i.e., they can only be used to measure values, not utilities. Another major drawback of these methods is that they cannot be used as the scaling method when a discrete choice conjoint strategy is adopted for measuring health status preferences, because their response variable is judgment -- not choice.

The Magnitude Estimation Method

Magnitude estimation is a psychophysical direct scaling method developed by Stevens in accordance with his power law interpretation of the psychophysical function (as explained earlier in this chapter). Although the origins of this method are usually attributed to Richardson (1929) in his work on the intensity of fading afterimages where judgments were always in terms of fractions, it was Stevens who operationalized and developed the method of magnitude estimation as it is used in its current form. Using magnitude estimation, the subject is given a standard stimulus and asked to provide a number or ratio indicating how much better or worse the stimulus under evaluation is compared to the standard stimulus. Since the scale is anchored at this standard stimulus, this method has also been referred to as
the “single anchored estimation method” (Kneppreth, Gustafson, Leifer, et al., 1974).

This method usually involves three steps. The first step is the same as that for rating scale methods, i.e., the subject rank orders the stimuli under consideration or for continuous stimuli the subject specifies the least and most preferable stimuli. The second step is the main one, where the actual comparison takes place. In this step, one stimulus is chosen as the standard or reference, and all other stimuli are compared to this standard stimulus in a random order in terms of how much more (or less) these stimuli are preferred compared to the standard stimulus. In the final step, the standard stimulus is assigned an arbitrary score and numerical scores are estimated for each of the evaluated stimuli according to the ratio stated in the second step (Kneppreth, Gustafson, Leifer, et al., 1974).

This method can also be used with certain variations. For instance, in case of continuous stimuli, one can estimate stimulus magnitudes instead of ratios. Using physical function as an example, the subject could be asked to specify walking how many blocks without any assistance would be preferred twice as much as the standard stimulus of walking one block without any assistance. Other variations pertain to the handling of the standard stimulus. For instance, the researcher may decide to present the standard stimulus only at the beginning of each experimental session rather than before each judgment. Further, there is the issue of the position of the standard stimulus, i.e., at the top, bottom, or middle of the scale. Indeed, published applications of this method in the health-care literature have been inconsistent in the selection of a standard health state -- in three studies, the
standard was taken from the end of the scale, defined as the least ill state, the healthiest state, or the absence of discomfort or dysfunction (Haig, Scott, and Wickett, 1986; Patrick, Bush, and Chen, 1973b; Rosser and Kind, 1978); whereas in another study the standard was taken from the middle of the scale (Kaplan, Bush, and Berry, 1979). It has, however, been shown that in such cases the resulting scales will “not be enormously different” (Jones and Woskow, 1962; Stevens, 1956).

Jones (1974a, 1974b) has noted that it is even possible to eliminate the standard stimulus completely and let each subject settle down to his own modulus — in this case the data from different subjects may be reduced to the same scale by application of appropriate constants. However, there may be problems in agreeing upon a common basis for all respondents’ judgments, which in turn would affect the reliability of the scale scores obtained. Indeed, Kneppreth, Gustafson, Leifer, et al., (1974) have cautioned against the use of such a variation. It is important to note that whatever variation is used, care must be taken to ensure that the subject is not constrained by the given scale. It has been reported that the use of logarithmic scales may help alleviate this problem, since a wide range of scores are possible with such scales while small differences can still be detected (Gustafson, Feller, Crane, et al., 1971).

In terms of the five criteria identified earlier, the magnitude estimation method: (a) depends on quantitative judgments — specifically judgmental ratings — from respondents, (b) involves two stimuli at a time (i.e., the standard stimulus and the stimulus under evaluation), (c) works with either discrete or continuous stimuli, (d) does not incorporate the
notion of risk (i.e., measures values, not utilities), and (e) provides responses on a ratio-level scale. With respect to the health status preference measurement strategies with which it can be used, the rating scale method can be used with all the strategies discussed except the rank order judgment strategies, discrete choice conjoint strategies, and the analytic hierarchy process.

Magnitude estimation has had limited applications in the health status measurement literature. Indeed, the only applications of magnitude estimation to the measurement of health status preferences was in the 1970s (Patrick, Bush, and Chen, 1973a). Most of these applications were by the San Diego group of researchers, who were interested in deciding upon a scaling method to use to measure health status preferences for the Quality of Well Being scale (Kaplan, Bush, and Berry, 1979; Patrick, Bush, and Chen, 1973b). In this regard, magnitude estimation was but one of several alternative scaling methods tested -- and subsequently rejected (Kaplan, Bush, and Berry, 1979).

The few studies which did estimate the reliability of the method found it to have acceptable reliability. One study reported an intra-rater reliability of 0.74-0.83 and an inter-rater reliability of 0.75-0.79 (Patrick, Bush, and Chen, 1973b), while another computed an intra-rater percentage agreement as 97.2% and an inter-rater percentage agreement as 88% (Rosser and Kind, 1978). One study attempted to validate the magnitude estimation method by correlating the scores obtained from it with court awards for damages in personal injury claims in British courts (Kind, Rosser, and Williams, 1982). It was found that the legal scale was highly correlated
(r=0.82) with the magnitude estimation scale (Kind, Rosser, and Williams, 1982).

In spite of this evidence supporting the reliability and validity of the magnitude estimation method, it has been rarely used in health status preference measurement. Kneppreth, Gustafson, Leifer, et al., (1974) have listed some of the drawbacks of the method, which may explain its lack of popularity in the health status preference measurement literature. First, the method is useful only when the subject is familiar with the estimation of ratios. In this regard, the respondent estimating the ratios may not be able to judge ratios well, since some people tend to overestimate ratios while others tend to underestimate them. This suggests that the method may not be feasible in all patient populations. Second, Kneppreth, Gustafson, Leifer, et al., (1974) have noted that the method is quite unpopular with respondents, who feel more uncomfortable estimating in ratios than in any other mode of preference measurement. Although this problem may be alleviated by training and discussion of the method with the respondent, this is slow process and yet might not be successful. Third, this method is not very sensitive to small differences in ratios. For instance, the respondent may not be able to differentiate between 14/16ths and 15/16ths, which is about a six-point difference on a 100-point scale (Kneppreth, Gustafson, Leifer, et al., 1974). This would result in lack of resolution or degree of error inherent in the method. Finally, experiments using functional measurement as a means of testing for equal intervals have shown that scales derived using the magnitude estimation method do not meet this empirical criterion (Anderson, 1976; Weiss, 1975).
Further, magnitude estimation methods do not permit the incorporation of risk, i.e., they can only be used to measure values, not utilities. Another major drawback of these methods is that they cannot be used as the scaling method when a discrete choice conjoint strategy is adopted for measuring health status preferences, because their response variable is judgment -- not choice.

Gamble Methods

Unlike the previous two scaling methods discussed, gamble methods were explicitly designed to incorporate the notion of risk. Developed in the field of decision theory, these methods are based on the axioms of expected utility theory (as discussed earlier). A gamble or wager or lottery is a situation where each one of a set of outcomes or consequences can occur with a given probability (Kneppreth, Gustafson, Leifer, et al., 1974). The traditional or standard gamble methods rely on the respondent's ability to choose between such a wager and an alternative with no risk involved (i.e., a sure thing). Apart from these standard gambles, there are other gamble methods which involve a choice between two gambles, i.e., there is no sure thing alternative. These other gamble methods are known as lottery equivalents (McCord and de Neufville, 1986). Both of these different types of gamble methods will be discussed in this section, although the emphasis will be on the standard gamble since this is the version which has been used most commonly in health status preference measurement. Indeed, there have been no published applications of lottery equivalents to the measurement of health status preferences.
There are two different ways in which to implement both these gamble methods in order to develop preference functions for respondents. First, the probabilities in the gamble can be varied until the respondent is indifferent between the two alternatives, i.e., the gamble and the sure thing for a standard gamble or the two gambles for the lottery equivalent. At this point, it is assumed that the expected utility of the two alternatives are equal. These approach is called the variable probability or probability equivalence approach. Second, the probabilities in the gamble can be held constant and one of the consequences can be varied until the respondent is indifferent between the two alternatives. This approach is called the constant probability or certainty equivalence approach.

Gamble methods are usually implemented in four steps. First, in case of discrete stimuli, the respondent must rank order all the stimuli or at least specify a most preferred and least preferred stimulus. For continuous stimuli, only the most and least preferred stimuli need to be specified. It should be noted that the certainty equivalence approach requires that stimulus levels and preference be related -- if they are not, the certainty equivalence approach cannot be used. This is necessary since the consequences are varied in this approach. For instance, if the respondent prefers a "mediocre" stimulus to "extreme" stimuli (e.g., hypothetically, if the respondent prefers a state of halfway-perfect health to one of perfect health), he or she will have a nonmonotonic preference function and the certainty equivalence approach will be inappropriate. The second step is to choose a scale and assign numerical scores to the most and least preferred stimuli. The scores may be arbitrary, since the final scores obtained have no
physical meaning -- only the distance between the points is useful. The next step involves the construction of a gamble or lottery for presentation to the respondent. Depending on the particular approach adopted to implement the gamble, the nature of the gamble will be different.

With the probability equivalence approach a gamble is constructed which involves a choice between two alternatives. The first alternative is a wager involving the most and least preferred stimuli while the second is the stimulus under evaluation, which is presented as a sure thing. For instance, the wager could be between the consequences of being in perfect health (with probability $p$) or immediate death (with probability $1-p$). It should be noted that it is not necessary to always use the most and least preferred stimuli as consequences of the wager -- the only requirement is that the sure thing alternative be more preferred to one of the consequences of the wager and less preferred to the other. It can be seen that if the sure thing is preferred to both the consequences of the wager, it will always be preferred to the wager regardless of the probabilities. The opposite holds true if both wager consequences are preferred to the sure thing. No point of indifference can be found in either case. Another necessary characteristic of the wager is that is should be exhaustive, i.e., the probabilities in the wager should add to one.

The respondent is asked to indicate the value of $p$ at which he or she would be indifferent between the sure thing and the wager. This however, is not accomplished by asking the respondent an open-ended question about the value of $p$. Instead, the sequence of probability values is varied in a converging ping-pong manner by alternating between high and low values.
(e.g., 10/90, 90/10, 20/80, 80/20, etc.). This convergent cascading addresses the concerns of Tversky and Kahneman, who pointed out the need to reduce the possibility of an anchoring bias where a strategy of constantly increasing/decreasing encourages respondents to overestimate or underestimate their indifference point (see Kahneman and Tversky, 1979, 1982; Tversky and Kahneman, 1981). It also reduces the framing effect associated with consistently increasing or decreasing probabilities, which are often interpreted by respondents as gains or losses relative to a point of reference (see Kahneman and Tversky, 1979, 1982; Tversky and Kahneman, 1981). Visual aids such as the probability wheel and the chance board have been developed to help respondents understand the notion of probability in making their evaluations (Furlong, Feeny, Torrance, et al., 1990). After the respondent's indifference point has been determined in this manner for the stimulus under evaluation, another stimulus is presented as the sure thing and the procedure is repeated. In this way, preferences for each of the stimuli under evaluation can be measured using the probability equivalence approach. The utility score for each stimulus evaluated in this manner is given by the formula:

\[ u \text{ (sure thing)} = p \ u_1 + (1 - p) \ u_2 \]  \hspace{1cm} (3.9)

where

- \( u_1 \) = utility of the most preferred stimulus
- \( u_2 \) = utility of the least preferred stimulus
- \( p, (1 - p) \) = final probabilities at the indifference point
Since all the unknown values in this formula are determined by the researcher, the utility of each stimulus under evaluation can be calculated. It should be noted that the computational task is greatly simplified by using zero as one end of the scale, since either the utility of the most or least preferred stimulus will be zero and will therefore be left out of the calculations.

The certainty equivalence approach differs from the probability equivalence approach with respect to the construction of the gamble. With the certainty equivalence approach, the first alternative of the gamble is a wager between the most and least preferred stimuli (as for the probability equivalence approach). However, in this case, the probability of occurrence of these stimuli is fixed, usually at 0.5 for each stimulus. The other alternative another stimulus, presented as a sure thing. In this case, the sure thing is varied until an indifference point is found between the two alternatives. At that point, the utility of the sure thing alternative can be calculated by the following formula:

\[
    u \text{ (sure thing)} = p_1 u_1 + p_2 u_2 ,
\]

where

- \( p_1 \) = fixed probability of the most preferred stimulus
- \( p_2 \) = fixed probability of the least preferred stimulus = \( 1 - p_1 \)
- \( u_1 \) = utility score of the most preferred stimulus
- \( u_2 \) = utility score of the least preferred stimulus
Since all these parameters are known (the values at which the probabilities are fixed are determined by the researcher; so are the arbitrary scores of the most and least preferred stimuli), the utility of the sure thing stimulus can be calculated. In order to determine the preference scores for the remaining stimuli, the wager is varied using two of the stimuli whose utilities have already been determined. In this way, preferences for each of the stimuli under evaluation can be measured using the certainty equivalence approach.

In terms of the five criteria identified earlier, the gamble methods: (a) depends on indifference judgments from respondents, (b) involves three stimuli at a time (i.e., the two consequences of the wager and the stimulus under evaluation or the sure thing), (c) works with either discrete or continuous stimuli (although the certainty equivalence approach works best with continuous stimuli), (d) incorporates the notion of risk (i.e., measures utilities, not values), and (e) provides responses on an interval-level scale (according to the axioms of expected utility theory). With respect to the health status preference measurement strategies with which they can be used, the gamble methods can be used with all the measurement strategies discussed with the exception of all the conjoint methodology strategies and the analytic hierarchy process.

Although interviewers and respondents have often reported that the gamble methods are difficult to administer and understand, many investigators favor these methods on account of the fact that the standard gamble is built on the axioms of expected utility theory and it forces the respondent to make preference judgments under conditions of uncertainty.
Further, the standard gamble has been said to yield interval-level scales (Luce and Raiffa, 1957), although such claims are more definitional as opposed to empirically demonstrated (Veit and Ware, 1982).

While the standard gamble has been viewed by some researchers as the "gold standard" in health status preference scaling methods on account of its theoretical grounding in expected utility theory (see Torrance, 1976a), other investigators have turned to other scaling methods on account of the drawbacks of the standard gamble. One of the major reasons for this is the accumulation of evidence that people exhibit patterns of preference that are incompatible with expected utility theory. For instance, Schoemaker (1982) has presented extensive evidence that people violate the axioms of expected utility theory at the individual level. In health-care, Llewellyn-Thomas, Sutherland, Tibshirani, et al. (1982) have found that changes in the gamble consequences significantly influenced reported scores for health states, a finding which contradicts the substitutability axiom of expected utility theory. The observation of significant departures of human behavior from expected utility theory led to Kahneman and Tversky proposing an alternative theory of decision-making, which describes how people actually make decisions under risk -- prospect theory (see Kahneman and Tversky, 1979, 1982; Tversky and Kahneman, 1981).

It should, however, be noted that the mere finding that people do not behave according to the axioms of expected utility theory is not sufficient cause to disavow the standard gamble method. Indeed, expected utility theory is a normative theory of how people should behave if they want to maximize their utility. Expected utility theory makes no claim of being a
descriptive theory of human behavior. Therefore, the observation that people do not behave according to the axioms of expected utility theory does not in any way contradict the postulates of expected utility theory.

However, there are other drawbacks of the standard gamble method. Hershey and Schoemaker (1985) have shown that serious discrepancies exist the probability equivalence and certainty equivalence approaches to utility measurement. These results have been replicated by Schkade and Johnson (1989). Since both these approaches satisfy the axioms of expected utility theory, they should yield the same results. Given Hershey and Schoemaker's finding, an appropriate question is which of the two approaches is the "correct" one. Unfortunately, this question has not been the subject of empirical investigation. Kneppreth, Gustafson, Leifer, et al., (1974) have also raised the concern that gamble methods may suffer from lack of acceptability by respondents, who may rebel at the idea of playing a "game." Moreover, if they continue, they may not concentrate on the task. Kneppreth, Gustafson, Leifer, et al., (1974) have noted that these methods should be used primarily in situations where the hypothetical situation posed is realistic to the subjects. This has implications for the selection of subjects for health status preference measurement studies, since this acceptability bias may be less if patients were selected as subjects since they may be able to better identify with the stimuli under evaluation.

Another problem with the probability equivalence approach especially is that different individuals may interpret or react differently to the same probability. It has been observed that people appear to make choices in wager conditions on the basis of perceived probabilities, rather
than the stated objective probabilities (Edwards, 1954; Slovic and Lichtenstein, 1968). Therefore, since preference functions are generated using the objective probabilities, bias may be introduced into the scaling method according to how the respondent perceives the probabilities. In an attempt to resolve this condition, Fishburn (1964) has suggested that a subjective function of objective probability should be determined for the respondents, and that this be used in the method when calculating preferences. However, this would involve considerably more time and effort, thus adding to an already lengthy and expensive scaling method.

The problem of “chaining” has also been noted in the case of the certainty equivalence approach. The chaining of measurements refers to the fact that later responses by any subject are based on previous ones (McCord and de Neufville, 1986). This may be a problem with the certainty equivalence approach because the certainty equivalent obtained in response to the first lottery is used to calculate the utilities of the remaining stimuli (as it is incorporated into the wager). Therefore, if any bias enters the assessment procedure at an early stage, it will remain in the process and taint subsequent measurements.

It is ironic that the greatest strength of gamble methods is also a drawback under certain conditions. It had been noted earlier that the gamble methods were the only scaling methods which allowed the incorporation of risk. However, subjects' attitudes toward risk may be a source of bias in the utilities measured using the gamble methods. This has also been referred to as a gambling effect bias. In this regard, subjects may be risk-seeking, risk-neutral, or risk-averse. On one hand, some respondents may be inveterate
gamblers (i.e., risk-seeking), who enjoy gambling and will tend to choose a wager over a sure thing even when the sure thing is more preferential. For these subjects, the thrill of the wager tends to increase the worth of the consequences of the gamble. On the other hand, there are other respondents who are extremely averse to risk and always choose the sure thing even when it is clearly less preferable. For these subjects, the fear of the wager tends to increase the worth of the sure thing. Kahneman and Tversky (1979) have referred to this latter effect as the "certainty effect." For both risk-seeking and risk-averse subjects, the standard gamble method may therefore give biased measures of utility. It has been shown that most respondents are risk-averse when faced with the prospect of a gain and risk-seeking when faced with the prospect of a loss. Applied to health status preference measurement, this finding suggests that people will usually prefer to remain in a less-than-perfect health state rather than risk ending up sicker or dead. In particular, a health state would have to be extremely undesirable before a person would seek to get out of it by undergoing some therapeutic intervention with even a moderate risk of an undesirable consequence. Therefore, preferences measured using the standard gamble method are usually higher than those measured using other scaling methods (Froberg and Kane, 1989b).

It should be noted that a difference between the preferences measured using the standard gamble and other scaling methods is to be expected, since the other methods do not incorporate the notion of risk. Therefore, the higher preferences obtained with the standard gamble could be considered to be the effect of the attitude of the respondent toward risk. In other words,
the difference between the standard gamble and other scaling methods could be considered to be the difference between utilities and values. This is, in fact, the view propounded by Torrance and his colleagues at McMaster University (for e.g., see Torrance, Zhang, Feeny, et al., 1992) as well as in the decision analysis literature. Further details on this view will be provided in the next chapter, when the distinction between value-based and utility-based preferences is reviewed.

However, the criticism of the gambling effect bias is yet important to note, because one alternative of the standard gamble is a sure thing while the other is a wager. Therefore, what is needed is a method which will incorporate respondents' attitudes toward risk while avoiding the inveterate gambler/certainty effect problem described earlier. McCord and de Neufville (1986) have developed "lottery equivalents" as such a scaling method. Further, Becker and Sarin (1987) have adopted lottery equivalents as the scaling method to be used in conjunction with their generalization of expected utility theory, called Lottery Dependent Expected Utility (LDEU) theory. LDEU has been empirically tested and found to have better predictive as well as descriptive validity as compared to expected utility theory (Daniels and Keller, 1990). The major difference between lottery equivalents and the standard gamble method is in the choice faced by the respondent. In the lottery equivalents method, the subject asked to make a choice between two wagers as opposed to being given a choice between a wager and a sure thing in the standard gamble. In this way, both options are probabilistic and hence incorporate the notion of risk. The method would also address the problem of inveterate gamblers, since both options are
risky. However, the method may not be acceptable to extremely risk-averse subjects, who may refuse to participate in the response task on account of their extreme risk aversion.

In spite of these drawbacks, gamble methods have been extensively used in measuring health status preferences. The major impetus behind this extensive use is that they are the only methods which permit the incorporation of risk -- and health outcomes cannot be known with certainty. As noted earlier, the McMaster group of researchers has pioneered the use of these methods in health-care public policy decision-making. McNeil, Weichselbaum, and Pauker (1978) were the first to apply these methods for the purpose of clinical decision-making. Gamble methods have been used to measure health status preferences for a variety of conditions, including chronic states, acute states, and health states worse than death. The intrarater reliability of these methods has been estimated at 0.77 (Torrance, 1976a), the one-week test-retest reliability at 0.80 (O'Connor, Boyd, and Till, 1985), and the one-year test-retest reliability at 0.53 (Torrance, 1976a). The low one-year test-retest reliability may either reflect true changes in preference over the year or the effect of a different interviewer in both the data collection efforts. It should be noted that all these applications have used the standard gamble method. There have been no published applications of the lottery equivalents method to health status preference measurement.
The Time Tradeoff Method

The time tradeoff method was developed by Torrance as a simple-to-administer alternative to the gamble methods specifically for the purpose of measuring preferences for health states (Torrance, Thomas and Sackett, 1972). The first two steps of this method are the same as those for the gamble methods -- ranking the stimuli and choosing a scale by assigning arbitrary numbers to the most and least preferred stimuli. The difference between the gamble methods and the time tradeoff method lies in the next step, when the subject is presented with a choice. Unlike the standard gamble, where the subject is asked to choose between a wager and a sure thing, the time tradeoff method asks the subject to choose between two sure things. The method usually asks the respondent how much time (i.e., years of life) he or she is willing to give up in order to be in a healthier state as compared to a less healthy one.

For instance, for a chronic health state which is considered better than death (according to the ranking provided by the respondent in the first step), the time tradeoff method would be operationalized by offering the subject two alternatives (i) health state $i$ for time $t$ (i.e., life expectancy of the individual with the chronic condition $i$ under evaluation) followed by death and (ii) perfect health for time $X<t$ followed by death. The time $X$ is varied in a ping pong manner (as described for the gamble methods earlier) until the respondent is indifferent between the two alternatives. At this point, the required preference score for the chronic condition $i$ under evaluation is given by $(X/t)$. 
The time tradeoff method can be altered to apply to other conditions, such as health states worse than death and temporary health states. A time tradeoff board can be used as a visual aid to help respondents make the judgment (see Furlong, Feeny, Torrance, et al., 1990). In terms of the five criteria identified earlier, the time tradeoff method: (a) depends on indifference judgments from respondents, (b) involves two or three stimuli at a time (i.e., stimulus under evaluation and the reference stimuli -- there is one reference stimulus for evaluating chronic states better than death and temporary health states, but two reference stimuli for evaluating chronic states worse than death), (c) works with either discrete or continuous stimuli, (d) does not incorporate the notion of risk (i.e., measures values, not utilities), and (e) provides responses on an interval-level scale (according to the same argument as for the gamble methods, after which the time tradeoff method is patterned). With respect to the measurement strategies with which it can be used, the time tradeoff method can be used with all the health status preference measurement strategies discussed earlier with the exception of all the conjoint methodology strategies and the analytic hierarchy process.

The time tradeoff method has been shown to provide reliable data. Intra-rater reliability coefficients which have been reported range between 0.77 and 0.88 (Torrance, 1976a, 1982b) and test-retest reliability has been reported by different investigators at several different intervals: (a) one-week -- 0.87 (O'Connor, Boyd, and Till, 1985) (b) four-week -- 0.81 (Churchill, Torrance, Taylor, et al., 1987), (c) six-week -- 0.63-0.80 (Churchill, Torrance, Taylor, et al., 1987), and (d) one-year -- 0.62 (Torrance, 1976a).
Once again, the one-year test-retest reliability coefficient is moderate, and this may be either because of real changes in preference over the course of the year or because of the effect of a different interviewer in the retest as compared to the test. Churchill, Torrance, Taylor, et al., (1987) have provided some evidence of the validity of the time tradeoff method using the method of known group differences. They asked end-stage renal patients to evaluate their own health using the time tradeoff method, hypothesizing a particular order of scores based on the severity of illness of the patients. These hypotheses were borne out, suggesting that the time tradeoff method can make clinically meaningful distinctions between health states (Churchill, Torrance, Taylor, et al., 1987).

Some of the drawbacks of the time tradeoff method are similar to those of the gamble methods, i.e., the need to have well-trained interviewers, lengthy interviews (and the consequent chance of respondent fatigue), and resource-intensive. Another potential limitation is that it does not directly account for nonlinearities in the time-utility function in which a person places a much higher premium on immediate rather than long-term survival. Along the same lines, it does not account for Sutherland, Llewellyn-Thomas, Boyd, et al.'s (1982) observation that subjects are willing to spend only some "maximal endurable time" in a given health state.

The Contingent Valuation Method

The contingent valuation method was developed in the field of welfare economics as a method to place a measure of numerical worth on intangible items or goods which are not traded in regular markets. Within
this method, there are two general approaches for eliciting responses from subjects -- the willingness to pay approach and the willingness to accept approach. In the first approach, subjects are usually asked to estimate how much (in terms of some monetary amount) they would be willing to pay in order to receive a given amount of the item in question. In the second approach, subjects are usually asked to estimate how much (in terms of some monetary amount) they would be willing to accept in return for giving up a certain amount of the item in question.

Although the two approaches should yield the same results in the absence of income effects, the studies that have compared hypothetical willingness to pay and hypothetical willingness to accept estimates have found large differences in the estimates which cannot be explained by income effects (Bishop and Heberlein, 1979; Brookshire, Randall, and Stoll, 1980; Hammack and Brown, 1974; Knetsch and Sinden, 1984; Rowe, d’Arge, and Brookshire, 1980). This has led to a recommendation from most experienced researchers using the contingent valuation method to elicit willingness to pay measures as opposed to willingness to accept measures (see Cummings, Brookshire, and Schulze, 1986; Mitchell and Carson, 1989). Indeed, most of the studies using the contingent valuation method in healthcare have elicited willingness to pay measures from subjects.

The contingent valuation method is very easy to implement. Subjects are exposed to the stimuli (preferably in random order) and a willingness to pay/accept measure is obtained. The response format can either be of an open-ended or discrete valuation type. In open-ended valuation questions, the respondents are asked to state their maximum willingness to pay/accept
for the stimulus, while in discrete valuation questions there are a limited amount of response alternatives to choose between. In a number of studies where open-ended and discrete valuation questions have been compared, it has been found that willingness to pay has been higher in discrete valuation questions (Kristrom, 1989, 1990; Sellar, Chavas, and Stoll, 1985).

In terms of the five criteria identified earlier, the contingent valuation method: (a) depends on quantitative judgments -- specifically pricing judgments -- from respondents, (b) involves only one stimuli at a time (i.e., stimulus under evaluation), (c) works with either discrete or continuous stimuli, (d) does not incorporate the notion of risk (i.e., measures values, not utilities), and (e) provides responses on a ratio-level scale. With respect to the measurement strategies with which it can be used, the contingent valuation method can be used with all the health status preference measurement strategies discussed earlier with the exception of the rank order judgment strategies, the discrete choice conjoint strategies, and the analytic hierarchy process.

However, contingent valuation studies may be prone to several biases, the most significant of which are incentives to misrepresent responses and implied value cues. In terms of incentives to misrepresent responses, respondents may alter their responses if they feel that it is in their best interest to give a lower or higher valuation than the true value. Although this bias may be present in any direct scaling method, it is particularly relevant for the contingent valuation method because the response is usually given in terms of some monetary amount. Implied value cues exist when some information of the contingent valuation instrument implies a certain
value for the stimulus under evaluation. For instance, starting-point bias is present when the respondent's valuation is affected by some potential willingness to pay/accept amount is given in the study. Range bias may be present when there is a range of potential willingness to pay/accept amounts given in the instrument which affects the responses.

The contingent valuation method has not been frequently applied to the measurement of health status preferences, perhaps on account of these biases. There were a few applications in the early 1980s (Mushkin and Dunlop, 1979; Thompson, Read, and Liang, 1984), which were not very supportive of the method. One-year test-retest reliability coefficients were very low (0.25) participation rates were very low, and subjects were reported to have been hostile to the question of estimating their willingness to pay for health benefits (Thompson, Read, and Liang, 1984). However, there seems to be an resurgence in interest in contingent valuation as a measure of benefit, to be used as the denominator in cost-benefit pharmacoeconomic analyses. This interest appears to be spearheaded by the efforts of Magnus Johannesson (see Johannesson, 1992; Johannesson, Johansson, Kristrom, et al., 1993; Johannesson, Johansson, and Jonsson, 1992; Johannesson and Jonsson, 1991; Johannesson, Jonsson, and Borgquist, 1991), although some researchers at McMaster University have also been active in this area (see Gafni, 1991; O’Brien, Novosel, Torrance, et al., 1995; O’Brien and Viramontes, 1994).

Apart from these direct scaling methods, there are some other ones which have been used occasionally to measure health status preferences.
These include the methods of equivalence, fractionation, multiple judgment, and ratio partition. Of these, only the equivalence method will be briefly discussed since only this method has been applied to the measurement of health status preferences. The method of *equivalence* (also called the person tradeoff method) has been used in a few studies to measure health status preferences (Berg, 1973; Patrick, Bush, and Chen, 1973b). This method is an adaptation of the method of adjustment or equivalent stimuli developed in the field of psychophysics on the basis of the research conducted by Plateau (Plateau, 1872). Although this method has been operationalized in different ways, the common underlying task for the respondent is to decide how many people in a particular health state are equivalent to a specified number of people in health state B. Each judgment is then treated as a ratio of the standard health state. These ratios are then averaged to obtain a preference scale for each health state under evaluation.

In terms of the five criteria identified earlier, the equivalence method: (a) depends on indifference judgments from respondents, (b) involves two stimuli at a time (i.e., the stimulus under evaluation and the standard stimulus), (c) works with either discrete or continuous stimuli, (d) does not incorporate the notion of risk (i.e., measures values, not utilities), and (e) provides responses on a ratio-level scale. With respect to the measurement strategies with which it can be used, the equivalence method is probably best used with only the holistic preference measurement strategies.

However, the reported inter-rater reliability of this method was only 0.60 (Patrick, Bush, and Chen, 1973b). Further, Patrick, Bush, and Chen (1973b) have noted that the method may be too complex for use outside a
laboratory setting. Finally, it should be noted that some respondents reported they were confused and even offended by the unrealistic assumptions and emotive nature of the task (Patrick, Bush, and Chen, 1973b).

**Indirect Scaling Approaches**

Indirect scaling approaches can be considered to be methods in which there is one extra step between the raw data and the final scale as compared to direct scaling approaches. The raw data are only ordinal in level of measurement, i.e., show order or direction of preference as opposed to the magnitude of preference. In order to convert the raw ordinal-level data into the final cardinal-level scale, it is necessary to make certain assumptions. Depending on the particular assumptions made, different scaling methods are possible under this approach.

**The Paired Comparisons Method**

One of the most popular indirect scaling methods is the paired comparisons method. As noted earlier, it is the most commonly used method for obtaining scales in accordance with Thurstone's (1927a) Law of Comparative Judgment. It is usually implemented in three steps. At first, it is necessary to construct pairs of stimuli to be evaluated. Although it is not necessary to construct all possible pairs of stimuli, enough pairs must be constructed so as to ensure a complete ordering of stimuli. In the second step, the subject is presented with these pairs and asked to make a judgment,
which may be of preference, intensity, or any other desired attribute of the stimuli. The final step involves converting the raw data into the final scale. From the judgment data collected from the second step, one may obtain the proportion of trials on which each stimulus was judged more preferred, intense, etc. to each of the others. From these proportions, one may derive a subjective scale by assuming one of the five cases of Thurstone’s (1927a) Law of comparative judgment. Since the details of this Law were described earlier, they will not be repeated here.

It is important to note the range of the subjective scale derived as a result of this method will only extend over the range of the stimuli evaluated. Therefore, it is desirable to include a wide range of stimuli. However, there must be some degree of confusion between adjacent stimuli in order to apply Thurstone’s (1927a) Law of Comparative Judgment. This in turn makes it very difficult to design an investigation spanning a very broad range of stimuli without inordinately increasing the number of pairs to be evaluated -- which would run the risk of respondent fatigue. Another point to note is that the subject must be transitive in his or her preferences (i.e., if he or she prefers stimulus A to stimulus B and stimulus B to stimulus C, he or she should prefer stimulus A to stimulus C) in order to apply this method.

In terms of the five criteria identified earlier, the method of paired comparisons: (a) depends on preference judgments from respondents, (b) involves two stimuli at a time, (c) works with either discrete or continuous stimuli, (d) does not incorporate the notion of risk (i.e., measures values, not utilities), and (e) provides observed responses on an ordinal-level scale which are subsequently transformed into an interval-level final scale.
However, with respect to the measurement strategies with which it can be used, the method of paired comparisons is probably best used with only the rank order judgment conjoint strategies.

The method of paired comparisons was used in the earliest study on health status preference measurement in the modern period of health status development i.e., from the early 1970s to the present (Fanshel and Bush, 1970). However, these researchers used other scaling methods in their subsequent investigations on health status preference measurement (see Kaplan, Bush, and Berry, 1976, 1978, 1979; Patrick, Bush, and Chen, 1973a, 1973b). A possible reason for this could be that the method requires a large number of comparisons to be made. In this regard, it is necessary to construct \( n(n - 1)/2 \) pairs for complete judgment designs, where \( n \) is the number of alternatives to be evaluated. If each health state was described in terms of only three dimensions, each with four levels, the total number of alternatives (assuming that there were no "nonsense" alternatives) would be 12. Therefore, the total number of pairs to be judged by each subject would be 12(11)/2 or 66, which is prohibitively large and may cause respondent fatigue and even threaten the reliability of the data collected due to this fatigue. Therefore, the use of the paired comparisons method may be unwieldy when there are several stimuli to be evaluated.

Although there are other scaling methods under the indirect scaling approach category, only one more will be discussed here -- the *pick-one method*. There have been no published applications of this method to the measurement of health status preferences. The reason for discussing this
particular method is that it is one of the scaling methods which will be used in this dissertation.

The Pick-one Method

As the name implies, the pick-one method involves choosing one out of several alternatives from a choice set. The method may be implemented in four steps. The first step involves constructing the stimuli, the second the construction of choice sets. Since the details of these steps have already been discussed in the section on discrete choice conjoint strategies, they will not be repeated here. In the third step, the subject is shown the required number of choice sets and makes his or her selection. The final step involves converting the raw ordinal-level data into an interval-level scale. In order to do this it is necessary to assume that the respondent’s underlying decision-making process is best represented by some probabilistic choice model, such as the MNL choice model. Again, since the details of this step were provided in the discussion of discrete choice conjoint strategies, they will not be repeated here.

In terms of the five criteria identified earlier, the pick-one method: (a) depends on choices to be made by respondents, (b) involves multiple stimuli at a time -- the limits to the number of stimuli depend on the specifics of the experimental design, (c) works with either discrete or continuous stimuli, (d) does not incorporate the notion of risk (i.e., measures values, not utilities), and (e) provides observed responses on an ordinal-level scale which are subsequently transformed into an interval-level final scale. However, with respect to the measurement strategies with which it can be used, the pick-
one method was designed for use with only the discrete choice conjoint strategies.

This section has reviewed several different scaling methods for measuring health status preferences. It is important to note that several other methods have been proposed for assessing -- as opposed to measuring -- health status preferences. These assessment methods will only be mentioned here, simply in order to acknowledge them as alternative methods to use when resource constraints do not permit the direct or indirect measurement methods to be used. Assessment methods include assigning arbitrary scores to health states, using a panel of experts to assign scores to different health states, obtaining scores from the published literature, and deriving scores from administrative or social decisions.

Regarding the first two of these methods, the scores so obtained would be confined to only the judges who generated them and to the particular study for which they were generated. This specificity makes comparisons between different studies difficult or even impossible. Regarding obtaining scores from the published literature, the applicability of the scores to the purpose of the current investigation needs to be thoroughly determined. In this regard, it is necessary to determine the characteristics of the judges, instruments, and treatments used in the published studies. Moreover, most published estimates of preference scores do not specify the length of time over which they were calculated. This is important to know because the preference for health states is dependent on how long the health state in question is expected to last (as discussed earlier). Obtaining such information may be difficult since many details are not published. Finally,
using the "revealed preference" method or deriving preference scores from administrative or social decisions runs the risk of confounding the politics of decision-making with the true preferences underlying these decisions. Further, the human capital approach which has been used frequently to estimate the benefit of many therapeutic alternatives have been criticized for discriminating against individuals who do not trade a product or service in the market -- homemakers, retirees, etc. (Avorn, 1984). In light of these observations, it is always desirable to measure health status preferences by primary data collection efforts for any particular investigation. This is the approach which will be adopted in this dissertation.

Given this variety of scaling methods, a natural question would be which is the "best" one. Unfortunately, there is no clear answer to this question. While deciding upon a particular scaling method, the following points need to be considered:

1. the decision about the scaling approach can only be made after a decision has been reached on the particular measurement strategy to be used for the study;
2. if it is considered necessary to incorporate risk as part of the scaling method, gamble methods need to be used;
3. all scaling methods can be made to incorporate the effect of time by the use of appropriate instructions to respondents on the nature of the scaling task;
4. if the interest of the study is on the choice made by respondents or the process used by respondents in making a choice, the pick-one scaling method needs to be used;

5. although the methods have been shown to be reliable, the precision of these methods is still a question. For instance, the standard errors of the mean for the rating scale, probability equivalence standard gamble, and time tradeoff methods are 0.09-0.15, 0.13, and 0.13 respectively.\(^{192,195}\)

Consider the time tradeoff method, a 95\% confidence interval around a point estimate of 0.60 would be 0.60 ±/− 0.26, which is considerably large. This would suggest that individual measures may not be very precise, and that it may be preferable to work at a group level rather than the individual level;

6. the results of the different scaling methods do not necessarily converge. Froberg and Kane (1989b) have provided a review of the studies which have either examined the functional relationships between the methods or compared the mean scale scores derived from each method. Due to this lack of convergence, the selection of an appropriate scaling method depends on the purpose of the study and the way in which the results will be used;

7. from a data collection point of view, the rating scale method is the most flexible in that it can be used with a variety of data collection modes, including the mail and personal interview. Of the other scaling methods, magnitude estimation, contingent valuation, paired comparisons, and the pick-one -- especially the paired comparisons and pick-one methods -- may be amenable to data collection through the mail, albeit with a low
response rate. However, the gamble and time tradeoff methods require personal interviews;

8. from a feasibility point of view, the gamble and time tradeoff methods are the most expensive and time-consuming and required visual aids to help the respondent understand the task required of him or her. Rating scale and magnitude estimation methods are the least expensive and easiest to understand. Equivalence and contingent valuation methods have confused and even offended respondents -- therefore these methods need to be implemented with great caution. Paired comparison methods run the risk of burdening the respondent with too many pairs of stimuli to evaluate. Finally the pick-one method is easy for respondents to understand but requires expertise in designing the choice sets.

**Evaluation of Existing Research in Health Status Preference Assessment**

Against this backdrop of the methodology for measuring health status preferences, the existing research in this area can be evaluated. In this regard, the work of three groups of researchers needs to be evaluated -- the San Diego, McMaster, and EuroQol groups of researchers. While it is acknowledged that there have been other researchers who have attempted to measure preferences for health states, only these three groups have worked with a comprehensive health status classification system at the general (as opposed to disease-specific) level. Therefore, the work of these three groups of researchers comes closest in spirit to the emphasis of this dissertation. This section will only focus on the methodological aspects of the preference
measurement methods used by these groups of researchers (specifically the measurement strategy and scaling approach), since other relevant criticisms of the work of these researchers has already been given in Chapter I.

To reiterate, patients’ preferences for health states are affected by a multitude of factors, including the relative importance of health status dimensions, interactions between these dimensions, patient attitudes toward risk taking, the point in time patients are expected to experience the health state, the length of time patients are expected to stay in the particular health state, and the sequence of health states preceding and following the health state under evaluation. In order to provide true meaning to health status scores, the role of each of these factors and their effect on patients’ preferences for health states needs to be assessed.

The San Diego Group

As described earlier, the San Diego group of researchers were responsible for the development of the Quality of Well Being (QWB) Scale. Although this instrument does provide information about the relative importance of health status dimensions, it does not allow for interactions between these dimensions. Therefore, the information provided about the relative importance of dimensions may be misleading, since interactions could change the preferences for health status dimensions. Indeed, the QWB Scale was purposely constructed so as to have independence between the dimensions. It is possible that this independence might have been achieved at the cost of loss of accurate representation of health status preferences.
Further, in an attempt to verify the interval-level properties of their scale and provide support for their additive model form, Kaplan, Bush, and Berry (1979) conducted a functional measurement analysis on a 2x2 factorial design extracted from their data. Since the ratings yielded parallel curves when plotted as a function of level of social activity, with a separate curve for each of two symptom/problem complex levels, they argued that their scores were on an interval scale (Kaplan, Bush, and Berry, 1979). Although Kaplan, Bush, and Berry (1979) did not specify the model form they were testing [as Veit and Ware (1982) have pointed out], if one reads their previous studies it can be assumed that they were testing an additive model form. In that case, the parallel curves of the functional measurement would seem to support an additive model form, i.e., no interactions between health status dimensions.

However, it is necessary to place enough ordinal constraints in the design in order to provide an adequate test of parallelism. As Veit and Ware (1982) have noted, in the 2x2 design of Kaplan, Bush, and Berry (1979) there were not enough restrictions on the order of the data points to constrain an interval scale. Indeed, many monotonic transformations of scale scores would reproduce the order of the four data points. Therefore, it can be concluded that the test that Kaplan et al. conducted was not a true test of the model form and level of measurement of the QWB Scale. This would support the contention that not including interactions between the health status dimensions represented by the QWB might be misleading in terms of the preferences obtained for different health states.
Moreover, the scaling method used by this group of researchers for the final QWB Scale was a rating scale method. As noted earlier, rating scale methods do not incorporate the notion of risk into the evaluation of health states. Given the importance of incorporating risk into such evaluations (as discussed earlier), this represents another drawback. However, it should be noted that gamble methods of scaling cannot be currently used as part of any conjoint methodology strategy (the strategy used by these researchers was a rating scale judgment conjoint strategy). Therefore, although the non-incorporation of risk might be a drawback there is little that these researchers could have done, given that they had decided to use a rating scale judgment conjoint strategy.

The handling of the issue of time by this group of researchers was much more satisfactory. Regarding the point in time respondents were expected to experience the health state under evaluation, these researchers specifically tested for this effect by including five levels of age which "... correspond to the natural transitions from infancy to old age (Patrick, Bush, and Chen, 1973a p.9)." It was found that there were no significant differences between respondents on the basis of these age groups (Kaplan, Bush, and Berry, 1976). Regarding the expected duration of stay in the health state under evaluation, this was specified in the instructions to the respondents. Finally, the issue of health states following the one under evaluation has been one of the strengths of the General Health Policy Model, because the QWB Scale is the point in time or static component of the model while the prognosis is the dynamic aspect. By summing over all expected health states over the lifetime of an individual, the well-years so computed
will account for the health states expected to precede and follow the health state under evaluation.

In summary, the QWB Scale itself may be based on an incorrect model form which in turn may lead to biased estimates of health status preferences. This problem cannot be rectified by the strengths of the General Health Policy Model, and represents a serious drawback of the QWB Scale.

The McMaster Group

As described earlier, the McMaster group of researchers were responsible for the development of the Health Utility Indices. Although three different versions of these indices have been developed (Marks I, II, and III), there has been no published literature on the latest version (i.e., Mark III). Therefore, this evaluation will focus on the first two versions.

Both versions used the conditional utility function-based procedure to measure preferences for health states. Both versions provide information about the relative importance of health state dimensions. More importantly, both versions explicitly allow for some interactions between health status dimensions, since the underlying model form is a multiplicative one. Further, the existence of the multiplicative model form was empirically tested and found to hold, which provides more faith in the underlying model form. Although a multiplicative model form does not allow all interactions between health status dimensions, Keeney and Raiffa (1976) have noted that in cases of more than four dimensions, these higher-order interactions are not likely to be significant. Further, to scale the parameters of the multilinear model forms which would allow the existence of such
interactions would require much more data to be collected, and may even be prohibitive in terms of resources.

The first version of the Health Utility Index did not incorporate the notion of risk into the scaling method -- the scaling methods used in this version were the rating scale and time tradeoff methods. However, the second version used the standard probability equivalence gamble method to obtain scale scores, which explicitly incorporates subjects' attitudes toward risk in the evaluation. This is a major strength of the Mark II version of the Health Utility Index.

Regarding the point in time respondents were expected to experience the health state under evaluation and the expected length of duration of the health state under evaluation, both these factors were specified in both versions. Since the expected duration of stay in the health state under evaluation was a lifetime in both cases, the issue of expected future health states is not relevant in these cases. It should, however, be noted that the notion of staying in a particular health state for a lifetime is unrealistic.

The only major drawback with the measurement methodology of the McMaster group of researchers is an inherent characteristic of the measurement strategy that they adopted-- the conditional utility function-based procedure does not have an error theory. As noted earlier, the lack of error theory is problematic in that no measurement error is assumed in the estimation of model parameters. In this regard, all that can be obtained is a point estimate of the model parameters. There is no way to construct a confidence interval around each point estimate. This means that there is no way of knowing how much faith one can have in the point estimate.
parameters of the model. This may be a serious drawback, given the potential uses of health status preference measures for public policy and clinical decision making.

In summary, the McMaster Health Utility Indices (especially the Mark II version) are methodologically sound. The Mark II version explicitly incorporates risk into the evaluation, and this factor coupled with the permissibility of some interactions between health status dimensions makes it one of the premier health status measurement instruments available at present. The major drawback is the lack of error theory in the measurement strategy adopted.

The EuroQol Group

Of the three groups of researchers whose work is reviewed here, the EuroQol group is the newest entrant into the field of health status preference measurement. It should, however, be noted that some of the individual members of the EuroQol group have had distinguished careers in the field of health status preference measurement as individual researchers; however, as a group, the EuroQol has been in existence for less than a decade. Given the short duration of its existence, it is understandable that there has not been much published literature on the specifics of the methodology adopted by these researchers to measure health status preferences -- especially since the health status descriptive classification used by these researchers has already been revised once.

The few articles which were published on the first version of the EuroQol used a holistic preference measurement strategy with a rating scale
scaling method. Sixteen health states were selected which were considered to be commonly occurring in clinical practice. Respondents were asked to rate these health states on a 100-point rating scale with 0=worst imaginable health state and 100=best imaginable health state.

Health status preference measurement using the revised version of the EuroQol proceeded in essentially the same manner as for the first version. However, much more experimentation with different preference measurement methodologies seems to characterize the current focus of investigation by the EuroQol group of researchers (Busschbach, Hessing, and de Charro, 1993; Sintonen, 1993). For instance, the group has reported comparisons of preference scores obtained using different scaling methods, including the time tradeoff, standard gamble, equivalence, and magnitude estimation scaling methods apart from rating scale methods. This multitude of scaling methods is noteworthy in that it potentially allows greater understanding of respondents' subjective preferences, since each scaling method has its own strengths.

However, less thought seems to have been given to experimenting with different measurement strategies, since all these experiments were conducted with the same "common core" of sixteen health states which -- like those selected for the first version of the EuroQol instrument -- were based on a judgment sample of the theoretical universe of 243 possible health states (actually 245 health states, counting the states of death and unconsciousness). Still, it should be noted that there have been some attempts to determine the model form of the preference function, using regression-based techniques (as explained in Chapter 1). This essentially falls
within the purview of social/clinical judgment analysis, and consequently share the drawbacks of this particular procedure (as detailed earlier in this chapter).

Since experimentation with health status measurement with the EuroQol group of researchers is still in progress, it seems premature to evaluate the research of this group. Still, a few points need to be noted. First, in terms of measurement strategy, although the move from a holistic strategy to social/clinical judgment analysis is a step in the right direction, cause-and-effect conclusions cannot be drawn in the absence of experimental designs. In this regard, the use of the functional measurement approach may warrant investigation by the EuroQol group. Moreover, the issue of multicollinearity is a serious one in health status preference measurement, and the use of social/clinical judgment analysis might lead to biased parameter estimates (i.e., attribute or dimension preference scores) in such situations.

Second, in terms of scaling methods, the effort to investigate with several different methods is creditable. Although no clear "best" method seems to have been decided, the exercise itself will provide valuable insight into respondents' preference functions for health status.

Finally, one last shortcoming of all three groups of health status preference measurement researchers needs to be reiterated -- the sole use of judgmental responses as opposed to choices to infer preference functions. As noted earlier, preferences may be inferred via judgmental as well as choice data, and, given the likelihood of response mode effects, it is possible that the two basic types of response modes (i.e., judgments and choices) may
lead to different health status preference functions. In this regard, all three groups of researchers seem to have operated upon the assumption of procedural invariance -- which has been shown to not hold in practice.

Having reviewed the methodology for measuring health status preferences as well as the research conducted by the three most prominent and influential groups of health status preference measurement researchers, the next question that needs to be addressed is: "How do the different preference measurement methodologies compare with one another?" This is an important issue, especially as it pertains to this dissertation, since it gets at the essence of convergent validation -- if convergence is to be expected among the different types of measurement methodologies investigated in this study, previous studies on this issue need to be reviewed so that theoretical and empirical support for the hypotheses presented in Chapter II can be provided. Therefore, it is necessary to review the studies which have investigated the comparative relationship between the various preference measurement methodologies investigated in this dissertation. This is the subject of the next chapter.
HEALTH STATUS PREFERENCE MEASUREMENT:
A COMPARISON OF DISCRETE CHOICE CONJOINT
AND CONDITIONAL UTILITY MODELING

Volume II

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By
Zafar Mohomed Anwar Hakim, B.Pharm., M.S.

* * * * *

The Ohio State University
1995

Dissertation Committee: Approved by
Dev S. Pathak 
Suzan Kucukarslan 
Jon C. Schommer

Adviser
College of Pharmacy
CHAPTER IV

COMPARISONS OF PREFERENCE MEASUREMENT METHODOLOGIES IN THE PUBLISHED LITERATURE

Chapter Overview

The purpose of this chapter is to review the published literature on topics relevant to the research questions investigated in this dissertation. The chapter begins with an introduction to the issues involved in comparing alternative preference measurement methodologies. After delineating the specific issues explored in this study, the chapter proceeds to review three types of studies in the literature -- studies comparing top-down versus bottom-up measurement strategies, studies comparing value-based versus utility-based preference scores, and studies investigating the nature of response mode effects (specifically judgments versus choices) on preference measurement methodology. The next section deals with the important issue of respondent evaluations of preference measurement methodologies, and reviews the few studies which have addressed this issue. Finally, the last section summarizes the main findings of the preceding review with the intent of providing the link between the published literature and the specific questions which were investigated in this dissertation.
Issues in Comparing Preference Measurement Methodologies

Research comparing different preference measurement methodologies is varied and multidisciplinary. In recognition of the sheer volume of published literature on such comparative studies, it was decided to limit the scope of this literature review to only those studies which compared the methodologies under investigation in this dissertation.

The issue of comparing alternative preference measurement methodologies is one which has been recognized in the literature for over a quarter of a century. Since the seminal work of von Neumann and Morgenstern (1944, 1947), which provided an axiomatic foundation for the measurement of cardinal utility, there have been several preference measurement methodologies which have been developed (see Anderson, 1970, 1971, 1980, 1981; Keeney and Raiffa, 1976; Luce and Tukey, 1964; Louviere and Woodworth, 1983). Since each of these different measurement methodologies purported to same the same construct (i.e., preferences -- either as utilities or values), the natural question that arises is, “How do these different methodologies compare with one another?”

The evaluation of this question was (and still is) quite difficult, since preferences are an abstract construct which cannot be observed readily. Therefore, there is no way of knowing for certain whether the construct being measured by these methodologies was indeed preferences. To further complicate matters, even if these different methodologies do measure preferences, there is no way of knowing which is the “correct” way, if the results obtained by these different methodologies are not in agreement with
each other. This is because there is no external standard or criterion to use as a reference point for comparison, i.e., there is no gold standard.

In the absence of such a gold standard, researchers have proposed and utilized a variety of bases for comparison of such methodologies. In this regard, conformance with axioms of the underlying theory, linear convergence among measures, agreement in terms of preferential ordering of alternatives measured using different methodologies, verification of functional relationships between the different methodologies, comparison of parameter estimates obtained using the different methodologies, and relative success in predicting preferences in holdout validation tasks have been used for the purpose of comparing the alternative preference measurement methodologies. The vast majority of studies in the published literature have only considered a subset of these bases for comparison.

Before reviewing the relevant studies in this section, a brief explanation about the rationale behind each of these bases for comparison is necessary. Regarding conformance with axioms of the underlying theory, some researchers have held that if respondents behave in accordance with the axioms underlying the measurement theory, the preference scores obtained via the measurement methodology will be valid; therefore, if there are discrepancies between preference scores obtained via different measurement methodologies, the scores obtained in conformance with axioms would be the valid ones.

The principle of linear convergence among measures is based on the tenets of the multitrait-multimethod matrix introduced by Campbell and Fiske (1959), and holds that if the different measures of preferences are valid
(i.e., measuring the same construct), they should be unique up to a positive linear transformation and have high linear correlations with one another.

Regarding agreement in terms of preferential ordering of alternatives measured using different preference measurement methodologies [also called ordinal convergence by Fischer (1977)], this logical basis for comparison simply means that different measurement methodologies may give different preference scores to different alternatives, but as long as they do not change the preferential ordering of alternatives they are all valid. It should be noted that this is one of the weaker bases of comparison, since it implicitly assumes procedural invariance. The issue of procedural invariance and the consequences of its breakdown in practice have already been discussed and will not be repeated here.

One of the more stringent bases for comparison is the verification of functional relationships between different preference measurement methodologies. This is because these functional relationships are based on a body of theoretical literature and empirical testing of theoretical relationships is the essence of the scientific method (Rudner, 1966; Hunt, 1990). By empirically verifying the existence of functional relationships between different preference measures, it is possible to infer with some degree of confidence that not only is the theoretical relationship between the two measures verified by empirical evidence, but also each of the measures is valid.

The comparison of parameter estimates of the levels of the different attributes or dimensions is not a very stringent test, simply because it does not possess the property of falsifiability. In other words, if the parameter
estimates obtained from the different preference measurement methodologies do not converge, no meaningful conclusion can be reached. This is because -- as posited by change of process theory (Mellers, Ordonez, and Birnbaum, 1992), which will be presented later in this chapter -- this comparison totally ignores the role of the combination function or model form in determining preferences. Thus, it is possible that different methodologies may yield nonconvergent parameter estimates but highly convergent holistic alternative preference scores because each methodology may induce a different combination function -- this is nothing but the response mode effect introduced in Chapter II.

Finally, there is the issue of comparing the relative success of different preference measurement methodologies in predicting preferences in holdout validation tasks (also called cross-validation). This is another strong comparative test, mainly for two purposes. First, there is an upward bias in any multiple regression technique which uses the same data to estimate parameters and then to measure its predictive ability. This is known as capitalization on the chance characteristics of the estimation sample. Cross-validation using holdout validation tasks usually causes the measure of predictive validity to "shrink" on account of correlated measurement errors (Horst, 1966). Second, unless one multiple regression model contains variables which are entirely a subset of another (i.e., nested models), there is no straightforward technique for comparing their relative abilities to explain variations in the dependent variable. Since different methods often choose sets of attributes or dimensions which do not meet this restriction, comparing them in terms of multiple r's is impractical. However, cross-
validation essentially involves substituting one large independent variable, usually a weighted sum of the holdout sample's independent variables, and the estimation sample's parameters to obtain a simple $r$ instead of a multiple $r$. Thus, the resulting correlation matrix can be readily analyzed using a simple t-test for correlated correlation coefficients (Alpert, 1971). Further, the purpose of most preference measurement exercises is to use the preference model obtained from the study for the purpose of predicting respondents' preferences for other stimuli (i.e., which were not a part of the study). In this regard, comparison of predictive validity gets at the true purpose of most preference measurement exercises. It is important to note that the validation set used to predict preferences must be separate from that used to estimate the model parameters -- this is necessary to avoid capitalization on the chance characteristics of the sample, which might make the preference measurement methodology appear overly successful in predicting preferences when in truth this is just an artifact of the particular sample used in the estimation.

**Issues Relevant to the Comparisons of Interest in this Study**

Against this backdrop, the different studies which have compared the various preference measurement methodologies can be compared. However, as noted earlier, discrete choice conjoint methodology is a relatively new methodology. Therefore, there have not been many published empirical studies comparing this methodology to other preference measurement methodologies (the notable exceptions are the ones comparing it to rating
scale judgment conjoint strategies, which were reviewed in Chapter III). In order to facilitate the comparison among the three preference measurement methodologies investigated in this study, it is instructive to view each methodology in terms of its component measurement strategies and scaling approaches. The conditional utility function-based strategy with a rating scale scaling method (CUF-RS) is a “bottom-up” measurement strategy with a rating scale scaling method, the conditional utility function-based strategy with a standard gamble scaling method (CUF-SG) is a “bottom-up” measurement strategy with a standard gamble scaling method while discrete choice conjoint methodology (DCCM) is a “top-down” measurement strategy with a pick-one scaling method. It should be noted that it was decided to adopt the broader “top-down” versus “bottom-up” distinction as opposed to the specific types of measurement strategies (i.e., CUF versus DCCM) because of the paucity of published studies directly comparing the two specific measurement strategies.

Further, it is also instructive to differentiate between the components of the different scaling methods in terms of response modes and incorporation of attitudes toward risk. In this regard, the rating scale scaling method uses judgmental ratings as the response mode and measures preferences under certainty (i.e., measures values, not utilities). The standard gamble scaling method uses indifference judgments as the response mode and measures preferences under risk (i.e., measures utilities, not values). Finally, the pick-one scaling method uses choices as the response mode and measures preferences under certainty (i.e., measures values, not utilities). These distinctions are presented in Table 1.
Table 1: Distinctions Between the Preference Measurement Methodologies Investigated in this Study

<table>
<thead>
<tr>
<th>Measurement Methodology</th>
<th>Measurement Strategy</th>
<th>Response Mode</th>
<th>Value/Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUF-RS</td>
<td>Bottom-up</td>
<td>Judgmental Ratings</td>
<td>Value</td>
</tr>
<tr>
<td>CUF-SG</td>
<td>Bottom-up</td>
<td>Indifference Judgments</td>
<td>Utility</td>
</tr>
<tr>
<td>DCCM</td>
<td>Top-down</td>
<td>Selection Choices</td>
<td>Value</td>
</tr>
</tbody>
</table>

Legend:  
CUF-RS = conditional utility function-based strategy with rating scale scaling method  
CUF-SG = conditional utility function-based strategy with standard gamble scaling method  
DCCM = discrete choice conjoint methodology, i.e., conjoint methodology strategy with pick-one scaling method
Viewed in this manner, it can be seen that there are three major types of comparative studies which are of interest to this dissertation: (a) studies comparing top-down and bottom-up preference measurement strategies, (b) studies comparing value-based preference scores and utility-based preference scores, and (c) studies comparing judgmental and choice response modes. Each of these three major types of studies will be reviewed in this section, with special emphasis placed on the studies which used any of the methodologies investigated in this dissertation. Finally, on the basis of these reviews, the hypothesized relationships between the preference measurement methodologies investigated in this study can be explained and understood.

**Top-down Versus Bottom-up Preference Measurement Strategies**

The issue of comparing preference measurement strategies has been a contentious one in the psychometric literature ever since the 1950s, when the first studies on bootstrapping (or statistically inferred measurement strategies for measuring preferences) were published (for e.g., see Hoffman, 1960; Meehl, 1954). These first studies were concerned with the comparison of statistically inferred (usually social/clinical judgment analysis) with holistic preference measurement strategies. It was found that statistically inferred measurement strategies were usually more effective than holistic judgments of experts in several instances requiring expert judgment, e.g., medical diagnosis. On the basis of these early studies, the social/clinical judgment analysis became a popular technique for providing what Hoffman
(1960) termed a "paramorphic" (i.e., algebraic representation which was not necessarily descriptively accurate) representation of human judgment.

In the mid-1960s the expectancy-value models of Rosenberg (1956) and Fishbein (1963), which were popular in the field of social psychology, were introduced into the field of behavioral decision making as self-explicated or subjective evaluation models (for e.g., see Hoepfl and Huber, 1970, Huber, Sahney, and Ford, 1969). These expectancy-value models can be viewed as the forerunners -- at least in terms of measurement strategy -- of the additive rating scale procedure [which was popularized by Edwards (1971, 1977) in the 1970s] in the behavioral decision making literature. With the introduction of these models to the behavioral decision making literature, the debate between top-down and bottom-up strategies for preference measurement could have said to have begun, although this debate was phrased more in terms of "subjective" versus "objective" weighting of attributes or dimensions in decision making. In this regard, the subjective weights were the self-explicated weights or preference scores given by respondents to the attributes or dimensions, while the objective weights were the weights inferred statistically from judgments of multiattribute or multidimensional alternatives via regression or analysis of variance estimation techniques.

Further, it should be noted that most of these studies used the additive rating scale (ARS) procedure as the bottom-up strategy and social/clinical judgment (SCJ) analysis as the top-down strategy. Now, while neither of these particular procedures was used in this study, it is still worthwhile to briefly review the literature comparing them since there are...
Some commonalities in the difference between these procedures and the procedures investigated in this dissertation and will therefore be helpful in understanding the rationale behind the research questions investigated in this dissertation. After each individual study is reviewed, a short summary of: (a) the types of strategies compared, (b) the nature of the preference scores obtained (i.e., values or utilities), and (c) the particular response modes used in the study will be presented.

One of the earliest of such comparative studies between preference measurement strategies was reported by Hoffman (1960), in his classic article on the “paramorphic” representation of clinical judgment. In his study, Hoffman (1960) reported “relatively good correspondence” between the individual attribute preference scores obtained via SCJ analysis and self-explication using the point distribution scaling method (i.e., dividing 100 points among the attributes used to describe the alternative in accordance with the relative importance of the attributes). He also found that the distributions of the subjective weights were more flat or uniform that the distribution of the statistically inferred weights. However, since Hoffman (1960) only presented the graphic plots of weight pairs for selected subjects, no conclusion beyond a simple subjective impression that the weight profiles seem essentially to agree is possible. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) the ARS procedure used judgmental distribution, while SCJ analysis used judgmental ratings as the response mode.
Hoepfl and Huber (1970) conducted another of the earlier studies in this area. Although their primary intent was to test the predictive ability of self-explicated models, they also constructed a Huber-hybrid additive model (although they did not specifically identify it by this name) and estimated the predictive ability of both in terms of predicting students' evaluations of hypothetical university faculty members using a 0-100 point rating scale. Since the focus of this literature review is only on the comparison of different measurement strategies, only the comparative aspects of this study, i.e., on the comparison between the ARS procedure and Huber-hybrid additive modeling will be reviewed here. In this regard, Hoepfl and Huber (1970) found that Huber-hybrid additive modeling outperformed the ARS procedure. In terms of parameter estimates, they found that the preference scores for attributes or dimensions obtained by the ARS procedure were significantly more uniform (i.e., flat) than those yielded by the multiple regression used in conjunction with Huber-hybrid additive modeling (Hoepfl and Huber, 1970). This latter finding was previously reported by Hoffman (1960) and Shepard (1964).

Two major drawbacks of this study need to be pointed out. First, there was no control for an order effect -- subjects always performed the overall evaluation task before the subjective evaluation task. This could have potentially confounded the results. Second, and more important, there was no cross-validation. This drawback could have seriously biased the results in favor of the Huber-hybrid modeling procedure, since the weights used in this modeling procedure were estimated on the evaluations of the same sample which were used as the criterion for validation. This is nothing but
an example of the issue of capitalization on the chance characteristics of the sample, which was mentioned earlier. Therefore, the conclusion reported in the study that the Huber-hybrid modeling procedure was superior to the ARS procedure in terms of predictive ability is tentative at best.

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and Huber-hybrid modeling, (b) both methodologies yielded values, and (c) both methodologies used judgmental ratings as the response mode.

Summers, Taliaferro, and Fletcher (1970) performed a similar study, comparing the ARS procedure and SCJ analysis in terms of description and accuracy of human judgment. The SCJ analysis involved rating multiattribute alternatives on a 10-point scale, while the subjective weights in the ARS procedure were obtained by having the subjects distribute 100 points among the four attributes used to describe the alternatives (i.e., classify into four classes). An interesting aspect of this study was the use of subject-written protocols, describing the judgment process as perceived by the subjects themselves, concurrently with the above parameter elicitation procedures (Summers, Taliaferro, and Fletcher, 1970).

Like the Hoepfl and Huber (1970) study reviewed above, these researchers found that the preference scores for attributes or dimensions obtained by the ARS procedure were more uniform (i.e., flat) than those yielded by SCJ analysis. Further, the respondents indicated (both, via the subjective weights and the written protocols) that they used each attribute about equally in judging the alternatives, while the results of the SCJ analysis indicated that respondents focused on only one particular
dimension in judging the alternatives. Regarding accuracy of the two procedures, it was found that SCJ analysis reproduced the subjects' multiattribute judgments more accurately than the ARS procedure. Specifically, it was found that when the ARS procedure was used to reproduce subjects' multiattribute judgments, the variance accounted for was approximately 20% less than that accounted for by SCJ analysis. It should be noted that this is hardly surprising, since there was no holdout sample used; thus the discussion about capitalization on chance which was presented with respect to the Hoepfl and Huber (1970) study is also relevant here. Finally, another drawback of this study was that there was no control for order effects -- subjects always made multiattribute judgments before giving self-explicated weights. Therefore, like the Hoepfl and Huber (1970) study, the conclusions reported by the study authors need to be interpreted with caution.

Finally, it should be noted that Brehmer and Qvarnstrom (1976) have taken objection to the methodology used in this study -- specifically, Summers, Taliaferro, and Fletcher (1970) found little correspondence between the two sets of weights expressed in terms of the proportion of variance accounted for by the attributes or dimensions. Brehmer and Qvarnstrom (1976) argued (and demonstrated) that if the SCJ weights were redefined in terms of slope coefficients, a close correspondence between the two sets of weights was obtained. In support of this argument, Brehmer and Qvarnstrom (1976) noted that further evidence for their hypothesis was provided by the results of Slovic (1969), which showed that greater correspondence between the two sets of weights would be obtained if the
SCJ weights were defined in terms of the difference between the grand mean and the mean for each level of each dimension.

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) the ARS procedure used classification into four classes (i.e., judgmental distribution) and SCJ analysis used judgmental ratings as the response mode.

Huber, Daneshgar, and Ford (1971) also compared SCJ analysis, the ARS procedure (with self-explicated and equal weights), and Huber-hybrid modeling (using additive and multiplicative model forms) in terms of their respective abilities to predict job ratings and job choices, using persons seeking professional employment in public schools as subjects for the study. These researchers were also interested in the effect of experience of subjects on the predictive ability of the different procedures. Judgments of multiattribute alternatives as well as individual attributes were obtained on a 0-100 point rating scale (Huber, Daneshgar, and Ford, 1971).

Keeping in mind the emphasis of this literature review, only the results of comparing measurement strategies will be discussed here, not the results on comparing the different model forms in terms of predictive accuracy. It was found that experience of subjects was an important moderator of the predictive accuracy of different measurement strategies. Specifically, for experienced subjects, SCJ analysis was the best predictor of job ratings among all the different measurement strategies investigated in this study, followed by Huber-hybrid modeling and finally the ARS procedure. However, for inexperienced subjects, Huber-hybrid modeling
was the best predictor for job ratings, followed by the ARS procedure and finally SCJ analysis. In terms of predicting job choices, the ARS procedure was the most accurate for both experienced as well as inexperienced subjects, followed by Huber-hybrid modeling and finally SCJ analysis (Huber, Daneshgar, and Ford, 1971). However, this study suffered from the same drawbacks as the previous studies reviewed in this section, i.e., the order of presentation of strategies was not randomized and there was no cross-validation of the results using a holdout sample.

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure, Huber-hybrid modeling, and SCJ analysis, (b) all methodologies yielded values, and (c) all methodologies used judgmental ratings as the response mode.

The issue of experience in self-explicated versus statistically inferred strategies was also addressed in a study by Slovic, Fleissner, and Bauman (1972). It is interesting to note that this was one of the earliest comparative studies which used functional measurement as the statistically inferred measurement strategy -- most of the other comparative studies used SCJ analysis as the statistically inferred (or top-down) measurement strategy. This study, however, did not use predictive accuracy as the basis for comparison. Instead it used correspondence between the individual attribute preference scores using the two strategies. It was found that inexperienced subjects exhibited greater correspondence between the two measures than experienced subjects. These results were explained on the basis of routine decision making behavior by experts as opposed to students (who served as
the inexperienced subjects), who were recently trained in the mechanics of the evaluation process and might consequently have paid more attention to all attributes in decision making merely on account of unfamiliarity with the task. It should also be noted that subjective weights were more variable across subjects than were the statistically inferred weights (Slovic, Fleissner, and Bauman, 1972).

Like the other studies reviewed in this section so far, this study also did not control for order effects (subjective weighting tasks always followed overall evaluations) and did not use a holdout sample for cross-validation. It should also be noted that this study was not designed to be a methodological comparison; it was intended to serve as a demonstration of the use of statistically inferred strategies in measuring preferences. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and functional measurement, (b) both methodologies yielded values, and (c) functional measurement used judgmental ratings, while the ARS procedure used classification into eight classes (i.e., judgmental distribution) as the response mode.

Alpert (1971) conducted a comparison of the ARS procedure and SCJ analysis in terms of predictive ability of judgmental ratings. The results of his analysis favored SCJ analysis over the ARS procedure. An important point to note about this study was the use of cross-validation. Indeed, this was one of the earliest studies in the context of comparing preference measurement methodologies that used cross-validation as a basis for comparison. However, no attempt was made to control for an ordering,
effect. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) both methodologies used judgmental ratings as the response mode.

Blood (1971) obtained measures of satisfaction with five attributes of job situation, overall job satisfaction, and a rank-ordering of the importance of the five attributes from 380 clerical workers. After reordering each subject’s attributes from most important to least, Blood (1971) computed statistical measures of importance for the five attributes, now defined not by their content area but by their subjectively assigned importance rank. Blood (1971) reasoned that if both subjective and regression-based measures of importance were valid they should agree in terms of the preferential ordering of attributes in terms of importance. However, the results of his analysis indicated no relationship between these two measures. These results were interpreted as a demonstration of the invalidity of subjective weighting or self-explicated procedures in general (Blood, 1971).

John and Edwards (1978) have taken exception to Blood’s (1971) conclusions. They note that since Blood (1971) obtained data for each different job situation from a different subject, the measures taken from subjects may not be comparable across subjects. However, using statistical measures of importance is appropriate only if the meaning of the scale scores is invariant across all respondents. Hence, John and Edwards (1978) note that without some evidence that the scores obtained have the same meaning for all the respondents of Blood’s (1971) study, the conclusions of this study should be treated with caution. Apart from this issue, Blood’s (1971) study
suffered from the same drawbacks as most previous studies in this area -- lack of controlling for an ordering effect and absence of cross-validation with a holdout sample.

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) information pertaining to the response modes used in this study were was not provided in the article.

Green, Carmone, and Wind (1972) reported an empirical comparison between a rank-order judgment conjoint strategy and the ARS procedure and found that the ARS procedure exhibited reasonable accuracy (mean Spearman correlation = 0.85) in predicting the actual observed rankings of multiattribute alternatives. Like most of the other studies reviewed in this section, this study did not control for ordering effects (subjective weighting always followed the conjoint strategy) and did not use a holdout sample for cross-validation. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and rank-order judgment conjoint methodology, (b) both methodologies yielded values, and (c) the rank order conjoint methodology used classification into four classes, while the ARS procedure used judgmental ratings as the response mode.

Permut (1973) compared the ARS procedure with SCJ analysis in terms of their rankings of the importance of ten different dimensions of instructor quality. Although he found that the ARS procedure predicted the observed ranks with moderate accuracy (mean Spearman's correlation =
0.55), it should be noted that this was substantially less than the accuracy reported by Green, Carmone, and Wind (1972). An important point to note about this study -- pointed out by John and Edwards (1978) -- was the prevalence of negative statistical weights for dimensions which were presumed to be positively correlated with instructor quality. This suggests the possibility of multicollinearity among the dimensions, which could have biased the regression estimates. However, since Permut (1973) did not report the intercorrelations among the ten dimensions, this hypothesis cannot be verified. Like most of the other studies reviewed in this section, this study did not control for ordering effects (subjective weighting always followed rating the multidimensional alternatives) and did not use a holdout sample for cross-validation. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) SCJ analysis used judgmental ratings, and the ARS procedure used classification into ten classes as the response mode.

Nystedt and Magnusson (1975) compared the ARS procedure and SCJ analysis in terms of their ability to predict overall evaluations of multiattribute alternatives. Being careful to use a double cross-validation procedure for estimating the multiple correlations resulting from the regression analysis, they found that the overall multiattribute evaluations were more highly correlated with the ARS scores as opposed to the SCJ scores. This study was noteworthy in that it used a double cross-validation procedure; however, like the other studies which have been reviewed in this section, it did not control for an order effect (respondents gave subjective
weights only after giving overall multiattribute evaluations). With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) both methodologies used judgmental ratings as the response mode.

In a frequently cited study, Cook and Stewart (1975) compared seven different scaling methods as part of the ARS procedure in terms of their correspondence with overall (or holistic) preference scores. The motivation behind their study was to test whether the particular scaling method used made a difference with respect to how closely the computed preference scores for any multiattribute alternative corresponded with the overall preference score provided by respondents. Six of the seven scaling methods used judgments as the response mode, while one used choices as the response mode. In spite of the variety of scaling methods used in this study, it was found that there were no significant differences between the seven methods in terms of the computed preference scores obtained via an ARS procedure using these scaling methods. However, the computed scores of each of these seven methods was significantly different than the overall preference scores of multiattribute alternatives. Like most of the other studies reviewed in this section, this study did not control for ordering effects (subjective weighting always followed rating the multidimensional alternatives) and did not use a holdout sample for cross-validation. Although this study did not compare preference measurement methodologies, it was included in this review to emphasize the finding that the particular scaling method used as part of the ARS procedure does not
affect the results of the comparative analysis -- irrespective of whether the scaling methods use judgments or choices as the response mode.

Armstrong, Denniston, and Gordon (1975) conducted a between-subjects comparative analysis between the ARS procedure and SCJ analysis. This study was notable in that it was not a preference measurement study; instead, it investigated the relative merits and demerits of decomposition of a problem in the context of problem-solving. Also, each problem had a "correct" answer -- therefore the accuracy of both strategies could be tested against this answer. The accuracy ratio of each strategy was used as the evaluative criterion. Armstrong, Denniston, and Gordon (1975) operationally defined the accuracy ratio as the ratio of the estimated multiattribute score over the actual multiattribute score (or vice versa, as long as the larger of the two numbers was in the numerator). It was found that for nine out of the 13 respondents, the ARS procedure was significantly more accurate that SCJ analysis; also, the estimates obtained using the ARS procedure were also less likely to have large errors. It was concluded that decomposing a problem led to more accurate estimates of the alternatives under evaluation. Further, the results also showed that the benefits of decomposition were the greatest when subjects' knowledge of the problem area were the poorest. However, like most of the other studies reviewed in this section, this study did not cross-validate the results with a holdout sample.

Scott and Wright (1976) conducted a validation study of SCJ analysis in an industrial purchasing context using alternatives described with two, three, and six attributes. Although the primary purpose of this study was not to compare self-explicated and SCJ analyses, one of the aspects they
considered was the correspondence between parameter weights as determined by subjective weighting and SCJ analysis. In this regard, the Pearson correlations between the two sets of weights were reasonably high in most cases, i.e., mostly above 0.5 (Scott and Wright, 1976).

These researchers also reported a particularly telling comparison between the two sets of weights -- each subject's self-explicated and statistically inferred weights were arranged in descending order, and the ratio of each separate attribute weight to the sum of all the weights was computed. In the two-attribute case, both sets of weights showed that the more important attribute was weighed twice as heavily than the second attribute. In the three-attribute case, the statistically inferred weights indicated that the most important attribute was twice as important as the second most important attribute, but the self-explicated weights showed little difference between these two attributes. However, in the six-attribute case, there were some striking differences -- for instance, the self-explicated weights indicated that the most important attribute was about 50% more important than the fourth most important attribute, while the statistically inferred weights estimated the ratio between these two attributes to be 500% (Scott and Wright, 1976). These results quite dramatically replicate the findings of most previous studies comparing parameter weights obtained via the ARS procedure and SCJ analysis, where it was shown that self-explicated weights are more uniform across attributes as opposed to statistically inferred weights.

Scott and Wright (1976) attributed these results to a lack of self-insight by respondents, because all the other validation tests of SCJ analysis
conducted as part of their study supported the validity of parameter weights derived via SCJ analysis. Speculating about why respondents appeared biased toward overestimating the importance of trivial attributes, Scott and Wright (1976) noted that when asking respondents to provide self-explicated importance weights for different attributes, one was essentially asking the respondent to report the covariation between changes in those attributes and consequent changes in evaluation of alternatives because of the aforementioned changes in attributes. Since such data were not expected to be easily retrievable from memory, hasty decision making by respondents might be responsible for the "lack of self-insight" bias of self-explicated weights. This hypothesis is in accordance with Wright's (1974) treatise of the "harassed decision maker."

Another possible reason forwarded by Scott and Wright (1976) for this observed discrepancy was the effect of social desirability bias on self-explicated weights, in that respondents may have certain notions about societal norms regarding how many and which particular attributes a decision maker should consider in making a decision. It should also be noted that this social desirability bias hypothesis was independently suggested by Nisbett and Wilson (1977) as an argument against self-explicated weighting procedures in general in their influential paper on the validity of self-explicated data. Specifically, Nisbett and Wilson (1977) suggested that people are taught by culture or subculture about how attributes should be weighed in their evaluations -- therefore, when subjects are asked to indicate their subjective preference weights for different attributes, these weights cannot be regarded as indicative of the evaluation process, but rather as
evidence for the ability to use learned rules of evaluation. This hypothesis has received empirical support in a study by Brookhouse, Guion, and Doherty (1986), which manipulated social desirability as a factor in an experiment comparing the ARS procedure and SCJ analysis. In support of Scott and Wright's (1976) and Nisbett and Wilson's (1977) hypothesis, Brookhouse, Guion, and Doherty (1986) concluded on the basis of their results that social desirability bias was more closely related to self-explicated as opposed to statistically inferred parameters. It should be noted that there have been estimation procedures developed which control for social desirability bias in experimental research -- for instance, Holbrook and Moore (1979) have proposed a procedure for assessing the convergent validity of top-down and bottom-up strategies in the case of “socially sensitive perceptions.”

With respect to the particulars of the methodologies compared in both, the Scott and Wright (1976) as well as the Brookhouse, Guion, and Doherty (1986) studies: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) both methodologies used judgmental ratings as the response mode.

In opposition to Nisbett and Wilson's (1977) hypothesis, Surber (1985) reported an analysis which supported the conclusion that self-explicated weights differ from statistically inferred weights because they reflect the overall effect of an attribute on judgment, not just the importance weights of the attribute. This premise is consistent with the use of an averaging model as a combination function instead of an additive model (see Anderson, 1980, 1981).
Wiley, MacLachlan, and Moinpour (1976) compared the ARS procedure and SCJ analysis in terms of preference scores for multiattribute alternatives, parameter weights (i.e., importance weights for each attribute), and ideal points on attributes. With respect to all three indicators, there were significant differences between the ARS procedure and SCJ analysis. Although the median correlation between preference scores for multiattribute alternatives obtained via the two procedures was 0.50, there was no evidence of consistency between the attribute importance weights obtained via the two procedures. These results are consistent with the ones reviewed above. The notable extension of this study was the comparison of the two procedures in terms of their respective ideal points. The rationale behind this comparison is that as long as both procedures agree on the "goal," any differences in preference scores obtained via the two procedures maybe interpreted as indicative of the invalidity of at least one of the procedures. If, however, the two procedures disagree on the "goal," it can be argued that systematic differences between the preference scores obtained via the two procedures are expected, and consequently that neither procedure is invalid per se. The fact that this study found little correspondence between the ideal points obtained via both the procedures supports the latter argument.

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) both methodologies used judgmental ratings as the response mode.
Schmitt (1978) compared the ARS procedure and SCJ analysis in terms of correspondence between preference scores for multiattribute alternatives under changing task situations. Apart from the regular consistency index (i.e., the correlation between actual and predicted judgments), these researchers also used a matching index as a basis for comparison of the two procedures. The matching index is the correlation between predicted values based upon actual attribute-criterion relationships and predicted values based on the subject's weights (either self-explicated or statistically inferred). Schmitt (1978) noted that matching represented the degree to which the subject "knows" the actual relationship among the variables. In this study, subjects received information concerning the actual relationship from labels attached to each attribute as well as from outcome feedback conditions (where comparisons were made between actual judgments and the "correct" criterion values).

The results indicated that the scores obtained via the ARS procedure had lower correlations with the "correct" criterion value that the scores obtained via SCJ analysis. Outcome feedback did, however, have a greater negative effect on consistency as well as matching indices based on scores obtained via SCJ analysis as opposed to scores obtained via the ARS procedure. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) both methodologies used judgmental ratings as the response mode.

Cattin and Weinberger (1979) reported the results of convergent validation between the ARS procedure and rating scale judgment conjoint
methodology, using number of attributes per alternative as an experimental factor. Replicating the results of previous studies, they found that the linear correlation between the preference scores of multiattribute alternatives obtained via the two procedures decreased as the number of attributes used to describe an alternative increased. However, they also found that self-explicated weights did fairly well in predicting the judgmental rating provided by respondents. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and rating scale judgment conjoint methodology, (b) both methodologies yielded values, and (c) the ARS procedure used classification into six or nine (depending on the number of attributes used to describe alternatives) classes, while rating scale judgment conjoint methodology used judgmental ratings as the response mode.

Heeler, Okechuku, and Reid (1979) compared the ARS procedure, rank-order judgment conjoint methodology, and an information display board in terms of the individual attribute preference scores obtained via each procedure. In contrast to most other comparative studies, these researchers used preference measurement methodology as a between-subjects factor in their experiment. Keeping in mind the focus of this literature review, only the results comparing the ARS procedure and rank-order judgment conjoint methodology shall be reviewed here. It was found that the two procedures yielded different preferential rank-orderings of the attributes used to describe each alternative. Unfortunately, Kendall's tau was not reported by the researchers, so one cannot conclude whether these differences in observed rank-orders were significantly greater than that
expected by chance. However, eyeballing the results, it is apparent that there are some serious differences -- for instance, although both procedures agree on which attribute is the most important, the attribute that is ranked second in importance via the ARS procedure is ranked sixth using rank-order judgment conjoint methodology. However, this study did find a significant linear convergence between the preference scores for multiattribute alternatives, as obtained via the two procedures.

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and rank-order judgment conjoint methodology, (b) both methodologies yielded values, and (c) the ARS procedure used classification into ten classes (since each alternative was described in terms of ten attributes), while rank-order judgment conjoint methodology used judgmental rankings as the response mode.

Nutt (1980) compared the ARS procedure and a variant of SCJ analysis proposed by Gustafson, Pai, and Kramer (1971) in terms of the preference scores of attributes and variance in preference scores of attributes. In order to operationalize the ARS procedure, Nutt (1980) used three different scaling methods -- a linear rating scale, a log rating scale, and rank-order weighting. It should be noted that this study also compared the different scaling methods used as part of the ARS procedure in terms of similarities in decisions made on the basis of these preference scores (assuming an additive model form) -- however, the results of this analysis will not be reviewed here since they do not pertain to the focus of this literature review. Consistent with the results of previous studies, Nutt (1980)
found that the statistically inferred preference scores of attributes were distinct from self-explicated weights. However, in contrast to the results of Cook and Stewart (1975), there were significant differences observed among the preference scores of attributes obtained via the different self-explicated scaling methods. The linear rating scale had the least variance associated with the attribute preference scores. In comparison to the linear rating scale, the statistically inferred attribute preference scores had 20% more variance, the log rating scale had about 200% more variance, while the rank-order weighting method had nearly 300% more variance (Nutt, 1980).

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and a variant of SCJ analysis, (b) both methodologies yielded values, and (c) the ARS procedure used judgmental ratings and rankings, while the variant of SCJ analysis used judgmental ratios as the response mode.

Nutt (1981) reported another comparison between the ARS procedure and SCJ analysis, although in this case he did not use the variant of SCJ analysis. Being careful to control for order effects by randomly varying the order of presentation of the two procedures, Nutt (1981) evaluated the accuracy of the predictions made by either procedure against an external managerially relevant standard. It was found that SCJ analysis was significantly more accurate than the ARS procedure in terms of predictive ability as evaluated against this external standard. Since respondent evaluations also favored SCJ analysis over the ARS procedure (these results will be reviewed later in this chapter), Nutt (1981) recommended the use of SCJ analysis over the ARS procedure for decision analysis. With respect to
the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) both methodologies used judgmental ratings as the response mode.

Neslin (1981) compared the ARS procedure with rating scale judgment conjoint methodology in terms of attribute preference scores as well as predictive ability using a holdout sample. An interesting aspect of this study was that it tested two different model forms in the rating scale judgment conjoint methodology -- and main effects only (i.e., additive) model form and a main effects plus selected interactions (i.e., partial multiplicative) model form. Like other comparative studies reviewed in this section, there discrepancies between the two sets of parameter estimates, in terms of not only the preference scores but also the preferential rank-ordering (Neslin, 1981).

As a first measure of predictive ability, Pearson correlations were calculated for each individual between actual and predicted evaluations. It was found that both the main effects only as well as the main effects plus selected interactions models outperformed the self-explicated model. Further, the main effects only model was slightly better than the main effects plus selected interactions model in this measure. As a second measure of predictive ability, the three procedures were compared in terms of their relative abilities to correctly rank order respondents' evaluations of the alternatives. In this regard, once again both the main effects only as well as the main effects plus selected interactions models outperformed the self-explicated model; however, in contrast to the results of the first measure, the
main effects plus selected interactions model was slightly better than the
main effects only model for this measure of predictive ability (Neslin, 1981).

This study was noteworthy in that it incorporated the role of
interactions in the comparative analysis. In doing so, it demonstrated the
difficulty of handling interactions in the ARS procedure, which may be a
drawback -- especially as applied to the measurement of health status
preferences. With respect to the particulars of the methodologies compared
in this study: (a) the measurement strategies compared were the ARS
procedure and rating scale judgment conjoint methodology, (b) both
methodologies yielded values, and (c) both methodologies used judgmental
ratings as the response mode.

In an important study, Akaah and Korgaonkar (1983) empirically
compared the estimated attribute preference scores and the predictive
validity (using holdout samples) of self-explicated, Huber-hybrid, rating
scale judgment conjoint, and hybrid conjoint models. Within each of these
types of strategies, these researchers also compared different model forms.
For instance, in the self-explicated or ARS procedure, the different model
forms compared were weighted and unweighted additive; in the Huber-
hybrid models, the different model forms compared were additive hybrid,
and addilog hybrid, and multiplicative hybrid; in the rating scale judgment
conjoint models, the different model forms compared were the main effects
only and the main effects plus selected interactions model forms; finally, in
the hybrid conjoint models, the different model forms compared were the
main effects only and the main effects plus selected interactions model forms
(Akaah and Korgaonkar, 1983).
A multivariate analysis of variance testing the equality of the normalized attribute weights obtained via the different strategies indicated that there were significant differences in attribute preference scores. To examine the nature of the differences, these researchers computed and then rank-ordered the mean attribute preference scores for each of the model forms tested in this part of the analysis. It should be noted that this part of the analysis did not include the hybrid conjoint models and the model forms which included interactions, since the presence of interaction and/or self-explicated terms in these models did not make their main effects or individual attribute preference scores comparable to those of the additive or main-effects only model forms. Further, the unweighted self-explicated model form was not included in this part of the analysis since it assumed equal individual preference scores. It was found that the Huber-hybrid models yielded similar ranks, the conjoint main-effects only model and the Huber-hybrid models yield similar ranks for the top three attributes, while the weighted self-explicated model differed in preferential rank-ordering from the other models investigated in this part of the analysis (Akaah and Korgaonkar, 1983).

In terms of the predictive ability of the different strategies investigated in this study, the results of the study supported the following conclusions: (a) the rating scale judgment conjoint, and hybrid conjoint models outperformed the Huber-hybrid and self-explicated models in terms of predicting preference scores for multiattribute alternatives used in model calibration (i.e., in the estimation sample), especially when they included interactions; (b) however, when it came to predicting preference scores for
multiattribute alternatives used in the holdout sample, all strategies evaluated were about equal, with only the rating scale judgment conjoint main-effects only model significantly outperforming the self-explicated and Huber-hybrid models; and (c) in terms of predicting the rank-order of multiattribute alternatives in the holdout sample, the rating scale judgment and hybrid conjoint models outperformed the self-explicated and Huber-hybrid models. The authors interpreted these results as generally favorable of rating scale judgment and hybrid conjoint modeling as opposed to self-explicated and Huber-hybrid modeling strategies in terms of predictive ability (Akaah and Korgaonkar, 1983).

Although this study was quite comprehensive in its inclusion of different preference measurement methodologies, two potential limitations of this study need to be noted. First, it did not control for any order effects -- in this regard, the self-explicated tasks always preceded the multiattribute alternative evaluation tasks. Further, the holdout tasks were always the very last thing the respondents did -- the possibility of fatigue in performing the holdout evaluations cannot be ruled out. Indeed, Akaah and Korgaonkar (1983) themselves commented on the presence of substantial "noise" in the holdout data. Second, these researchers made no attempt to control for multiple comparisons when comparing the different model forms in terms of attribute preference scores and predictive ability. Therefore, the probability of committing at least one Type I error was quite high. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure, Huber-hybrid modeling, rating scale judgment conjoint methodology, and hybrid conjoint
modeling, (b) all methodologies yielded values, and (c) all methodologies used judgmental ratings as the response mode.

Leigh, MacKay, and Summers (1984) performed an empirical comparison of the test-retest reliability and predictive ability of the ARS procedure and conjoint methodology using several scaling methods including rank-order judgments, rating scales, and contingent valuation. Since there were no significant differences in test-retest reliability and predictive ability of the conjoint models using the different scaling methods, data from all the conjoint models were aggregated for further analysis. In terms of the test-retest reliability for attribute preference scores obtained via the two strategies, there were no significant differences between the ARS procedure and the combined conjoint models. However, the results for the mean squared difference between the test and the retest attribute preference scores suggested significantly higher reliability (i.e., a lower mean squared difference) for the self-explicated preference scores. Further, the self-explicated multiattribute alternative preference scores produced a significantly higher test-retest reliability than the combined conjoint models. In terms of predictive ability, there were no significant differences between the two strategies (Leigh, MacKay, and Summers, 1984).

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and conjoint methodology, (b) both methodologies yielded values, and (c) the ARS procedure used pricing judgments, while the conjoint methodology used rank-order judgments, rating scale judgments, and pricing judgments as the response mode.
Adelman, Sticha, and Donnell (1984) compared the relative effectiveness of multiattribute weighting procedures as a function of the number of attributes used to describe the alternatives and the distribution of correct attribute preference scores. The aspect of this study that is pertinent to the present discussion is the comparison of the individual attribute preference scores obtained via SCJ analysis and the ARS procedure implemented using ratio weighting, paired comparisons, dividing 100 points among the attributes in relation to their importance, and the standard gamble scaling methods. In this regard, three dependent variables were used -- the standard deviation of the individual attribute preference scores generated by each procedure, the accuracy of each multiattribute score (defined as the Pearson correlation between the true scores as determined by an external criterion and the predicted scores), and respondent evaluations of each of the procedures. Keeping in mind the focus of this section of the literature review, only the results of the first two dependent variables will be discussed here; results about respondent evaluations will be discussed in a subsequent section of this chapter.

Regarding the results for standard deviations of individual attribute preference scores, it was found that the ARS procedure with ratio weighting and the method of paired comparisons yielded the largest standard deviations among the five scaling methods studied. Regarding the results for the accuracy of the multiattribute scores, it was found that accuracy decreased as the number of attributes used to describe alternative increased, regardless of the particular preference measurement strategy or scaling method. Further, ratio weighting and the method of paired comparisons
were found to be more accurate when the true distribution of individual attribute scores was peaked, while the other scaling methods studied were found to be more accurate when the true distribution of individual attribute scores was flat (Adelman, Sticha, and Donnell, 1984).

The major contribution of this study lies in the fact that it demonstrated that measures of precision (such as standard deviation) and measures of accuracy do not always go hand-in-hand. Therefore, these results underscore the need to be careful in selecting a dependent variable for any comparison of preference measurement methodologies since the particular dependent variable chosen could dictate the choice of one methodology over another. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values for the most part, although the standard gamble used as part of the ARS procedure yielded utilities, and (c) the ARS procedure used judgmental ratios, binary choices, indifference judgments, and judgmental classification (i.e., dividing 100 points among the attributes in relation to their importance), while SCJ analysis used judgmental ratings as the response mode.

Jaccard, Brinberg, and Ackerman (1986) evaluated the convergence between individual attribute preference scores using self-explicated weights, paired comparisons, and conjoint methodology (among other measures not pertinent to this discussion). These researchers found very low linear correlations and Kendall's rank-order concordances between the individual attribute preference scores obtained via these procedures. It should be noted
that the emphasis in this study was on the individual attribute preference scores, as opposed to the multiattribute preference scores. Just because there is low convergence at the individual attribute level does not necessarily mean that there will be low convergence at the multiattribute alternative level. This is because different measurement methodologies may induce different model forms or combination functions from respondents. Unfortunately, multiattribute preference scores were not provided by the authors of this study. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and rank-order judgment conjoint methodology, (b) both methodologies yielded values, and (c) the ARS procedure used judgmental ratings and binary choices, while rank-order judgment conjoint methodology used judgmental rankings as the response mode.

Reilly and Doherty (1989) reported an empirical comparison between the ARS procedure and SCJ analysis. Like other studies reviewed in this section, they found that the preference scores of multiattribute alternatives obtained via SCJ analysis were more highly correlated with preference scores of multiattribute alternatives in a holdout sample than the preference scores of multiattribute alternatives obtained via the ARS procedure. However, an interesting finding of their study was that in spite of this finding, a fair portion of the subjects were able to correctly identify their judgment policy. These researchers concluded that self-explicated and statistically inferred scores may measure different constructs (Reilly and Doherty, 1989).
It should be pointed out that the Reilly and Doherty (1989) study may suffer from a selection bias. A total of 40 subjects initially provided both self-explicated and multiattribute judgments. The initial correlations were run on these 40 subjects. Eleven of these 40 subjects returned to the investigators for feedback about their policies (the remaining 29 had their policies mailed to them). Out of these 11, seven correctly identified their judgment policies. This is the result which Reilly and Doherty (1989) took as evidence that a fair portion of their respondents could correctly identified their judgment policies. Now, it is possible that these 11 subjects were specially motivated about judgment policy in general -- after all, they did come back on their own for feedback about their own policies. Thus it is possible that these subjects may different from the other subjects in that in wanting to learn more about their own judgment policy, they were more insightful than regular subjects about judgment policy in general. Therefore, without further evidence and replication of these findings, this result of Reilly and Doherty (1989) needs to be treated with caution.

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) both methodologies used judgmental rankings as the response mode.

Reardon and Pathak (1989) compared the ARS procedure and Huber-hybrid modeling in terms of their ability to predict contingent valuation measures. Specifically, the ARS procedure used two model forms -- subjectively weighted and unweighted. Moreover, three types of Huber-hybrid models were compared -- additive hybrid, addilog hybrid, and
multiplicative hybrid. Finally, it should be noted that the alternatives for which contingent valuation measures were provided were generated by means of a main-effects only orthogonal factorial (a design commonly used in conjoint methodology). The results generally supported the additive and addilog Huber-hybrid models as being the best predictors of the contingent valuations in both estimation as well as the holdout samples.

The noteworthy aspect of this study was the use of contingent valuation measures as a criterion for preference measurements. Contingent valuation, as noted earlier, has a firm basis in the theory of welfare economics, and thus provides a theoretically justifiable criterion for preference measurement. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and Huber-hybrid modeling, (b) both methodologies yielded values, and (c) both methodologies used judgmental rankings as the response mode.

Usher (1991) compared the ARS procedure and conjoint methodology in terms of their respective abilities to predict multiattribute preference scores in a holdout sample. The ARS procedure was implemented with a variety of scaling methods, including equal weighting, rank-order weighting, and constant sum weighting. It was found that conjoint methodology significantly outperformed equal and rank-order weighting, but not constant sum weighting, in terms of predictive ability (Usher, 1991). Although this study also compared the two methodologies in terms of respondent evaluations, these results will be reviewed in a subsequent section of this chapter. With respect to the particulars of the methodologies
compared in this study: (a) the measurement strategies compared were the ARS procedure and conjoint methodology, (b) both methodologies yielded values, and (c) the ARS procedure used judgmental rankings and classification judgments, while conjoint methodology used binary choices as the response mode.

There have also been some studies published in the literature which have compared top-down and bottom-up strategies in terms of their relative ability in predicting an external criterion. One of the earliest of such studies was by Oskamp (1967), who reported a study of clinical diagnosis using the MMPI. Oskamp (1967) had experts rate 200 MMPI profiles for whether the patient was being hospitalized for psychiatric or medical reasons. Statistical weights for each of the 13 MMPI scales were obtained via SCJ analysis. Self-explicated weights were directly assessed for each subject as a "... subjective report of what proportional weight he thought he was attaching to each variable" (Oskamp, 1967 p.412). Since the profiles were of actual patients, "true" weights were obtained statistically using the dichotomy of actual diagnosis (i.e., psychiatric versus medical reasons) as the criterion variable (Oskamp, 1967).

Unfortunately, the analyses comparing the overall multiattribute scores resulting from the various weighting schemes were not reported. However, Oskamp (1967) did note that the average number of scales which each subject reported using (i.e., subjective weights) matched the average number of significant statistically inferred weights. Both these numbers in turn were close to the "true" number of weights, as determined by the criterion variable. Further, the ordering and the relative magnitudes of both
the self-explicated as well as the statistically inferred weights matched the "true" weights closely and to approximately the same degree, although the statistically inferred weights were closer to the "true" weights than were the self-explicated weights.

Lathrop and Peters (1969) evaluated the decomposition principle in the context of course evaluations. Students gave judgmental ratings of a number of attributes for each course and an overall course evaluation. The individual attribute and overall evaluation ratings were separately averaged and the averages treated as objective value measures. Students who were not enrolled in these courses then served as subjects in one of two conditions. In the first condition, they were given the average score of a class on each attribute and asked to make an overall judgmental evaluation; in the second condition, they were asked to assign weights to each of the individual attributes. It was found that across a number of conditions, the second condition (i.e., bottom-up) afforded better prediction of the overall evaluative ratings than did the first condition, despite the fact that the self-explicated weights were decidedly nonoptimal as compared to the weights derived from multiple regression (Lathrop and Peters, 1969). With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and SCJ analysis, (b) both methodologies yielded values, and (c) both procedures used judgmental ratings as the response mode.

Stillwell, Barron, and Edwards (1983) used a real-world bank’s credit scoring model as the criterion in their comparison of the ARS procedure (using a variety of scaling methods) and rating scale judgment conjoint
methodology. These researchers presented the results of their analysis in two parts. In the first part of the analysis, they compared the individual attribute preference scores obtained via each procedure. It was found that the rating scale judgment conjoint methodology agreed with the criterion weights in terms of preferential ordering of attributes, while the ARS procedure did not -- regardless of scaling method. The second part of the analysis involved comparing the decisions made on the basis of the overall preference scores for multiattribute alternatives and the decisions made by the credit scoring model of the bank. In this regard, it was found that the ARS procedure outperformed the rating scale judgment conjoint methodology. Further, in agreement with Cook and Stewart (1975), these researchers found very little differences between the different scaling method used as a part of the ARS procedure, at least in terms of the basis for comparison investigated in this study. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and rating scale judgment conjoint methodology, (b) both methodologies yielded values, and (c) the ARS procedure used judgmental ratios, three types of rank-order judgments, and distribution judgments (i.e., divide 100 points among the attributes in accordance with their importance), while rating scale judgment conjoint methodology used judgmental ratings as the response mode.

Currim and Sarin (1984) reported a particularly pertinent empirical comparison for this literature review -- their study compared both the procedures which were investigated in this dissertation, i.e., conjoint methodology and the conditional utility function-based procedure.
Although their study used rank-order and rating scale judgment conjoint methodology and not discrete choice conjoint methodology, their study is still very significant to the present one on account of the relative similarities between rating scale judgment and discrete choice conjoint strategies in terms of parameter estimates and predictive validity (as reviewed earlier). The basis of comparison for this study was the relative predictive performance of the two procedures in holdout samples (Currim and Sarin, 1984).

Specifically, Currim and Sarin (1984) compared statistically inferred (i.e., top-down) and algebraically computed (i.e., bottom-up) preference measurement strategies. Within each type of strategy, they compared three different types of models -- additive conjoint (see Luce and Tukey, 1964), measurable value (see Dyer and Sarin, 1979a), and axiomatic utility (see Keeney and Raiffa, 1976; von Neumann and Morgenstern, 1944, 1947). Therefore, in all, the comparison involved six different models. A between-subjects design was used, with the between-subjects factor being the type of strategy (i.e., top-down or bottom-up).

Currim and Sarin (1984) considered two types of holdout samples, the first for decisions under certainty and the second for decisions under risk. For the case of decisions under certainty, there were two holdout validation samples. The first consisted of five choice questions (wherein respondents selected either of the two alternatives or expressed indifference between the two alternatives in the choice set), while the second consisted of a ranking task requiring the respondent to rank order four alternatives. Currim and Sarin (1984) reported the results of four tests for decisions under certainty:
(a) the ability of the model to pick the preferred alternative from a pair of alternatives (with respect to the first holdout sample), (b) the correlation between an individual's rank ordering of the four alternatives and his model's predicted ranking of these four alternatives (with respect to the second holdout sample) as measured with Kendall's tau, (c) the ability of the model to predict the most preferred alternative from the set of four alternatives (with respect to the second holdout sample), and (d) the ability of the model to predict the ranks of all four alternatives correctly (with respect to the second holdout sample).

The results of both the holdout validation samples favored the top-down strategies over the bottom-up strategies. First, considering only the top-down strategies, there were no significant differences among the three models (i.e., additive conjoint, measurable value, and axiomatic utility) in terms of predictive ability for any of the four tests mentioned above. All these models were fairly accurate in predicting the preferred alternative from a pair of alternatives, having 75% to 81% correct predictions. Regarding the correlations between individuals' rank-ordering and the models' predicted rank-ordering of the four alternatives in the second holdout sample, the calculated Kendall's tau coefficients ranged from 0.66 to 0.72 (which are sufficiently high). Regarding the ability of the models to predict the most preferred alternative from the set of four alternatives, these models correctly predicted between 60% to 74% of such cases. Finally, regarding the ability of the models to predict the ranks of all four alternatives correctly, the models predicted between 59% and 66% of all ranks correctly (Currim and Sarin, 1984).
Next, considering only the bottom-up strategies, there were no significant differences between the three different models in terms of predicting the preferred alternative from a pair of alternatives (the range was 59% to 68%). However, there were significant differences among the three models in terms of correlations between individuals’ rank-ordering and the models’ predicted rank-ordering of the four alternatives in the second holdout sample -- the calculated Kendall’s tau coefficients ranged from 0.21 for the measurable value model to 0.24 for the axiomatic utility model to 0.54 for the additive conjoint model. Further, all three models were about equivalent with regard to predicting the most preferred alternative from the set of four alternatives, with ranges from 46% (for the measurable value model) to 55% (for the additive conjoint model). Finally, on the criterion of proportion of all ranks correctly predicted, the additive conjoint model significantly outperformed the other two models -- 56% to 30% (for the measurable value model) and 36% (for the axiomatic utility model). Comparing the results of the top-down to the bottom-up strategies, Currim and Sarin (1984) concluded that the top-down strategies were superior to the bottom-up strategy, at least in the case of decisions under certainty.

For comparing the predictive ability of models under risk, the holdout sample consisted of four decisions made under risk using the standard gamble using the probability equivalence approach. The basis for comparison was the Pearson correlation between the observed $p$ value and the $p$ values predicted by the different models. Again, the top-down strategies outperformed the bottom-up strategies. Specifically, the Pearson correlations with the top-down strategies ranged from 0.62 for the additive
conjoint model to 0.84 for the axiomatic utility model, while on the other hand, the Pearson correlations with the bottom-up strategies ranged from 0.08 for the measurable value model to 0.16 for the additive conjoint model. Therefore, the results of decisions under risk supported those of decisions under certainty (Currim and Sarin, 1984).

In discussing the relative merits and demerits of the two types of strategies, Currim and Sarin (1984) considered five criteria -- information requirements, experimental design considerations, implementation, model construction, and interpretive power of the results. Regarding information requirements, they noted that for an individual-level analysis, top-down strategies required a larger number of responses to gain degrees of freedom; in contrast, the number of responses required in bottom-up strategies was small. Regarding experimental design considerations, top-down strategies were clearly simpler to design, compared to bottom-up strategies. Regarding implementation, again top-down approaches were preferred since they could be implemented easily using written instructions and stimulus cards; in contrast, bottom-up strategies usually required personal interactive interviews (which are usually more expensive). Regarding model construction, both strategies are about equivalent, since standard statistical packages can be used to estimate models in top-down strategies and it is possible to program computers to perform the same operations for models in bottom-up strategies (although such packages are not widely available). Finally, regarding interpretation, bottom-up models can potentially provide a more clear description of respondents' decision making processes on account of their easy incorporation of interactions as well as information.
about risk coefficients. Although this is also possible for top-down strategies, most standard computer packages do not allow such flexibility. The authors concluded by noting that all these factors -- including the results of the predictive ability of the strategies -- should be taken into account while selecting a particular preference measurement strategy (Currim and Sarin, 1984).

With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the conditional utility function-based procedure and conjoint methodology, (b) both methodologies yielded values as well as utilities, and (c) the conditional utility function-based procedure used indifference judgments, while conjoint methodology used judgmental rankings and judgmental ratings as the response mode.

Although there has been a vast body of literature comparing top-down and bottom-up preference measurement strategies in a variety of contexts, it is noteworthy that this body of literature has not affected the field of health status preference measurement. Indeed, there is quite a paucity of published empirical comparisons between difference preference measurement strategies in the context of health status preference measurement. It appears that health status preference measurement researchers are more concerned with the scaling method as opposed to the measurement strategy. This maybe cause for concern, since -- as noted by Froberg and Kane (1989a) -- the scaling method is nothing but an artifact of the measurement strategy.
Llewellyn-Thomas, Sutherland, Ciampi, et al. (1984) published an empirical comparison between the ARS procedure and rating scale judgment conjoint methodology in the context of measuring patients' preferences for different aspects of voice function and sound of voice of laryngeal cancer patients undergoing radiation therapy. The two strategies were compared on the basis of test-retest reliabilities and between-occasion reliability of mean scores reported at the beginning and end of the course of therapy. For both of these indicators, the rating scale judgment conjoint methodology yielded higher reliability coefficients than the ARS procedure. Further, in agreement with most of the studies reviewed above, the ARS procedure yielded more flat individual attribute preference scores than the rating scale judgment conjoint strategy. This study is notable mainly because of the fact that it compared alternative preference measurement methodologies in the context of health status preferences. On account of the fact that it replicated the results of other studies in different contexts, this study showed that the results obtained in other contexts may also apply in the context of health status preference measurement. With respect to the particulars of the methodologies compared in this study: (a) the measurement strategies compared were the ARS procedure and rating scale judgment conjoint methodology, (b) both methodologies yielded values, and (c) both procedures used judgmental ratings as the response mode.

Summary

Having reviewed a substantial amount of studies in the published literature which have compared top-down and bottom-up preference
measurement strategies, it is prudent to summarize the major results of these studies and note some of the shortcomings which need to be addressed in future research. First of all, one of the more robust findings of these studies is that individual attribute preference scores are usually flatter when measured using bottom-up strategies as opposed to top-down strategies. The reasons for this discrepancy are less clear. It is possible that making each attribute explicit in the bottom-up strategies forces decision makers to consider the "true" importance of each attribute; on the other hand, the social desirability hypothesis (Brookhouse, Guion, and Doherty, 1986; Nisbett and Wilson, 1977; Scott and Wright, 1976) cannot be ruled out as a causative factor for these observed discrepancies between individual attribute preference scores.

Second, and on a related note, the linear correlation between individual attribute preference scores determined using top-down and bottom-up preference measurement strategies is typically low. However, the linear correlations between preference scores for multiattribute alternatives determined via the two strategies is usually higher. The results of the study reported by Wiley, MacLachlan, and Moinpour (1976) are particularly noteworthy in this regard. These dual observations are not necessarily contradictory to each other -- it is possible that different measurement strategies invoke different combination functions from respondents in their information integration processes. Therefore, although the individual attribute preference scores may be uncorrelated, the final multiattribute preference scores may be more highly correlated since the individual attribute preference scores were combined in different ways on account of
different combination functions or model forms of respondents' preference functions. Although not specifically formulated with this application in mind, change of process theory (Mellers, Ordonez, and Birnbaum, 1992) -- which will be reviewed later in this chapter -- can accommodate such explanations.

This has implications for the choice of the particular dependent variable in comparative analysis of preference measurement methodologies - it is more instructive to compare overall multiattribute preference scores as opposed to comparing individual attribute preference scores. This is because individual attribute preference scores do not account for interactions among the attributes used to describe any alternative under evaluation. Indeed, in the presence of interactions among the attributes of an alternative, individual attribute preference scores have little meaning by themselves. This is especially true for the case of health status preference measurement, since the likelihood of having clinically significant interactions among health status dimensions or attributes is quite high (as explained in Chapter II).

Further, with respect to the overall multiattribute preference scores obtained as a result of both strategies, the scores obtained via top-down strategies are usually higher than those obtained via bottom-up strategies. This could be on account of the fact that when respondents evaluate multiattribute alternatives, they inherently consider the interactions among the attributes or dimensions used to describe the alternatives. On the other hand, interactions need to be explicitly accounted for while using the bottom-up strategies -- since this explicit accounting may be different from the implicit accounting for interactions, it is possible that these differences
would be manifest in different preference scores for the multiattribute alternatives. It should be noted that this particular discussion even applies within a given response mode and for a given attitude toward risk -- i.e., the difference between top-down and bottom-up strategies both of which use judgmental ratings under riskless conditions could be explained by the explicit versus implicit accounting for interactions hypothesis outlined above. Further, whether the implicit accounting for interactions during top-down strategies is sufficient to overcome the expected differences between value-based and utility-based preference scores (as reviewed later in this chapter) is still open to question. However, the particular preference measurement methodologies compared in this study should shed some light on this last question.

In terms of predictive ability of the top-down versus bottom-up strategies, the published studies have shown an increasing sophistication in measurement techniques. Most of the initial comparative studies evaluated predictive ability of these strategies did not use a holdout sample and cross-validation procedures, and hence ran the risk of capitalizing on the chance characteristics of the estimation sample, which in turn favored the top-down measurement strategies in terms of predictive ability. However, the later comparative studies made it a point to use holdout sample to evaluate the predictive ability of both strategies -- in fact, some studies have even used double cross-validation samples (for e.g., see Nystedt and Magnusson, 1975).

The results of these comparisons of predictive accuracy of top-down versus bottom-up strategies are equivocal. The few studies which have used an external criterion as the standard have usually favored bottom-down over
top-down strategies in terms of ability to predict this external criterion (for e.g., see Lathrop and Peters, 1969; Stillwell, Barron, and Edwards, 1983). However, other studies, which have used the holdout sample approach have usually favored top-down over bottom-up strategies in terms of predictive ability (for e.g., see Alpert, 1971; Currim and Sarin, 1984; Reilly and Doherty, 1989; Usher, 1991).

In drawing a conclusion from these equivocal results, two points should be considered. The first point deals with the veracity of the external criterion used, if any -- i.e., just how true a measure of the underlying construct is this external criterion? The second point deals with the issue of holdout samples in general -- holdout samples are usually multiattribute alternatives; therefore, the cognitive processes used to evaluate the holdout samples would be similar to those used in top-down strategies, thus “stacking the deck” in favor of these strategies in terms of predictive ability using the holdout sample approach. It should be noted these points to consider move in opposite directions, i.e., the first works against bottom-up strategies, while the second works against top-down strategies. As noted earlier, both these points should be kept in mind in drawing conclusions about the relative predictive abilities of top-down versus bottom-up preference measurement strategies.

A few general points need to be noted about the studies evaluated in this section. First, most of the earlier studies were not designed for comparative purposes -- they were meant to test the validity of the bottom-up strategies for preference measurement by the principle of convergent
validation with overall multiattribute preference scores. This is possibly a reason why most of these studies did not bother using holdout samples.

Second, several of the studies reviewed in this section did not control for order effects. This is a serious drawback of these studies, since learning effects could have operated and thus have been potential confounding variables in these studies. It is necessary to control for such order effects via randomization of task order in comparative within-subjects comparative studies of preference measurement methodologies, unless there are compelling theoretical or practical reasons for not doing so.

One major drawback of SCJ analysis stems from its emphasis on representative design, and the consequent existence of substantial multicollinearity. As Ward (1962) and Darlington (1968) have noted, different statistically inferred importance weights from multiattribute alternatives with correlated attributes or dimensions seldom lead to the same conclusions with respect to the relative weighting of information or even to the preferential ordering of variables. Although some of the studies reviewed in this section have experimentally manipulated the attributes so that they have zero or near-zero intercorrelations, many of the studies did not report any such manipulations. It is possible, therefore, that these other studies did not control for multicollinearity -- consequently, the results of these studies should be treated with caution.

Finally, the majority of the studies reviewed in this section assumed an additive model form for respondents’ preference functions. It is possible that different results may have been obtained had the researchers allowed (and tested) for the existence of different model forms. This only serves to
reinforce the idea introduced earlier, of using multiattribute preference scores instead of individual attribute preference scores as the dependent variables in comparative studies of preference measurement methodologies.

**Value-based Versus Utility-based Preference Scores**

The distinction between preferences measured under certainty (i.e., values) and preferences measured under risk (i.e., utilities) was introduced in Chapter II. This section will review the theoretical rationale for the differences between value and utility, introduce the theoretical relationships between value functions and utility functions, and review the published empirical studies which have compared value and utility functions. Unlike the issue of top-down versus bottom-up strategies, this issue has been addressed in the health status preference measurement literature. Indeed, as explained below, it forms a critical piece in the health status measurement methodology adopted by the McMaster group of researchers, who have arguably been one of the more influential forces in the field of health status measurement.

As noted earlier, Multiattribute Preference (MAP) theory is concerned with establishing an axiomatic basis for various MAP model forms. In other words, it seeks to explain the nature of the combination function \( C \) in Figure 3. There are three model forms which are most commonly encountered in practice -- additive, multiplicative, and multilinear. MAP theory also specifies the independence conditions which are necessary and sufficient for the existence of these model forms. Within the general framework of MAP
theory, two fundamentally different measurement approaches have been
developed to model preferences for multiattribute alternatives, depending
on whether the decision problem involves uncertainties or not, and further,
the type of uncertainty (i.e., risk versus ambiguity versus ignorance).

The first modeling approach is concerned with modeling preferences
under conditions of certainty. Included in this approach are two types of
measurement theories -- conjoint measurement (Krantz, 1964; Krantz and
Tversky, 1971a; Luce and Tukey, 1964; Tversky, 1967b) and difference
measurement (Dyer and Sarin, 1979a). Both, conjoint measurement and
difference measurement theory specify the conditions under which a riskless
value function \( v \) can be constructed which preserves the preference order
among multiattribute alternatives and which can be expressed as a simple
aggregate of single attribute value functions \( w \). The best known model forms
under these measurement theories are the additive and multiplicative model
forms (Barron, von Winterfeldt, and Fischer, 1984; Dyer and Sarin, 1979a,

As derived under conjoint measurement theory, the additive model
form is represented by:

\[
v(x) = \sum w_i v_i(x) \quad (4.1)
\]

and the multiplicative model form by:

\[
v(x) = \prod w_i v_i(x) \quad (4.2)
\]
where

\[ i = 1 \ldots n \]

\[ x \in X = \text{multiattribute alternative} \]

\[ w_i = \text{scaling constant} \]

\[ x_i = \text{level of } x \text{ on the } i\text{-th attribute} \]

\[ v_i(x) = \text{i-th single attribute value function} \]

\[ v(x) = \text{overall value function for multiattribute alternative } x \]

As derived under difference measurement theory, the additive model form is represented by:

\[ v(x) = \sum w_i v_i(x) \quad (4.3) \]

and the multiplicative model form by:

\[ 1 + w v(x) = \prod [1 + w w_i v_i(x)] \quad (4.4) \]

where

\[ i = 1 \ldots n \]

\[ x \in X = \text{multiattribute alternative} \]

\[ w = \text{interaction parameter } (-1 < w < 0 \text{ or } w > 0) \]

\[ w_i = \text{scaling constant } (0 \leq w_i \leq 1) \]

\[ x_i = \text{level of } x \text{ on the } i\text{-th attribute} \]

\[ v_i(x) = \text{i-th single attribute value function} \]

\[ v(x) = \text{overall value function for multiattribute alternative } x \]
Further, as specified by difference measurement theory, the interaction parameter \( w \) is related to the scaling constants \( w_i \) by the equation:

\[
1 + w = \prod (1 + w_i)
\]

(4.5)

and

if \( \sum w_i > 1 \), then \(-1 < w < 0\),

if \( \sum w_i = 1 \), then \( w = 0 \) and the additive model form holds, and

if \( \sum w_i < 1 \), then \( w > 0 \).

The difference between conjoint measurement and difference measurement lies in the particular interpretation of the value function \( v \) -- in the case of conjoint measurement, \( v \) is an ordinal additive or multiplicative value function, while in the case of difference measurement, \( v \) is an cardinal strength of preference additive or multiplicative value function. In this regard, the value function is inferred on the basis of rank-order judgments of multiattribute alternatives in the case of conjoint measurement, as opposed to strength of preference judgments between alternatives in the case of difference measurement. This should not be interpreted to mean that the value function in conjoint measurement yields non-interval-level data; the ordinal judgmental data are converted via a monotone transformation for estimation purposes, so that the resulting statistically inferred parameters will be intervally-scaled. On the other hand, the value function in difference measurement is directly measured on a cardinal scale, and therefore does
not require any transformation of the raw data to make it into an intervalu-
scaled function.

However, although the value function \( v \) is an appropriate guide for
preferences among sure things, nothing in either conjoint measurement or
difference measurement theory guarantees that the expectation of the value
function \( v \) is appropriate for selecting among lotteries with multiattribute
alternatives. It is in this regard that the second modeling approach within
MAP theory is of use -- expected utility theory, which is concerned with
modeling preferences under conditions of risk. Based on the seminal work of
von Neumann and Morgenstern (1944, 1947), which was subsequently
extended by Savage (1954), expected utility theory provides an axiomatic
basis and rationale for constructing a risky utility function \( u \) which
preserves the preference order for riskless alternatives and at the same time
its expectation preserves the preferences among lotteries for such
alternatives. Applied to the multiattribute situation, several decomposition
forms of \( u \) have been developed (see Keeney and Raiffa, 1976). However, the
best known model forms are the additive and multiplicative model forms.
Indeed, Keeney and Raiffa (1976) have noted that most practical decision
problems can be adequately represented by either an additive or
multiplicative model form.

As derived under expected utility theory, the additive model form is
represented by:

\[
u(x) = \sum k_i u_i(x)
\]  

(4.6)
and the multiplicative model form by:

\[ 1 + k u(x) = \prod [1 + k k_i u_i(x_i)] \quad (4.7) \]

where

\[ i = 1 \ldots n \]

\[ x \in X = \text{multiattribute alternative} \]

\[ k = \text{interaction parameter (-1 < } k < 0 \text{ or } k > 0) \]

\[ k_i = \text{scaling constant (0} \leq k_i \leq 1) \]

\[ x_i = \text{level of } x \text{ on the } i\text{-th attribute} \]

\[ u_i(x) = \text{i-th single attribute utility function} \]

\[ u(x) = \text{overall utility function for multiattribute alternative } x \]

Further, as shown by Keeney (1974), the interaction parameter \( k \) is related to the scaling constants \( k_i \) by the equation:

\[ 1 + k = \prod [1 + k k_i] \quad (4.8) \]

and

if \( \Sigma k_i > 1 \), then \(-1 < k < 0\),

if \( \Sigma k_i = 1 \), then \( k = 0 \) and the additive model form holds, and

if \( \Sigma k_i < 1 \), then \( k > 0 \).
It should be noted that, as measured by expected utility theory, the utility function encodes the decision maker's strength or intensity of preference as well as attitude toward risk. Indeed, utilities measured according to expected utility theory confound a decision maker's strength or intensity of preference and attitude toward risk, i.e., these two components of the utility function cannot be separated according to expected utility theory. However, further insight into the decision maker's attitude toward risk can be inferred from the shape of his or her utility function, via several different methods such as Pratt's (1964) measure of local risk aversion, Jensen's inequality (DeGroot, 1970), or Keeney and Raiffa's (1976) ordering between the certainty equivalent and the expected outcome of lotteries. In this regard, a utility function which is linear represents an attitude of strict risk-neutrality, a utility function which is convex represents a strictly risk-seeking or strictly risk-prone attitude, while a utility function which is concave represents an attitude of strict risk-aversion.

Richard (1975) has drawn a key distinction between the additive and multiplicative multiattribute utility functions in terms of their incorporation of attitude toward risk -- the additive multiattribute utility function assumes that a decision maker is strictly multiattribute risk neutral for every possible pair of attributes, whereas the multiplicative multiattribute utility function assumes that a decision maker is either: (a) strictly multiattribute risk-averse with respect to every pair of attributes and regards the attributes in question as substitutes for each other or (b) strictly multiattribute risk-seeking or risk-prone with respect to every pair of attributes and regards the attributes in question as complements to one another.
Torrance, Boyle, and Horwood (1982) have noted that the intuitive interpretation of this characterization is that substitute attributes are such that an improvement in one is relatively satisfying, while an improvement in two or more is not that much better. On the other hand, in the case of complementary attributes, an improvement on any one attribute alone is not very useful, while a simultaneous improvement on several is much better. Further, a decision maker's attitude toward risk can be related to the interaction parameter \( k \) in the multiplicative multiattribute utility function. Specifically: (a) if \(-1 < k < 0\) (i.e., if \( k \) is negative), the decision maker is strictly multiattribute risk-averse and (b) if \( k > 0 \) (i.e., if \( k \) is positive), the decision maker is strictly multiattribute risk-seeking or risk-prone (Richard, 1975).

An intuitive way of determining the attitude toward risk of the decision maker is to plot his or her utilities against values (i.e., utilities on the y-axis and values on the x-axis). The position of the points in two-dimensional space relative to the 45° angle will be indicative of the attitude toward risk of the decision maker. Specifically, since the 45° angle represents an attitude of risk-neutrality (i.e., \( u = v \)), if the resulting curve in two-dimensional space lies below the 45° angle the decision maker is risk-prone or risk-seeking, while if the resulting curve in two-dimensional space lies above the 45° angle the decision maker is risk-averse. These relationships are illustrated in Figure 5.
Legend:
RA = risk-averse
RN = risk-neutral
RS = risk-seeking

Figure 5: Graphical Illustration of Attitude Toward Risk
[adapted from Torrance, Zhang, Feeny, et al., (1992)]
The preceding introduction has briefly described the two modeling approaches within MAP theory. As noted earlier, preferences measured under conditions of certainty are known as value-based preferences or values (i.e., the \( v \) function), while preferences measured under conditions of risk are known as utility-based preferences or utilities (i.e., the \( u \) function). Both types of preference functions map an attribute set into real numbers and are unique up to a positive linear transformation, but \( v \) results from an axiomatization of riskless decision making while \( u \) holds under an axiomatization of risky decision making. Behaviorally, the difference between a value function and a utility function has been described in terms of the incorporation of attitude toward risk -- specifically, a value function encodes the strength of preference while a utility function encodes the strength of preference and attitude toward risk (Dyer and Sarin, 1982). The purpose of this section is to investigate the theoretical basis for estimating a relationship between the value and utility functions and to review the few published empirical studies which have attempted to do so.

Measurement theoreticians have established that the value and utility functions are related to one another by a strictly increasing transformation, (for e.g., see Keeney and Raiffa, 1976; Raiffa, 1968b). Indeed, the theoretical literature supports the hypothesis that the relationship between a riskless value function \( (v) \) and a risky utility function \( (u) \) should be of the form

\[
    u = h(v)
\]  

(4.9)

where \( h = \) some monotonic (and usually nonlinear) transformation.
Intuitively, the transformation \( h \) may be viewed as reflecting the subject's attitude toward risk. This follows from the fact that the utility function \( u \) reflects both intensity of preference for certain outcomes as well as attitude toward risk and the value function \( v \) reflects only the intensity of preference for certain outcomes -- thus, the transformation function \( h \) should reflect only attitude toward risk. Thus, Dyer and Sarin (1982) have proposed that true attitude toward risk -- which they refer to as relative risk attitude -- be defined in terms of the shape of the transformation function \( h \). Specifically, if, for some interval of \( v \): (a) \( h \) is linear, then the decision maker is relatively risk-neutral (b) \( h \) is convex, then the decision maker is relatively risk-prone or relatively risk-seeking, and (c) \( h \) is concave then the decision maker is relatively risk-averse.

However, neither conjoint measurement theory, nor difference measurement theory, nor expected utility theory by themselves provide a rationale for any specific closed form relationships between the value and utility functions. As von Winterfeldt (1979) and Barron, von Winterfeldt, and Fischer (1984) have noted, in principle, the shape and aggregation form of the value and utility functions can be quite different, e.g., \( v \) may be multiplicative while \( u \) may be additive, all \( v \)'s may be linear in attribute scales while all \( u \)'s may be nonlinear, or vice versa.

However, when particular decomposed model forms (e.g., the additive or multiplicative model forms discussed above) are assumed, it is possible to establish closed form functional relationships between value and utility functions. A common way to relate these functions is by the uniqueness theorems of their respective measurement theoretic
representations. For instance, if both $u$ and $v$ are additive and order preserving, then there exist constants $a$ and $b$ (where $a > 0$) such that:

$$u = av + b \quad (4.10)$$

Dyer and Sarin (1979a) have used this equation to derive closed form functional relationships between value and utility functions for the cases where both functions are either additive or multiplicative. Other measurement theoreticians have conducted similar research, relating utility and value functions by some special class of transformations (for e.g., see Barron, von Winterfeldt, and Fischer, 1984; Bell and Raiffa, 1988a; Keeney and Raiffa, 1976; Pollak, 1967; von Winterfeldt, 1979). The general conclusion of this body of theoretical research is that value and utility functions are related to one another by either a linear, logarithmic, or exponential transformation, depending on the particular model form of the value and utility functions and assuming that both value and utility functions are either additive or multiplicative.

Barron, von Winterfeldt, and Fischer (1984) have studied the theoretical relationships between additive and multiplicative value and utility functions. Their classification is summarized in Table 2. According to this table, the value and utility functions are related by a linear transformation if both are additive in nature; a logarithmic transformation if the value function is multiplicative and the utility function is additive in nature; and an exponential transformation if the utility function is multiplicative in nature -- regardless of the whether the value function is additive or multiplicative in nature.
Table 2: Theoretical Relationships Between Value and Utility Functions

<table>
<thead>
<tr>
<th>Value Function ($v$)</th>
<th>Additive $u = v$</th>
<th>Multiplicative $u = \ln(1+wu)/(1+w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additive $u = v$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiplicative $1+ku = (1+k)^x$</td>
<td>1+ku = (1+wu)$^{\ln(1+wu)/(1+w)}$</td>
<td></td>
</tr>
</tbody>
</table>

Legend:  
$\ln$ = natural log  
$k$ = scaling constant for utility function  
w = scaling constant for value function

[adapted from Barron, von Winterfeldt, and Fischer (1984)]
These results have also been formally derived by von Winterfeldt, Barron, and Fischer (1980) using a method of proof based on the theory of functional equations (Aczel, 1966). Since the theoretical derivation is not of primary interest to this literature review it will not be reviewed here. What is of concern is the expected theoretical relationships between value and utility functions, so that hypotheses may be drawn about the functional relationships between the preference measurement methodologies investigated in this dissertation. In this regard, a two step procedure needs to be followed: (a) first, determine the nature of the model form of the value and utility functions, and (b) look up the appropriate cell in Table 2 corresponding to the value-utility combination determined in step (a) above in order to decide which particular type of transformation is applicable.

Having established the different types of theoretical relationships between value and utility functions, it is now necessary to review the relevant published literature which has empirically tested such functional relationships between values and utilities. One of the earliest studies to assess both value and utility functions on the same subjects (11 inmates from a state prison) was reported by Tversky (1967a). As part of his dissertation research, Tversky (1967a) applied additive conjoint measurement to selling price judgmental evaluations of cigarette packs and candy bags under conditions of certainty and risk. The additive model provided an excellent fit to both, value as well as utility functions. He found that all 22 value and utility functions (11 subjects x 2 attributes) were almost linear, so that after transforming one of them linearly, \( u = v \). This is consistent with the expected
functional relationship shown in Table 2, for the case when both the value as well as the utility functions have an additive model form.

Tversky’s (1967a) analysis also found that the utilities and values differed markedly for the same commodities, with the utilities exceeding the values for all but one subject. Although this observation is hardly surprising in light of the behavioral interpretation of value and utility functions provided earlier in this section (and indeed has been replicated in many subsequent studies in diverse research situations), it needs to be evaluated in the context of the state of preference measurement methodology in the mid-1960s. In this regard, the descriptive validity of the expected utility model had not been subject to much empirical investigation. Therefore, Tversky’s (1967a) potential explanation of these findings as indicative of a positive utility for gambling is quite remarkable. Indeed, this study could be considered as the beginning of a series of experiments conducted by Tversky and his colleagues in the late-1960s through the 1970s which empirically tested the descriptive validity of the expected utility model, and culminated with the development of prospect theory (Kahneman and Tversky, 1979).

What is important to note for the purpose of this dissertation, is that this study was one of the earliest to demonstrate a difference between utility-based preference scores and value-based preference scores, with utilities being higher than values.

As part of his dissertation research, Fischer (1976, 1977) conducted an experimental comparison between riskless and risky preference function assessment methods. Ten subjects evaluated 27 hypothetical alternatives using both holistic riskless judgmental ratings and holistic risky indifference
judgments. The order of presentation of these two types of judgmental tasks was randomized across subjects. After these holistic multiattribute judgments were made, each subject constructed decomposed riskless and risky preference models, following the ARS procedure and the conditional utility function-based procedure. Therefore, each of the ten subjects implicitly or explicitly evaluated the 27 multiattribute alternatives using five different preference measurement methodologies: (a) holistic riskless multiattribute judgmental ratings, (b) holistic risky multiattribute indifference judgments, (c) riskless additive explicitly decomposed judgmental ratings using the ARS procedure, (d) risky additive explicitly decomposed indifference judgments using the conditional utility function-based procedure, and (e) risky multiplicative explicitly decomposed indifference judgments using the conditional utility function-based procedure. It should be noted that in practice, one set of judgments was used to construct both the additive and multiplicative models using the conditional utility function-based procedure. Each subject was run individually and under close supervision of the experimenter. Fischer reported the results of his experimental comparison in two articles, the first dealing with the holistic preference measures (Fischer, 1976) and the second with the explicitly decomposed preference measures (Fischer, 1977).

Regarding the holistic preference measures [i.e., (a) and (b) above], Fischer first computed what he termed the "ordinal reliability" of the two measures -- the within-subject rank-order concordance coefficient (or Kendall's tau) between the two measures. These coefficients ranged from a high of 0.98 to a low of 0.69, with a median of 0.90. It was concluded that
these coefficients were "... encouragingly high" (Fischer, 1976 p.134). It should be noted that this notion of "ordinal reliability" totally ignores the role of response mode effects on preference measurement. This is because Fischer implicitly assumed that the "ordinal reliability" should be 1.0, since both measurement methodologies are normatively equivalent, i.e., should yield the same preferential ordering of alternatives. However, since the different measurement methodologies used different response modes to measure respondents' preferences, they may not be descriptively equivalent. Therefore, although Fischer (1976) attributed the less than perfect (i.e., less than 1.0) concordance coefficients to random respondent error, this could have been a systematic effect of response mode.

The holistic judgments were also subjected to axiomatic as well as statistical tests for the functional form of the value and utility functions. In other words, the model form of the value and utility functions was subject to axiomatic and statistical testing. The axiomatic analysis tested the hypothesis that subjects' preferences were ordinally additive in the conjoint measurement sense (i.e., according to the simple independence, joint independence, and cancellation axioms of conjoint measurement theory). The results of the axiomatic analysis showed that in only one out of the ten subjects could an additive model form be conclusively rejected in the case of the riskless holistic evaluations, i.e., holistic value functions; on the other hand, the additive model form could be conclusively rejected for three of the ten subjects in the case of the risky holistic evaluations, i.e., holistic utility functions (Fischer, 1976).
Apart from the axiomatic analysis, Fischer (1976) also subjected the holistic riskless and risky judgments to multiple regression analysis to determine whether the respondents' value and utility functions could be accurately predicted by linear statistical models and also to test the combination function (i.e., model form) of the value and utility functions via inferential statistics. Considering the riskless holistic judgmental ratings, it was found that additive models generally afforded excellent predictions, with adjusted $R^2$ values ranging from 0.88 to 0.99. Still, five out of the ten subjects exhibited statistically significant departures from additivity, i.e., significant interaction terms in the regression equation. Similar results were obtained for the risky holistic indifference judgments, although the $R^2$ values were generally lower and the standard errors larger than for the riskless holistic judgmental ratings. Specifically, adjusted $R^2$ values ranged from 0.79 to 0.99; however, six out of ten subjects exhibited statistically significant departures from additivity, i.e., significant interaction terms in the regression equation (Fischer, 1976). Therefore, from a descriptive point of view, additive models may seriously misrepresent these subjects' judgmental processes. This merely reinforces Anderson and Shanteau's (1977) warning on the potential drawbacks of using $R^2$ in explanatory model testing.

Fischer (1976) observed that the most striking finding of the analysis of holistic riskless and risky judgments was the apparent contradiction between the results of the axiomatic and inferential statistical tests for determining the model form of the value and utility functions of the ten respondents. Therefore, it appears that different analysis procedures may
support the existence of different model forms for preference functions. Given this disparity, the question that arises is which analytical procedure is the "better" one. Unfortunately, there is no simple answer to this question. However, in terms of the analytical procedures used in this study, it should be remembered that the axiomatic theory of conjoint measurement does not have an error theory, unlike that of inferential statistics. Therefore, in the case of the axiomatic analysis, one cannot know what degree of confidence to place in the results of the analysis. However, both the analytical techniques used in this study supported the existence of different model forms of preference functions under conditions of certainty and risk. Therefore, it is possible that one of the effects of risk (or uncertainty in general, although this notion does require extending the scope of the results of this study) is to change the combination function (i.e., function C in Figure 3) respondents use to integrate information in preference measurement tasks.

Regarding the explicitly decomposed preference measures [i.e., (c), (d) and (e) above], Fischer (1977) investigated the validity of the ARS and conditional utility function-based procedures as preference measurement strategies, in accordance with the principles of convergent validation (outlined in Chapter II), in terms of ordinal as well as linear convergence. With respect to ordinal convergence, Fischer (1977) computed within-subject rank-order concordance coefficient (or Kendall's tau) between the preference measures obtained via the five different methodologies. It should be noted that since Kendall's tau is designed to measure the concordance between only two variables, Fischer (1977) computed ten sets of concordance
coefficients, for each of the measures taken two at a time. It was found that the median within-subject concordance coefficients between the various measures were generally in the mid- to high-0.80's. The same results generally held at the individual level, although two respondents had concordance coefficients in the 0.60's (Fischer, 1977). Once again, the issue of response mode effects was not addressed. The potential drawbacks related to response mode effects were mentioned earlier, and so will not be repeated here.

With respect to linear convergence, Fischer (1977) tested the hypothesis that the holistic and explicitly decomposed preference measures should be linearly related. The rationale behind hypothesizing a linear relation was that if both measures were measured at an interval level of measurement (as they were intended to be measured), they should be linearly related. Fischer (1977) used the coefficient of determination ($R^2$) as an index of the strength of the linear association between the two measures. Considering the degree of linear convergence between the holistic risky [i.e., (b) above] and the explicitly decomposed risky preference measures [i.e., (d) and (e) above], it was found that there was strong evidence for linear convergence between these measures, taken two at time (median $R^2$ in all cases was above 0.85). This result was even stronger for the degree of linear convergence between the holistic riskless [i.e., (a) above] and the explicitly decomposed riskless preference measures [i.e., (c) above], with the median $R^2$ being above 0.90. With the exception of two subjects, these results generally held at the individual level also (Fischer, 1977).
Regarding the functional relationship between riskless and risky preference measures, Fischer (1977) found strong evidence for a linear relationship between value and utility functions for these respondents. Specifically, the median $R^2$ value for the linear convergence between the riskless ARS measure [i.e., (c) above] and the additive and multiplicative risky conditional utility function-based measures [i.e., (d) and (e) above, respectively] were 0.93 and 0.92 respectively. Once again, these results held at the individual level. In addition, the median $R^2$ values for the linear convergence between the holistic riskless and risky measures [i.e., (a) and (b) above], between the riskless holistic measure [i.e., (a) above] and the additive and multiplicative risky conditional utility function-based measures [i.e., (d) and (e) above, respectively], and between the risky holistic measure [i.e., (b) above] and the riskless ARS measure [i.e., (c) above] were all quite substantial, ranging from 0.84 to 0.87. The same result generally applied at the individual level, although a few $R^2$ values were in the mid-0.60's or low-0.70's (Fischer, 1977).

Commenting on these results, Fischer (1977) himself noted that from a theoretical standpoint, there was no reason to expect a linear relationship between value and utility functions and that the strong linear associations found in this study were "... somewhat unexpected" (Fischer, 1977 p.312). Indeed, as discussed earlier, the theoretical literature supports the hypothesis that the relationship between a riskless value function ($v$) and a risky utility function ($u$) should be of the form $u = h(v)$, (where $h$ = some monotonic -- usually nonlinear -- transformation). As noted previously, the transformation function ($h$) may be viewed as reflecting the subject's attitude
toward risk — if the transformation function is linear, the subject is risk-neutral; if the transformation function is concave, the subject is risk-averse; if the transformation function is convex, the subject is risk-seeking or risk-prone.

In recognition of these theoretical underpinnings of the relationship between value and utility functions, Fischer (1977) assessed the relationship between the riskless and risky holistic preference measures [i.e., (a) and (b) above] by means of a series of polynomial regressions. Specifically, he constructed a hierarchy of polynomial regression models (i.e., simple linear, linear plus quadratic, and linear plus quadratic plus cubic) to empirically test for the presence of nonlinearities in the relationship between riskless and risky holistic preference measures. The results of this polynomial regression indicated that five out of the ten subjects showed statistically significant departures from linearity. These results demonstrated that the relatively high level of linear convergence between the various preference measures did not preclude the possibility of systematic nonlinearities (Fischer, 1977). This merely serves to reinforce Anderson and Shanteau's (1977) warning about "weak inference with linear models" which was discussed previously.

It should be noted that Fischer (1977) himself explicitly cautioned against using the results of the linear convergence between the measures as the "final word" in his concluding paragraph: "... the high level of linear convergence ... might be interpreted as suggesting that applied decision analysts can ignore subtle distinctions between risky and riskless or additive
and multiplicative evaluation models. The author feels that this conclusion is unwarranted. ..." (p.313).

This study was significant in terms of its methodological contributions of the field of preference measurement. By having each respondent evaluate multiattribute alternatives using five different preference measurement methodologies, Fischer was able to perform a variety of analyses for both the holistic and explicitly decomposed preference measurement strategies. In this regard, this was the first published study to adopt the convergent validation approach to risky explicitly decomposed preference measures. The strength of the convergent validation procedure used in the study is remarkable, considering that the experimenter made no attempt to assure that the decomposed utility models would be consistent with the holistic utility models. In fact, as opposed to regular applications of decision analysis, no consistency checks whatsoever were employed. Although the issue of validating decomposed preference measurement strategies against holistic measures has been criticized in the decision analysis literature (see, for e.g., Stillwell, Barron, and Edwards, 1983), in the absence of an external criterion of validation, this particular validation process still has its merits. The major weakness of the above study is its small sample size and the fact that one cannot generalize from the context of the study (i.e., hypothetical job evaluations) to all possible situations in which such preference measurement strategies may be used. Still, the methodology used in this study was used as an exemplar in some subsequent studies and indeed has been very influential in shaping the methodology used in this dissertation.
Although Fischer (1976, 1977) performed polynomial regressions to test for nonlinearities in the relationship between the value and utility functions estimated in his study, he did not empirically test for particular functional relationships between these two types of preference functions. Therefore, although he found empirical evidence of significant nonlinearities for five out of his ten subjects, he could not specify the particular form of the functional relationship between the value and utility functions of these respondents. Accordingly, he was unable to infer the behavioral significance of these nonlinearities, i.e., whether the respondents were risk-seeking or risk-averse. This particular analysis was subsequently conducted on Fischer's data on holistic riskless and risky preference measures [i.e., (a) and (b) above] by Barron, von Winterfeldt, and Fischer (1984) using Barron and Person's (1979) Holistic Orthogonal Parameter Estimation (HOPE) procedure.

The results of this analysis indicated that for the riskless case, seven out of the ten subjects had additive model forms; on the other hand, for the risky case, only two out of the ten subjects had additive model forms. Further, in the risky case, seven out of the eight multiplicative models had negative $k$ values, indicating multivariate risk-aversion. The differences between these results and the results of the axiomatic and multiple regression analyses reported by Fischer (1976) serve to reinforce the fact that the determination of a particular model form for preference functions may be dependent on the estimation procedure used.

Further, on the basis of applying the HOPE procedure to the holistic riskless and risky preference measures, Barron, von Winterfeldt, and Fischer
were able to estimate \( \omega \) and \( k \) values for the multiplicative model forms for the value and utility functions respectively. Accordingly, they derived the theoretically predicted relationship between value and utility functions for each subject. Five out of the ten subjects had nonlinear relationships between their value and utility functions. This is the same number as that determined in the polynomial regression run by Fischer (1977). However, as noted earlier, the results of the polynomial regression did not allow easy interpretation of the behavioral significance of these nonlinearities. In contrast, the \( k \) values of the multiplicative model forms for the five subjects who had nonlinear relationships between their value and utility functions was negative, indicating an attitude of risk-aversiveness (Barron, von Winterfeldt, and Fischer, 1984).

As a first check on the validity of the theoretical relationships, plots of observed utilities and observed values were checked. As noted earlier, such graphs should be a straight 45° line for linear relationships between value and utility functions (i.e., risk-neutral subjects) and be above the 45° diagonal line for nonlinear relationships between value and utility functions indicating risk-aversiveness. Visual inspection of such graphs for the linear relationships generally supported the theoretical predictions for all the five subjects who had linear relationships between their value and utility functions. Regarding the nonlinear relationships, the theory provided good predictions of the relationships between value and utility functions for three of the five subjects who had nonlinear relationships between their value and utility functions. However, the data of the remaining two subjects deviated
substantially from theoretical predictions (Barron, von Winterfeldt, and Fischer, 1984).

The derived transformation function \( h \) was then applied to each subject's directly assessed holistic values to obtain predicted scores of holistic utilities. Finally, to determine the extent to which predictions of \( u \) were improved by modeling the relationship between \( u \) and \( v \), the "adjusted" utilities obtained as a result of the modeling process were compared to the "unadjusted" utilities obtained by arbitrarily assuming \( u = v \). The purpose of this comparison was to estimate if the theoretically-predicted exponential relationship between value and utility functions was better than a linear relationship between the two functions. By definition, for the five subjects who had linear relationships between their value and utility functions, the two scores were indistinguishable. However, for the other five subjects, using minimization of root mean squared error as the measure of goodness of fit, the "adjusted" utilities yielded a superior fit for four out of the five subjects (Barron, von Winterfeldt, and Fischer, 1984). Therefore, the theoretical relationships presented in Table 2 were supported by these data.

Barron, von Winterfeldt and Fischer's (1984) reanalysis of Fischer's (1976, 1977) data demonstrated that the theoretical relationships between value and utility functions had empirical support. This study was a significant one in that it provided theoretical justification for determining utility functions via a two-stage procedure, wherein value functions are determined first and then subsequently converted into utility functions via the appropriate transformation. This two-stage procedure has been cited as being easier and less expensive to implement than assessing utility functions
directly (Barron, von Winterfeldt and Fischer, 1984; von Winterfeldt, 1979). Indeed, this two-stage procedure has been adopted by the McMaster group of researchers in their research on measuring health status preferences (Torrance, Zhang, Feeny, et al., 1992).

Currim and Sarin's (1984) study comparing the predictive ability of top-down and bottom-up strategies had been reviewed in the previous section of this chapter. As noted previously, these researchers compared risky and riskless preference models within each type of strategy. The model form in all cases was additive. Apart from the results reviewed above, Currim and Sarin (1984) also estimated the functional relationship between the axiomatic utility and measurable value models in their study (no reason was given as to why they did not test the functional relationship between the additive conjoint model and the axiomatic utility model). Specifically, they tested two types of relationships between these models -- linear and exponential. It was found that the sum of squared error between the computed utilities and the actual stated utilities was minimum for 40 out of 43 subjects when an exponential functional relationship was assumed (Currim and Sarin, 1984).

Further, Currim and Sarin (1984) used the exponential model for all subjects to convert their riskless preference scores on the measurable value model to utilities. This transformed utility model was then used to predict indifference judgments for the four questions in the holdout sample (as described earlier). The correlation between the predicted and stated utilities was found to be 0.78. Observing the fact that the correlation of the "unadjusted" measurable value model was 0.64, Currim and Sarin (1984)
argued that certainty models could provide better predictions under conditions of risk if made to undergo a transformation so as to convert their values to utilities.

The fact that an exponential relationship provided a better fit to the data than a linear relationship is not in keeping with Barron, von Winterfeldt, and Fischer's (1984) theoretical classification, which was presented in Table 2. It is, however, possible that the axiomatic utility model used for this part of the study was not descriptively valid. Some support for this contention stems from the fact that when multiplicative and additive model forms were compared for the axiomatic utility models in terms of the proportion of correct predictions on the first holdout sample, the multiplicative model forms outperformed the additive model forms. Given this finding, it could be that the functional relationship that was estimated was between the "true" axiomatic utility model form (i.e., the multiplicative model form) and the riskless model. If this were indeed the case, the exponential relationship that was observed would be consistent with the theoretical classification given by Barron, von Winterfeldt, and Fischer (1984).

Apart from these few studies, there have not been many other published ones which have directly compared value and utility functions in the non-health-care literature. There have, however, been a few published studies which have assessed both types of functions but have not compared them directly. For example, a study by Eliashberg (1980) assessed two-attribute value and utility functions of 66 students in the context of housing selection. Although Eliashberg's (1980) study was meant to be a
demonstration of the assessment of value and cardinal utility preferences to management scientists and not a comparison between the two types of preferences, he did compare different assessment techniques to determine both value functions as well as utility functions in terms of ability to predict holistic evaluations under certainty. An indirect comparison between the two types of preference measures showed that the value functions outperformed the cardinal utility functions in terms of prediction of preferences under conditions of certainty in the holdout sample. It was concluded that the cost of the additional complexity in assessing cardinal utilities may not be warranted for predictive purposes in conditions of certainty (Eliashberg, 1980). No reason was provided in the article as to why it was decided to use a holdout sample of evaluation under certainty to determine the predictive accuracy of the cardinal utility models (which were developed for decision making under risk). Further, since the value and cardinal utility scores were not reported, one cannot even tell whether the preference scores were significantly different from one another.

Another study which did not directly compare value functions with utilities but still may have implications for the value versus utility distinction was reported by Fischhoff, Slovic, and Lichtenstein (1980). These researchers performed two tasks on a group of college students and members of the League of Women Voters, asking them to express their preferences with regard to the number of lives lost as a result of a catastrophic event. In the first task, subjects had to choose between a two-outcome gamble and a sure loss equal to the expected value of the gamble. In the second task, three shapes (linear, convex, and concave) of a function
representing social cost of fatalities were displayed, and subjects had to indicate which curve they thought society should evaluate lives in multifatality situations. In both tasks, the perspective adopted was that of a societal decision maker. Further, both tasks were framed in terms of loss of lives (Fischhoff, Slovic, and Lichtenstein, 1980).

It was found that the majority of the respondents preferred the gamble in the first task (which implies a concave disutility function) but chose the convex function in the second task. This same result was reproduced in a subsequent experiment (Slovic, Fischhoff, and Lichtenstein, 1982), in which the first task was a choice between two gambles. Thus, the certainty effect (i.e., always pick the sure-thing option due to aversion to gambling) was ruled out a reason for the discrepancy between value and utility functions. These researchers attributed these findings to inconsistency in preferences. However, Krzysztofowicz (1983a) noted that these results could be explained in terms of relative risk aversion. Specifically, if the social cost function $v$ was viewed as a construct distinct from the disutility function $u$, the observed convexity of $v$ and concavity of $u$ would imply that the preferences of most subjects were composed of increasing marginal social cost and an attitude of relative risk-seeking (Krzysztofowicz, 1983a).

As noted earlier, this study did not compare value and utility functions. In fact, it did not even assess either function. Still, it was reviewed in this section in order to show that the construct of relative risk attitude could explain some of the puzzling findings of studies in experimental psychology which are “untenable” with the rational theory of choice. This
particular example was chosen because it could potentially help in understanding the phenomenon of preference reversal.

Finally, an important study by Krzysztofowicz (1983a) needs to be reviewed. Krzysztofowicz (1983a) tested the relative risk aversion hypothesis in two experiments, in both of which he constructed value as well as utility functions. Assuming that the decision maker's relative risk attitude was constant over the range of outcomes in the analysis, he empirically constructed value functions for all his respondents. Then, he obtained each respondent's utility function by a superposition of the relative attitude of the subject on the subject's previously determined value function. This was done by determining a certainty equivalent $x_1$, such that $u(x_1) = 0.5$ is compared with an equal difference point $x_2$, such that $v(x_2) = 0.5$. Using theorems proposed by Dyer and Sarin (1982), this information was used to determine each subject's relative risk attitude and also the specific nature of the risk attitude (i.e., risk-neutral, risk-seeking, or risk-neutral). Using a theorem proposed by Pratt (1964), the appropriate transformation was determined and applied to the constructed value function, thus yielding a utility function for each respondent. In this way, each respondent's utility function was obtained by means of a two-stage procedure instead of via direct assessment (Krzysztofowicz, 1983a).

For his analysis, Krzysztofowicz (1983a) pooled the data from these two experiments with the results of two other studies from the literature (Allais, 1953; Allais, 1979; Sarin, Dyer, and Nair, 1980). These other studies had only reported value and utility functions at the individual level; the appropriate transformation was determined for these functions by
Krzysztofowicz (1983a). After verifying the hypothesis of constant relative risk attitude, the analysis tested whether the departure of the utility function from the value function is significant to warrant the concept of relative risk attitude i.e., identity hypothesis. A likelihood ratio test indicated that the identity hypothesis should be rejected. It was further determined that the relative risk attitude of all the subjects pooled together (n=34) was normally distributed. On the basis of one of the experiments which was conducted, wherein each subject provided two utility functions for the same attribute and decision context but for two different decisions (optimistic scenario and pessimistic scenario), it was found that subjects did not have an invariant relative risk attitude -- in other words, the attitude toward risk of subjects changed as a function of the decision situation. Specifically, subjects (on the average) were more relatively risk-seeking for the pessimistic decision situation as compared to the optimistic decision situation (Krzysztofowicz, 1983a).

On the basis of these results, Krzysztofowicz (1983a) concluded that value functions and utility functions were two different constructs not only theoretically, but also behaviorally. In this regard, he agreed with Dyer and Sarin (1979a) that value functions encode only the strength of preference, while utility functions encode the strength of preference as well as attitude toward risk. Thus, this study provided further empirical support to the fundamental theoretical relationships established by Barron, von Winterfeldt, and Fischer (1984). Moreover, Krzysztofowicz (1983a) concluded that the concept of relative risk attitude had empirical meaning and that it changed from one decision situation to another. This conclusion
essentially supports the postulate of prospect theory which holds that individuals are generally risk-seeking in the domain of losses and risk-averse in the domain of gains (see Kahneman and Tversky, 1979).

A point needs to be noted about Krzysztofowicz's (1983a) finding that the empirical distribution of the relative risk attitude parameter is normally distributed, which in turn suggests that subjects (as a group) do not have any tendency toward one type of relative risk attitude. This suggests that, in the aggregate, there is no need to draw a distinction between value and utility functions. It should be noted that Krzysztofowicz (1983a) was not the only decision theoretic researcher to come to the conclusion that the distinction between values and utilities may be spurious. Indeed, no less renowned researchers than von Winterfeldt and Edwards (1986) on the basis of their vast collective experience in decision making reached the same conclusion.

Von Winterfeldt and Edwards (1986) stated that the distinction between utility-based and value-based preferences was spurious on account of the following arguments: (a) in practice, the conditions of certainty are never satisfied (i.e., there is no such thing as a “true” value-based function), (b) risk aversion can frequently be explained by marginally decreasing value-based functions and/or regret attributes of a value-based function, (c) in the case of repetitive choices, risk aversion will be eliminated, and an argument can be made that all choices are repetitive, and (d) error and method variance within value and utility measurement procedures obscure any differences in practice which may be present in theory.
The remarkable fact about the similar conclusions reached by Krzysztofowicz (1983a) and von Winterfeldt and Edwards (1986) is that they seemed to be based on different sets of evidence and approached the issue from different levels of aggregation. Specifically, Krzysztofowicz's (1983a) conclusion was based on the findings of his one study (which pooled the data from four different studies conducted over collective period of 30 years), while von Winterfeldt and Edwards (1986) conclusion was more on the basis of incremental evidence collected -- strangely enough -- over 30 years of experience in the theory and practice of decision making. Especially notable in this regard was the fact that Edwards was trained in Bayesian decision theory, which makes extensive use of risk in its treatment of decision making (see for e.g., Edwards, Lindman, and Savage, 1963). Further, von Winterfeldt was the one who was instrumental in establishing theoretical closed-form functional relationships between value functions and utility functions (Barron, von Winterfeldt, and Fischer, 1984; von Winterfeldt, 1979).

Moreover, these different researchers approached the value versus utility distinction from two different levels of aggregation. In this regard, Krzysztofowicz (1983a) was mainly concerned with the value versus utility distinction at the aggregate level, maintaining that although there may be a distinction at the individual level (indeed, he concluded on the basis of his analysis that there was such a distinction at the individual level), the differences cancel out on aggregation. Von Winterfeldt and Edwards (1986), on the other hand, have taken issue with the value versus utility distinction at the individual level, arguing that there is no practical difference in
individual value functions and utility functions. Along these lines, it should also be noted that the notion of an identity relationship between value and utility functions (i.e., $u = v$) is a hallmark of the French school of thought in utility measurement, which has often been in rather acrimonious debate with the American school of thought on utility measurement over several issues, including the distinction between value and utility (see Allais, 1979 for a description of the viewpoints of the two schools of thought).

Given this debate, the need for differentiating between value functions and utility functions should be examined on a case-by-case basis. For the purpose of this dissertation, the context of measurement is health status preferences. The few studies measuring both value-based and utility-based preferences conducted in the context of health status preference measurement have not supported Krzysztofowicz's (1983a) von Winterfeldt and Edwards' (1986) argument (see, for e.g., Read, Quinn, Berwick, et al., 1984; Torrance, Zhang, Feeny, et al., 1992; Wolfson, Sinclair, Bombardier, et al., 1982). All these studies found significant differences in value-based and utility-based preference functions. Further, some have also empirically determined functional relationships between the two preference measures. An especially noteworthy point about these studies is that they used different measurement strategies to measure health state preferences, including holistic preference measurement and the conditional utility-based function procedure. Further, these studies were conducted with a wide variety of subjects, ranging from ambulatory to hospitalized patients to the general population. Moreover, these studies are not confined to a particular geographic location. This suggests that at least for the case of measuring
health status preferences, the value versus utility distinction should be recognized -- at least until evidence to the contrary can be presented. In the absence of such evidence, it is necessary to be guided by the existing published literature on health status preference measurement, which has found the distinction between value and utility to be empirically meaningful. It is this body of literature which is now reviewed.

It was noted at the beginning of this section that the issue of comparing value and utility functions had been addressed in the health status preference measurement literature. Given this observation, it is necessary to review the relevant literature in this regard, especially since one of the limitations of the main study comparing value and utility functions (i.e., Fischer, 1976, 1977) was that one could not generalize from its limited context to other contexts. The issue of comparing preference measurement methodologies has been an integral part of health status preference measurement ever since the notion of measuring preferences for health states was introduced into the literature, i.e., in the early 1970s.

It is quite fitting that one of the earliest published comparisons between value- and utility-based preferences as applied to the measurement of health status preferences was by George Torrance (1976a), who pioneered the measurement of utility-based preferences in the field of health status measurement. Torrance (1976a) empirically compared the standard gamble (SG), time tradeoff (TTO), and rating scale (RS) scaling methods in terms of feasibility, reliability, validity, and comparability. All three scaling methods were implemented as part of a holistic preference measurement strategy.
Given the focus of this literature review, the review shall concentrate mainly on the results of the comparability between the scaling methods. However, the results of the reliability and validity of the different scaling methods will also be briefly summarized. The results pertaining to the feasibility of the different scaling methods will be discussed in a subsequent section of this chapter, when the literature on respondent evaluations of preference measurement methodologies is reviewed.

The linear correlations at the individual-level, as measured by Pearson correlations, between the three scaling methods were 0.65 (SG and TTO), 0.40 (RS and TTO), and 0.36 (RS and SG). Although Torrance (1976a) interpreted these as evidence against the validity of the rating scale method, this conclusion can be contested on account of the fact that it is based on the premise that the standard gamble is the "gold standard" and the other scaling methods must show a high degree of linear convergence with the standard gamble in order to be considered valid. The arguments against using the standard gamble as the "gold standard" have already been presented, and therefore will not be repeated here. Further, these results of Torrance (1976a) have not been replicated in further studies. Indeed, other studies comparing rating scale and standard gamble scaling methods have found high linear correlations between the two scaling methods. Kaplan, Bush, and Berry (1979) argued that the low correlation found between the RS and SG scores could be traced to the use of a narrow range of data points and to the subjects' poor comprehension of the RS task and health states. At the aggregate level, the linear correlations between the SG and the TTO scaling methods was 0.98, while that between the TTO and RS scaling
methods was 0.89. No aggregate-level coefficient was reported for the linear correlation between the SG and RS scaling methods.

The essence of the comparability analysis between the three measures was to test whether the three scaling methods gave the same preference scores for holistic health states, and if not, were score related in some systematic way (i.e., by some functional relationship) so that conversion curves could be constructed. In order to answer these questions, Torrance (1976a) conducted both aggregate-level as well as individual-level analyses. It is unfortunate that Torrance (1976a) reported only the results of the comparison between the SG and TTO and the TTO and RS methods, but not the SG and RS methods (which are of primary interest for the purpose of this dissertation). Still, the comparability between the standard gamble and rating scale methods can be inferred from the comparison between the other two pairs of methods, as will be shown below.

Regarding the comparability between the SG and the TTO scaling methods, there were no significant differences between the preference scores obtained as a result of the two methods (as indicated by a paired t-test) at both, the aggregate as well as the individual level. At the aggregate level, the functional form of the relationship between the SG and TTO scaling methods was found to be linear (i.e., SG = TTO, $r^2=0.91$, $p<0.05$). Although the same relationship was expected to hold at the individual level, it was not supported by the data -- the relative imprecision of the scaling methods at the individual level and the consequent “noise” in the individual-level data were attributed for this observation (Torrance, 1976a).
Regarding the comparability between the RS and the TTO scaling methods, there were significant differences between the preference scores obtained as a result of the two methods (as indicated by a paired t-test, p<0.05) at both, the aggregate as well as the individual level. At the aggregate level, the functional form of the relationship between the RS and TTO scaling methods was found to be best estimated by a power function (i.e., $RS = 1 - (1-TTO)^{0.80}$, $r^2=0.80$, p<0.05). It should be noted that in the case of the functional relationship between the RS and the TTO, two different types of functions fit equally well -- a logarithmic function and a power function. Torrance (1976a) decided to select the power function for two reasons: (a) it was consistent with research in psychophysical scaling (see, for e.g., Stevens, 1968) and (b) it had the useful property that it could be easily modified to satisfy the health scale model [i.e., pass through the points (0,0) and (1,1)], and, when it was so modified it took on an extremely simple form. For the same reasons mentioned above, a corresponding functional relationship could not be supported by the data (Torrance, 1976a).

Although Torrance (1976a) did not report the results of the comparability between the SG and the RS scaling methods, he did note in passing that the holistic preference scores obtained as a result of the two scaling methods were significantly different (as indicated by a paired t-test, p<0.05). Further, even though the results of the functional relationship estimation between the RS and SG were not reported initially (i.e., in Torrance, 1976a), they were reported in a later paper (see Torrance, Boyle, and Horwood, 1982). It was found that the relationship between the SG and
the RS at the aggregate level could best be estimated by a power function \[ SG = 1 - (1-RS)^{1/n} \], no \( r^2 \) or \( p \)-values reported.

As noted earlier, the significance of study cannot be overstated. As the first published empirical comparison of preference measurement methods in the context of health status preference measurement, it served as an exemplar for several subsequent studies in health status preference measurement. Indeed a similar methodology (in terms of comparing different scaling methods) was followed in subsequent studies by the McMaster group of researchers, even after they began using the conditional utility function-based procedure as a measurement strategy for health status preference measurement. Further, given its methodological focus, its limited generalizability (it used only residents of the city of Hamilton, Canada for the RS and TTO; further, it used only educated, non-elderly subjects for the SG) is not a major problem.

The only potential drawback was the fact that the RS scaling method always preceded the TTO or the SG scaling methods. Therefore, the "negative" results obtained for the RS scaling methods (also see the subsequent section of this chapter, when the respondent evaluations for this study are discussed) could have been due to the fact that it was the first of the scaling methods that respondents performed. In this regard, respondents must have used it as a "warm-up" exercise -- this in turn could have been detrimental for its validity and reliability. However, in defense of this practice, it should be noted that this procedure is recommended for the measurement of utility, i.e., order preferences under certainty before measuring cardinal utilities for alternatives. The theoretical rationale for this
is given in Luce and Raiffa (1957) and an empirical demonstration of the
detrimental effects on measured utilities if this procedure is not followed is
provided by Laskey and Fischer (1987). Therefore, this might represent a
true drawback of this study, since it is a necessary condition to obtain
axiomatically valid utilities.

As noted, the McMaster group of health status researchers adopted a
similar methodology (in terms of comparing different scaling methods) in its
future work. The work on the Mark I version of the McMaster Health
Utilities Index (Torrance, Boyle, and Horwood, 1982) will not be reviewed
here because it did not measure utilities per se, since it did not include the
standard gamble as a scaling method. Therefore, this particular work
(although highly significant in the development of health status preference
measurement methodology) is of little use for this literature review, which is
focused on published studies which have modeled the relationship between
value and utility functions. Second, the research on the Mark I version of the
McMaster Health Utilities Index (Torrance, Boyle, and Horwood, 1982) did
not independently estimate a functional relationship between the scaling
methods it did use (i.e., RS and TTO), but used the Torrance (1976a)
relationship to compare values to utilities. Therefore, this research will not
be reviewed here.

However, the work on the Mark II version of the McMaster Health
Utilities Index (Torrance, Zhang, Feeny, et al., 1982) is highly pertinent to
this literature review, since it used the RS and SG scaling methods -- two of
the scaling methods which were used in this dissertation. In this study, they
used the conditional utility function-based procedure as the measurement
strategy, albeit with some modifications from the method suggested by Keeney and Raiffa (1976). These modifications will not be discussed in this section, since they do not pertain to the discussion of value-based versus utility-based preference scores. However, since much of the modifications were adopted for the methodology used in this dissertation, they will be discussed in terms of their usage in the present study in Chapter V.

Each respondent provided responses using the RS as well as the SG scaling methods. However, the RS scaling method was used to measure preferences for all 14 health states of interest in the study as well as for the individual attribute value functions; on the other hand, the SG scaling method was used only to measure preferences for four health states, and none of the individual attributes. This was because of the amount of time it took to complete the SG task. Like most other empirical studies which measured both values and utilities, it was found that the utility-based preference scores were much greater than value-based preference scores for the same health states (Torrance, Zhang, Feeny, et al., 1992).

In order to construct a utility function, a two-stage procedure was used, wherein a value function was first constructed using the RS scaling method and then this value function was converted into a utility function by applying the appropriate transformation. The particular form of this transformation was determined by an empirical analysis of the four health states for which both RS values and SG utilities were measured. Specifically, the task was to fit a curve through the four measured points such that it also went through the points (0,0) and (1,1). As explained above, this is necessary to satisfy the health scale model. Such a curve would then represent the
relative risk attitude of the average person (all analyses were reported at the aggregate level). Based on theoretical considerations (see Stevens, 1968) and past research (Torrance, 1976a), a power curve was fit to these data. At the aggregate level, the function was \( SG = 1 - (1-RS)^{2m} \). The fit yielded an \( R^2 \) of 0.70. This function was used to transform the RS values to utilities and then construct a multiattribute utility function for the average person. Cross-validation with a holdout sample yielded satisfactory results (Torrance, Zhang, Feeny, et al., 1992).

Although this brief discussion of the Mark II version of the McMaster Health Utilities Index does not do justice to the methodology used in the study, it does suffice for the purpose of this section. As noted earlier, more details about the methodology will be provided when the methodology of this dissertation is discussed in Chapter V.

Apart from the McMaster group of researchers, there have been a few isolated researchers who have compared values and utilities. Most of this research seems to have been prompted more from an application-oriented viewpoint, in that the interest was in checking for the differences between values and utilities as opposed to understanding why there should be a difference in the first place. Consequently, very few of these other published studies have estimated functional relationships between value and utility functions. Still, it is instructive to review these other studies on account of the fact that they do provide some measures of correlation between the two measures. Further, an idea of the magnitude of the difference between values and utilities in different health-care contexts can be obtained from these studies.
One of the earlier such studies was reported by Wolfson, Sinclair, Bombardier, et al., (1982). As part of their development of a functional status index for stroke patients, they compared the SG, TTO, and RS scaling methods for the measurement of health status preferences. No information was provided about the order in which respondents performed the scaling methods. Although they used four different groups of respondents, there were no practically significant differences between the respondent groups; therefore, they pooled the results of the four groups for the preference measurement analysis. It is important to comment on the rationale for pooling, i.e., no practically significant differences. Out of the 840 pairwise comparisons they computed, only 15 were statistically significant at the 0.05 level. On the basis of the observed magnitude of differences as well as the possibility that respondents may not have followed instructions, they concluded that the differences that were observed were not of practical significance (Wolfson, Sinclair, Bombardier, et al., 1982). There is, of course, another possible explanation for lack of practical significance -- i.e., lack of control of familywise error. In this regard, they did not control for the alpha-slippage which occurs in the analysis of multiple comparisons. Since they conducted 840 pairwise comparisons, the probability of committing at least one Type I error is virtually certain. Therefore, it is possible that some of the statistically significant differences they observed were "false positives." Either way, their decision to pool the data across respondent groups would be supported.

Regarding comparisons of the absolute scores, these researchers conducted a one-way analysis of variance with pairwise comparisons. As
before, they did not control for familywise error. It was found that the SG scores were significantly higher than the scores obtained via the TTO and RS scaling methods. Specifically, the SG scores were significantly higher than the RS scores for almost all levels of the functional status dimensions, while the SG scores were significantly higher than the TTO scores for fewer levels of the functional status dimensions. In contrast to the results of the Torrance (1976a) study, it was found that there were very few instances in which the RS scores and the TTO score were significantly different. Overall, there was more similarity between the scores obtained via the RS and TTO scaling methods than between either of these two scaling methods and the SG scaling method (Wolfson, Sinclair, Bombardier, et al., 1982).

Although these researchers did not report correlation coefficients between the different scaling methods, they did fit linear functional relationships between the scores obtained via the different scaling methods using regression. It was found that linear functions provided reasonably good fits to these data in all cases. For instance, the linear relationship between the SG and RS scores was given by: 
\[ SG = 0.32 + 0.88(RS); \text{ adjusted } R^2 = 0.76. \]
For the linear relationship between the SG and TTO scores, the regression equation was: 
\[ SG = 0.30 + 0.69(TTO); \text{ adjusted } R^2 = 0.84. \]
Finally, the linear relationship between the TTO and RS scores was given by: 
\[ TTO = 0.03 + 1.26(RS); \text{ adjusted } R^2 = 0.89 \] (Wolfson, Sinclair, Bombardier, et al., 1982).

This study offered some interesting contrasts to the Torrance (1976a) study. The most significant of these was the fact that the SG and TTO scores were significantly different in this study, as opposed to the Torrance (1976a)
study which found the two to be similar. Related to this issue was the closer correspondence between the TTO and RS scores obtained in this study as opposed to the Torrance (1976a) study. However, what is important for the purpose of this particular review is the fact that the SG and RS scores were found to be significantly different from each other in both studies -- specifically, in both studies, the SG scores were significantly higher than the RS scores. This observation is consistent with the behavioral interpretation of the two scores provided earlier in this section, i.e., the SG scores (i.e., utilities) encode strength of preference as well as attitude toward risk, while the RS scores (i.e., values) encode only strength of preference.

It is, however, unclear why Wolfson, Sinclair, Bombardier, et al., (1982) decided to fit a linear relationship between the SG and RS scores, when the theoretical literature would suggest a nonlinear relationship between values and utilities. Although the linear relationship they fit provided a very good account of their data as measured by adjusted $R^2$ values, Anderson and Shanteau's (1977) warning about “weak inferences with linear models” comes to mind -- i.e., linear models are remarkably robust to proven nonlinearities in the data, and consequently do a very good job of approximating inherently nonlinear relationships. The key word here is “approximating.” From a theoretical viewpoint, it is more instructive to fully describe the relationship, i.e., explain the relationship. Since the theoretical literature supports a nonlinear relationship between the two measures, it would be instructive to fit a nonlinear relationship to these data.

Although Wolfson, Sinclair, and Bombardier (1982) did not fit such a relationship to their data, Torrance, Boyle, and Horwood (1982) reported
that fitting a power curve to the Wolfson, Sinclair, and Bombardier (1982) data yielded an equation of: \( SG = 1 - (1 - RS)^{2.16} \). No adjusted \( R^2 \) values were provided for the fit of this nonlinear relationship to the data. Comparing this particular equation to the equation obtained by Torrance (1976a), it can be seen that both equations support the same general shape of the relationship between RS values and SG utilities, but not the same specific parameter value. Since the two equations support the same general shape of the relationship between RS values and SG scores, they both support the behavioral interpretation of the difference between values and utilities provided earlier. However, since they yield two different parameter values, it is possible that the specific attitude toward risk of the two samples is different. In this regard, since the plots of both data yielded a concave utility function, it stands to reason that both sets of respondents were risk-averse. If so, the different parameter values are only indicative of differing degrees of risk-averseness of the two sets of respondents. Specifically, since the parameter value of Torrance (1976a) was smaller in magnitude as compared to the parameter value of Wolfson, Sinclair, and Bombardier (1982) -- i.e., 1.61 versus 2.16 -- Torrance’s (1976a) respondents were less risk-averse than Wolfson, Sinclair, and Bombardier’s (1982) respondents. This follows from the fact that if respondents were risk-neutral, the parameter value would be equal to 1.0 -- i.e., \( SG = RS \). Therefore, lower (but positive) parameter values would indicate lesser degrees of risk-averseness.

Read, Quinn, Berwick, et al., (1984) also compared these same three scaling methods (i.e., SG, TTO, and RS) using a within-subjects design and testing for significant differences between the three scaling methods using
analysis of variance. Subjects were made to evaluate univariate (i.e., single attribute) as well as multivariate (i.e., multiattribute) alternatives. Single attribute alternatives were evaluated by all three scaling methods; multiattribute alternatives were evaluated only using the RS and SG scaling methods. Regarding order of presentation, none was specifically reported for the single attribute evaluations, while the RS scaling method always preceded the SG scaling method for the multiattribute evaluations. For all respondents, multiattribute evaluations preceded the single attribute evaluations (Read, Quinn, Berwick, et al., 1984).

Linear correlations, as measured by the Pearson correlation coefficient, between the three scaling methods for the single attribute evaluations were found to be reasonably high and significantly greater than zero (p<0.01) -- the correlation between the RS and SG scores was 0.63, the correlation between the RS and TTO scores was 0.65, and the correlation between the SG and TTO scores was also 0.65. A one-way analysis of variance on the single attribute evaluations showed that the SG scores were significantly higher than the TTO scores, which in turn were significantly higher than the RS scores. Direct comparisons between the SG and RS scores also showed significant differences between them (Read, Quinn, Berwick, et al., 1984). However, no mention was made in the article about efforts to control the familywise error; hence the possibility of a "false positive" test on some of the pairwise comparisons cannot be ruled out. However, the comparisons of interest to this literature review (i.e., SG and RS) would probably still be significantly different simply because of the magnitude of the difference between these scores.
Regarding the multiattribute evaluations, it is interesting to note that not all subjects produced identical preference orders for the alternatives. Since Kendall's Tau coefficient values were not reported, one cannot estimate the extent of agreement between the two scaling methods (i.e., SG and RS) in terms of preferential ordering. The linear correlation, as measured by the Pearsonian correlation coefficient, between the SG and RS scores for the multiattribute alternatives was moderately high -- 0.56 (p<0.01). The vast majority of subjects had higher SG scores than RS scores for the multiattribute alternatives. In all cases, there were significant differences between the SG and RS scores. However, it should be noted that the variability of responses among subjects was considerably greater for the SG scores as opposed to the RS scores. Further, no functional relationships were estimated between the different scores (Read, Quinn, Berwick, et al., 1984).

This study also found an interesting result in terms of the effect of the specifics of the medical condition under evaluation on the multiattribute preference scores obtained by different scaling methods. Examining the effects of severity of the condition under evaluation (i.e., moderate versus severe angina) and the duration of survival (i.e., five versus ten years) on the preference scores obtained via the scaling methods, these researchers found a significant interaction between severity of condition and duration of survival only for the RS scores -- in other words, only for the RS scores did the value of five additional years of life depend on the severity of angina. However, the practical significance of this finding is questionable, since it was shown to account for only 2% of the variability in the RS scores. Still, the
possibility that the difference between RS and SG scores is not only due to pure attitude toward risk does warrant further explanation. For instance, in this particular situation Read, Quinn, Berwick, et al., (1984) conjectured that an individual's willingness to engage in gambling might depend on the specifics of the alternatives which make up the gamble.

In providing a possible explanation for the divergent results obtained via the different scaling methods, Read, Quinn, Berwick, et al., (1984) conjectured that different scaling methods induce different processes of evaluation and that the components of such processes include: (a) attention to and perception of stimuli in evaluation tasks, (b) cognitive activity, such as recalling and taking into account of past events and life goals, and selecting reference points against which alternatives are to be evaluated against, and (c) emotional reactions to past, present, and future health states (Read, Quinn, Berwick, et al., 1984). In other words, these researchers supported a constructive view of decision making, in which the particular response given to a preference measurement task is contingent upon the specifics of the situation. This is in keeping with the literature on response mode effects and developments in behavioral decision theory in general, as reviewed later on in this section.

The major contribution of this study is not so much in terms of the fact that it demonstrated that preference scores obtained via different scaling methods were different as much as it went one step further and attempted to show the practical significance of such differences. In this regard, they noted that the practical significance of differences among the scaling methods depended upon whether or not they would lead to different medical choices.
To investigate this, these researchers used the SG and RS scores as inputs into a simplified decision tree for the medical problem investigated in the study. It was found that compared to the RS scaling method, the relatively higher scores provided by the SG scaling method changed the recommended course of action for 60% of the respondents. Although this demonstration may be criticized for using a simplified, didactic decision model, its potential significance to public policy and clinical decision making is clear -- that differences in preference measurement methodologies can have meaningful and clinically relevant implications for health-care decision making.

Another comparison among the SG, TTO, and RS scaling methods was reported by Richardson, Hall, and Salkeld (1990). In this comparison, it was found that the RS scores were the highest of the three scores for three of the four health states evaluated (only four health states were evaluated), followed by the SG scores, and finally by the TTO scores. In all cases, the RS scores had smaller standard deviations as compared to the SG and TTO scores. However, in terms of statistical significance, the only significant difference between the RS and SG scores was for the one in which the SG score was higher than the RS score (p<0.01). Further, there were no statistically significant differences between the SG and TTO scores, and only one statistically significant difference between the RS and TTO scores (Richardson, Hall, and Salkeld, 1990).

These results indicate that all three scaling methods give equivalent results. It should be noted that this is the only published study which has found such close correspondence between the RS and SG scores, and further
that the RS scores were -- for the most part -- higher than the SG scores. This latter finding is not consistent with the theoretical literature on value versus utility that was reviewed earlier on in this section. It should be noted that these statistical tests between the scores obtained via the three scaling methods were on the basis of standard t-tests. The use of standard t-tests for such testing can be questioned on two accounts: (a) since a within-subjects design was used instead of a between subjects design, a testing procedure which takes the correlation between the two measures should be used and (b) since there were three different groups, the number of t-tests conducted would be quite excessive. On the basis of these two arguments, it might have been more appropriate to use a repeated measures analysis of variance to test for significant differences between the scores obtained via the three scaling methods. The fact that the researchers did not control for family wise error is also noteworthy, since it does not rule out the possibility that at least one of the statistically significant differences observed was a "false positive."

Pearson correlation coefficients between the three measures -- taken two at a time -- were generally lower between the RS and SG scores as compared to the other sets of scores. The highest linear correlations were observed between the SG and TTO scores. At the aggregate level, the correlation between the RS and SG scores was 0.52, while the correlation between the RS and TTO as well as between SG and TTO was 0.70 (Richardson, Hall, and Salkeld, 1990).

Linear as well as power functional relationships were fit to the data. In the case of the RS and SG scores, the linear regression equation was: \( RS = 0.38 + 0.32(SG) \), adjusted \( R^2 = 0.27 \); while the power curve equation was: RS
519

= 1 - (1-SG)^{0.46}, \text{Adjusted } R^2 = 0.29. In the case of the RS and TTO scores, the linear regression equation was: \text{RS} = 0.31 + 0.46(\text{TTO}), \text{Adjusted } R^2 = 0.49; while the power curve equation was: \text{RS} = 1 - (1-\text{TTO})^{0.34}, \text{Adjusted } R^2 = 0.13. Finally, for the SG and TTO scores, the linear regression equation was: \text{TTO} = 0.18 + 0.68(\text{SG}), \text{Adjusted } R^2 = 0.49; no power curve equation was provided for the SG and TTO scores (Richardson, Hall, and Salkeld, 1990).

Like the previous results, the results of the functional relationships between the different measures are not in agreement with both, theoretical considerations and previous empirical research. In terms of theoretical considerations, the review provided earlier suggests that a nonlinear relationship would be expected between the RS and SG scores. The empirical evidence provided by Torrance (1976a) supports this premise. However, the fact that the \text{Adjusted } R^2 values in this study were higher for the linear relationship as opposed to the power function relationship goes against this premise.

The results of this study run counter to the results of most other studies published in the literature. While it might be instructive to conjecture why that might be so, in the absence of the actual instructions given to the respondents such conjectures are of limited use. Still, it is possible that any one or some combination of the following reasons might apply -- interviewer effects, unclear instructions, respondent confusion, or even the fact that the other studies were all "wrong." It should also be noted that the researchers did not control for an order effect -- respondents always did the RS scaling method first, followed by the TTO and finally the SG scaling method.

Hornberger, Redelmeier, and Peterson (1992) also compared these three scaling methods along with other health status measures in an attempt
to assess the effect of variability in methods to measure health status on the results of cost-effectiveness analyses using these methods. Keeping in mind the focus of this review, only those results pertaining to the comparison between the scaling methods will be discussed here. Two measurements were taken of each patient, separated by nine months. For all measures of health status used in the study except the RS, scores increased over the nine months. Regarding the mean scores obtained by each scaling method, the TTO scores were the highest, followed for the most part by the RS scores and finally the SG scores. However, none of the differences were statistically significant, as indicated by a paired t-test adjusting for multiple comparisons with the Bonferroni correction (Hornberger, Redelmeier, and Peterson, 1992).

Two types of correlations between the different measures were reported — Pearsonian correlations and Spearman rank-order correlations. Both correlation coefficients were quite low. Although exact values were not reported, it was stated that the Pearsonian correlations ranged between 0.05 and 0.51, with the majority being below 0.40. The values of Spearman correlations were reported as: 0.31 (SG and TTO), 0.14 (SG and RS), and 0.21 (RS and TTO). No attempt was made to derive functional relationships between the preference measures. The effects of the different measures were reflected in the cost-utility ratios that were computed — the ratio using the TTO scores was the lowest of the three preference measures ($39,338), while that using the SG scores was the highest ($45,254). The cost-utility ratio using the RS scores was $41,095 (Hornberger, Redelmeier, and Peterson, 1992).
The notable point about this study is its attempt to demonstrate the effect of differences in measures on the final results of analyses using these measures. Although similar in nature to what Read, Quinn, Berwick, et al. (1984) demonstrated, this study extended the scope of that analysis since used a decision tree based on actual clinical practice, as opposed to a contrived one for demonstration purposes, as used by Read, Quinn, Berwick, et al. (1984). The implications of this study, however, are similar to those of that study -- i.e., different measures can lead to different conclusions, and could thus affect health-care decision making.

A few potential drawbacks of this study should be noted. First, no attempt was made to control for an ordering effect -- for the preference scaling methods the SG preceded the TTO which was always followed by the RS scaling method. Second, and more seriously, only one health state was used in the analysis -- the respondent’s current health state. No attempt was made to operationalize the current health state. Therefore, it is possible that respondents were evaluating different states of health to begin with. Although all respondents were receiving the same treatment (i.e., hemodialysis), it is possible that their health status was different. Further, demographic and clinical information about the study participants was not provided in the article, so one cannot estimate how similar they were on these criteria. Therefore, the results of this study need to be treated with caution, simply because if patients are evaluating different states of health there is little to be gained in aggregating preference measures for these different states of health for analysis.
Nease, Kneeland, O'Connor et al., (1995) reported a comparison of the SG, TTO, and RS scaling methods as part of their investigation on the possible importance of incorporating patient preferences into clinical practice guidelines. Although the emphasis of their study was more application-oriented (i.e., on evaluating the role of preference assessment in clinical practice guidelines), they did report the following Spearman rank-order correlations between the three scaling methods -- 0.40 (RS and TTO; p<0.001), 0.31 (RS and SG; p<0.001), and 0.50 (SG and TTO; p<0.001). These correlations indicate that the preferential ordering of health states was different between different preference measures, although there was some agreement in rank-order preferences among the scaling methods. Further analysis revealed that any variation in agreement was random; no systematic bias was discovered (Nease, Kneeland, O'Connor et al., 1995).

It is interesting to note that these researchers treated the preference measures as ordinal in level of measurement (they used nonparametric statistics for their analyses), when the property of interval level of measurement is necessary for the use of these measures in pharmacoeconomic analyses. Further, they did not specify in what order subjects performed the preference measurement scaling methods. Finally, it should be noted that although the review of this particular study has been brief keeping in mind the purpose of this literature review, the results of this study have great significance for health-care policy.

Apart from these studies which compared the SG, TTO, and RS scaling methods, there have been a few published studies which have compared only the SG and RS scaling methods. In the process of
investigating whether health status preference scores were affected by variations in the descriptive scenario presented to subjects, by subjects' prior experience with health status scaling methods, and by interviewers who conducted the preference measurement session, Llewellyn-Thomas, Sutherland, Tibshirani, et al. (1984) reported that the mean preference scores obtained using the SG scaling method were substantially different and systematically higher than those obtained using the RS scaling method. This study did control for presentation of scaling method order effects and found that order of presentation of scaling method interacted with descriptive scenario presented to subjects. Specifically, for all cases when scenarios were presented in a standardized point-form format (i.e., bullet-form) and for cases when the RS scaling method preceded the SG scaling method when scenarios were presented in a narrative paragraph format the SG scores were found to be significantly higher than the RS scores. However, in cases when the SG scaling method preceded the RS scaling method for when scenarios were presented in a narrative paragraph format, there were no significant differences observed between the two sets of scores. These results were explained by the researchers in terms of an anchoring and adjustment heuristic, wherein subjects anchor their responses on the basis of the complex SG task and therefore may tend to use this score when subsequently faced with a RS task on the same scenario. This effect could have been compounded by the fact that subjects did not want to bother reading the narrative paragraph, which usually is more intensive than reading a bullet-point format (Llewellyn-Thomas, Sutherland, Tibshirani, et al., 1984).
This study did not discuss any other comparisons between the two scaling methods (e.g., correlations, functional relationships) on account of the fact that this was not the primary focus of their study. However, its contribution to the empirical literature is still significant in that it demonstrated that order effects may be operative in some instances of health status preference measurement.

As part of a study comparing the correspondence of patients’ anticipated preferences for treatment to actual preferences once they have experienced the treatment in question, O’Connor, Boyd, Warde, et al. (1987) reported a comparison between risky and riskless scaling methods. Subjects’ preferences were measured using five different scaling methods: (a) a regular rating scale, (b) a choice between two treatments with 100% chance of survival but certainty of some side effects, (c) a choice between two treatments with 60% chance of survival but certainty of some side effects, (d) a choice between two treatments with 60% chance of survival but a quantified uncertainty (i.e., risk) of some side effects, (e) a choice between two treatments with 100% chance of survival but a quantified uncertainty (i.e., risk) of some side effects (O’Connor, Boyd, Warde, et al., 1987). Therefore, measures (a), (b), and (c) were values, while measures (d) and (e) were utilities.

Significant differences (p<0.001) were only observed between measures (a) and (d). Pearsonian correlations between all measures were generally high -- above 0.8 -- with the exception of measure (a), which was generally uncorrelated with all the other preference measures (O’Connor, Boyd, Warde, et al., 1987). No other comparative results (e.g., functional
relationships, preferential orderings) were reported among the preference measures. Caution needs to be exercised before comparing these results to the results of other studies comparing health status preference measures. This is because the measure (a) is the only measure which is truly comparable with measures used in other studies. The risky preference measures used in this study were not standard gambles since they did not have a sure thing as an option.

Further, and more importantly for the purpose of this dissertation, each of the other four measures -- i.e., (b) through (e) -- were choice-based measures of preference. Since none of the other empirical studies reviewed so far used choice-based measures of preference, valid comparisons between such measures cannot be made. On the judgment versus choice issue in this study, however, it is interesting to note that there were no significant differences observed between measures (b) and (d) and between measures (c) and (e). Since these are riskless versus risky choice-based measures, this study found that the value versus utility distinction was spurious when both value- and utility-based measures were inferred from choices. Moreover, the only significant difference found in this study was between a measure of riskless judgments and risky choices -- measures (a) and (d). Finally, it should be noted that all the choice-based measures [i.e., (b) through (e)] were highly correlated with one another, but none of the choice-based measures were significantly correlated with the judgment-based measure [i.e., (a)]. Therefore, this study lends some support to the judgment versus choice distinction -- although the authors of this study did not seem to recognize this as a potential explanation for their results.
As part of their cost-benefit analysis of hepatitis-B vaccination, Jonsson, Horisberger, Brugera, et al. (1991) measured health status preferences for two states of health using the SG, TTO, and RS scaling methods. The results of this small analysis were similar to those of Torrance (1976a) -- the RS scores were substantially less than the SG and TTO scores, which in turn were similar to each other. This study did not report any other comparative analyses among the three scaling methods (e.g., correlations, functional relationships). However, in defense of this study, it should be pointed out that it was not designed as a preference measurement comparative study. Indeed, health status preferences were only actually measured in this study because the particular health states of relevance to the cost-benefit analysis did not fit with the existing health status valuation index they proposed using (i.e., the Rosser/Kind index) -- therefore the only reason they actually measured health status preferences was to get preference scores for the particular states of health they were interested in for their analysis. This suggests that if an "off-the-shelf" health status preference measure were available, they would have used it. Hence, given the low priority of actually measuring health status preferences in this study, it is understandable that no other comparative analyses were reported.

In a study conducted to determine the extent to which patients' subjective evaluations of their current health state are determined by their expectations, their actual health, and how they compare themselves with others, Llewellyn-Thomas, Theil, and McGreal (1992) measured health status preferences using the SG and RS scaling methods (as an operationalization of patients' subjective evaluations of their current health state). Although
their purpose was not to compare the SG and RS scores, these researchers did report the SG and RS scores for the two health states evaluated as part of the study. On the basis of these reports, the SG scores were consistently higher than the RS scores. These results are in general agreement with the published literature on this issue. Further, the SG scores had a smaller standard deviation compared to the RS scores. This finding is different from some of the published studies, but the magnitude of difference in this study was so small (about 0.02 units on a 0-1 scale) that it might not be practically significant. No other comparative results between the two scaling methods were reported (e.g., functional relationships, correlations). As for the Jonsson, Horisberger, Brugera, et al. (1991) study reviewed above, given the purpose of this study and the non-centrality of comparing preference measurement methodologies in this study, it is understandable that no other comparative analyses were reported.

Bass, Steinberg, Pitt, et al. (1994) conducted a comparative study of RS and SG scaling methods to investigate how patient preferences for treatment outcomes vary as a function of patient characteristics and scaling methods. Both these scaling methods were used as part of a holistic preference measurement strategy. Health states were described in a bullet-list format [probably in light of Llewellyn-Thomas, Sutherland, Tibshirani, et al.'s (1984) study]. In contrast with most other comparative studies, this study used a between-subjects design. The reason reported for using a between-subjects design was that pilot-testing showed that patients' concentration waned when both scaling methods were used in the same interview (Bass, Steinberg, Pitt, et al., 1994).
Consistent with other studies comparing the SG and RS scores, it was found that the SG scores were significantly higher than the RS scores. Further, the mean SG scores were highly correlated with the mean RS scores (Spearman rank-order correlation = 0.90). These researchers also fit a linear regression equation to the data by regressing RS scores on SG scores, yielding an adjusted $R^2$ of 0.80. No reason was given as to why a nonlinear functional relationship was not estimated. Finally, with respect to respondent characteristics such as age, gender, and race there were generally no significant differences observed between the preference scores obtained using the two scaling methods (Bass, Steinberg, Pitt, et al., 1994).

The results of this study are consistent with the other studies in this area, as reviewed in this section and therefore require no elaboration. It is, however, interesting to note that they treated the preference measures as ordinal in level of measurement (they used nonparametric statistics for their analyses), when the property of interval level of measurement is necessary for the use of these measures in pharmacoeconomic analyses.

The issue of comparing different preference measurement methodologies is an active field of research for the EuroQol group of researchers (Sintonen, 1993). These researchers have reported the results of a few comparative studies on preference measurement methodologies (see for e.g., Busschbach, Hessing, and de Charro, 1993a; Krabbe, Essink-Bot, and Bonsel, 1993; Nord 1992b) and are reportedly in the process of conducting several other such studies (see, for e.g., Busschbach, Bonsel, and de Charro, 1994; Sintonen, 1993). The one study comparing RS and SG scaling methods will be reviewed here. The other published studies only compared SG and
TTO scaling methods or RS and equivalence scaling methods; since they did not directly compare RS and SG scaling methods (which are of primary interest to this study), they will not be reviewed in this section.

The study in question (Busschbach, Hessing, and de Charro, 1993a) was an empirical comparison of the SG, TTO, RS and Rosser and Kind Index (RKI) -- which was scaled using magnitude estimation. The purpose of this study was to provide insight into the validity of the different measurement techniques on the basis of construct validity -- much along the lines of this dissertation. A within-subjects design was used, with the RS task always being performed first, while the order of the other three scaling methods was varied at random (Busschbach, Hessing, and de Charro, 1993a).

Although these researchers reported a positive correlation between the different scaling methods, no correlation coefficients were provided. Regarding differences in preference scores obtained using the different methods, it was found that the SG scores were the highest, followed by the TTO scores, and finally by the RS scores. The RS scores had the lowest amount of variability. These researchers did not attempt to fit any functional relationship between the preference measures, although a plot of SG utilities versus TTO values and RS values clearly indicated a concave utility function. The scores obtained with the RKI were "... rather deviant ..." (Busschbach, Hessing, and de Charro, 1993a p.51) from those obtained by the other scaling methods and further had the highest variability. On the basis of these results, it was concluded that the SG and TTO scaling methods had a "... relatively good ..." (Busschbach, Hessing, and de Charro, 1993a p.51) relationship, which could be explained on account of attitude toward risk
and time preferences for good health (i.e., discounting of health effects). It was further concluded that further research needs to be conducted on the RS in order to explain its lower values as compared to the other scaling methods. Finally, the RKI was considered to have the lowest construct validity (Busschbach, Hessing, and de Charro, 1993a).

These conclusions need to be re-examined in light of the above discussion on comparisons between different preference scaling methods. For instance, it is unclear as to why Busschbach, Hessing, and de Charro (1993a) explained the observed differences between the preference scores obtained by the SG and TTO scaling methods in terms of attitude toward risk (at least partially) and did not attribute the same reason to the difference between the SG and RS scores. After all, both the RS and TTO scale preferences under conditions of certainty -- in other words, both are values, not utilities. Therefore the relative risk attitude of respondents could also explain the difference between the SG and RS scores.

Further, it is also unclear from the descriptions provided in the paper whether the authors controlled for several methodological issues in health status preference measurement. In this regard, the labels used to anchor the end points of the SG and TTO scaling methods on the one hand and the RS scaling method on the other hand were different. Further, there is the question of the duration effect -- the SG and TTO scaling methods had subjects imagine they were in the state of health under evaluation for a period of 50 years. Although a particular duration was not specified for the RS scaling method, if the duration used in the actual EuroQol instrument was used then the effects of the duration effect need to be considered, since
the actual EuroQol instrument has subjects imagine they would be in the state of health under evaluation for a period of one year.

Finally, for the SG scaling method, the questions were posed in the context of an imaginary risky operation that could provide the necessary effects; however, for the TTO scaling method, the questions were posed in the context of an imaginary drug that could provide the necessary effects. No such corresponding contexts were provided for the RS and RKI methods. It is possible that the context could have affected the evaluation -- for instance, some people might have had bad experiences with operations and may consequently opt for the sure-thing in the SG without regard to the actual outcomes under evaluation. Since such effects were not controlled, it cannot be determined to what extent the resulting preference scores obtained were because of the health states under evaluation or because of the particular context under which the health states were evaluated -- or a combination of these two factors.

Summary

Summarizing the results of the comparative studies involving the SG, TTO, and RS scaling methods, it appears that the SG scores are significantly higher than the scores obtained by the RS scaling method [the study by Richardson, Hall, and Salkeld (1990) is a notable exception]. This is consistent with the theoretical premise of utility functions encoding strength of preferences as well as attitude toward risk versus value functions encoding only strength of preferences. The comparisons between the SG and TTO scaling methods have yielded mixed results -- some studies have found
the SG scores to be significantly higher than the TTO scores (for e.g., see Read, Quinn, Berwick et al., 1984; Wolfson, Sinclair, Bombardier, et al., 1982), while others found no significant differences between the two scores (for e.g., see Torrance, 1976a). Further, most studies have found high correlations between these measures, which can be taken as evidence for the fact that they are measuring the same construct, i.e., health status preferences. Finally, a power function provides a satisfactory description of the relationship between the RS and SG preference functions.

Taken as a whole, these studies support the following conclusions about the value versus utility distinction as it applies to health status preference measurement: (a) the theoretical differences between values and utilities are supported by empirical evidence from more than one study (b) this is true for both, the absolute magnitudes of the preference scores (where utilities are significantly higher than values) as well as functional relationships between the two preference functions and (c) both value-based and utility-based scores seem to measure the same construct, i.e., health status preferences. The remarkable fact about these conclusions is that they are based on data collected from a variety of subjects in varying health-care contexts. This would support the robustness of these conclusions.

Although there were a fair number of studies reviewed in this section, it is acknowledged that these are not all the studies which have compared different preference measurement methodologies. This is because, in keeping with the focus of this literature review, it was decided to focus only on studies which have compared the SG and RS scaling methods. Further, it
is also acknowledged that there are some studies published in the health-care literature which have measured health status preferences using both, the SG and the RS method which have not been reviewed here. This is because these studies have not directly compared the scores obtained by the SG and RS scaling methods (e.g., O'Brien, Elswood, and Calin, 1990).

Finally, it should be noted that although the emphasis in this literature review has been on determining relationships between different preference measurement methodologies yielding values and utilities for the purpose of validation of new methodologies to measure preferences, these relationships can be of interest for several other uses. First, by separating attitude toward risk from the strength of preference, one can gain additional insight into behavioral aspects of decision making under risk. Further, as noted by Dyer and Sarin (1982) and Krzysztofowicz (1983a), such a separation may facilitate comparison of interpersonal utilities -- an often-debated issue in welfare economics. In this regard, instead of comparing risky utility functions (which usually confound strength of preference and attitude toward risk), one can get more specific in the comparison in terms of strength of preference. Moreover, from an applied decision analysis point of view, it is instructive to investigate the degree of consistency of the functional relationship, i.e., across time, across decision makers, or across decision problems. This is because if there were some amount of consistency, one could apply a two-stage process to measure utilities -- first measure values and then convert these values into utilities using the functional relationship. Such a two-stage process would be quicker than estimating utilities directly, which may be associated with savings in resources. Indeed,
this was the very process adopted by the McMaster group of researchers in determining a utility function for the Mark II version of their Health Utility Index (Torrance, Zhang, Feeny, et al., 1992).

Judgmental Versus Choice Response Modes

It had been shown in Chapter II that the breakdown of the procedural invariance assumption is a major causative factor of the phenomenon of preference reversal. Given this, it is important to understand how exactly different response modes exert their effects on decision making. In order to do this, the literature on the effects of response mode needs to be reviewed. An interesting aspect of this literature is the apparent hesitance of researchers to acknowledge the very existence of such effects on decision making. Indeed, Slovic and Lichtenstein (1971) themselves in a major and often-cited review of research in behavioral decision making specifically noted: "... the distinction between judgments and decisions is a tenuous one and will not be maintained here; we shall use these terms interchangeably" (p.652). It is quite ironic that these were the very researchers who introduced the phenomenon of preference reversal to the discipline of decision making (Slovic and Lichtenstein, 1968) and subsequently provided an explanation for the phenomenon in terms of response mode effects (Lichtenstein and Slovic, 1971, 1973).

The literature on the effects of response mode on preferences is also noteworthy in that the impetus for such investigations came not from theoretical concerns about the effects of response mode on preferences but
from field observations about such effects. Indeed, most early studies on
decision making -- at least implicitly -- accepted the assumption of
procedural invariance. For instance, Coombs and Pruitt's (1960) study of
variance preferences was intended to be compared with Edwards' (1953,
1954a, 1954b, 1954c) studies of probability preferences, although Coombs
and Pruitt's study used choice as the response mode while Edwards' studies
used pricing judgments (more specifically, selling price or willingness to
accept judgments). Lichtenstein's (1965) later study, again using selling price
or willingness to accept judgments as the response mode, was intended for
direct comparison with Coombs and Pruitt's (1960) study.

It was not until the late 1960s that changes in preference ordering of
alternatives under evaluation as a function of response mode were noticed in
some empirical studies which used multiple response modes to measure
preferences (see Slovic and Lichtenstein, 1968). This observation
subsequently came to be known as the phenomenon of preference reversal.

In spite of the empirical work by experimental psychologists on the
phenomenon of preference reversal (Lindman, 1971; Slovic and Lichtenstein,
1968; Lichtenstein and Slovic, 1971, 1973), behavioral decision theoretic
researchers did not explore the response mode issue or the judgment versus
choice issue in much detail until the late 1980s. Most behavioral decision
theoretic researchers seemed to be more interested in the breakdown of the
other salient assumption of decision making, i.e., descriptive invariance as
opposed to procedural invariance. Accordingly, there were several
published studies on the effects of framing of alternatives and the effects of
the presentation of information about alternatives on decision making
processes and outcomes (for a review see Payne, Bettman, and Johnson, 1992). The few articles which noted the response mode issue in the late 1970s and early 1980s mainly did so in passing, and that also from a conceptual point of view (Bettman, 1981; Billings and Marcus, 1983; Einhorn and Hogarth, 1981; Einhorn, Kleinmuntz, and Kleinmuntz, 1979; Payne, 1982). It was not until the mid to late 1980s that attention began to be focused on response mode effects and response mode models were developed to account for the effects of response mode on decision making in general and the phenomenon of preference reversal in particular (Goldstein and Einhorn, 1987; Tversky, Sattath, and Slovic, 1988; Mellers, Ordonez, and Birnbaum, 1992).

It was noted in Chapter II that response mode effects were largely responsible for the phenomenon of preference reversal. There have been several explanations of how exactly different response modes are responsible for the phenomenon of preference reversal. These explanations can generally be divided into three major theoretical expositions, which subsume most of the different explanations. These theories can best be understood by considering them with respect to the framework of subjective measurement presented in Figure 3. Specifically within Figure 3, the differences between the three major types of theories forwarded to account for response mode effects can be understood in terms of the relative emphases placed on the three different functions shown as important in the framework of subjective measurement -- $H$, $C$, and $I$.

The first major theoretical exposition of response mode effects was Expression Theory. Forwarded by Goldstein and Einhorn (1987), expression
theory rests on a tripartite conception of the decision making process, consisting of the following stages: (a) stimuli are first encoded, (b) encoded stimuli are evaluated, and (c) evaluated stimuli are expressed onto a response scale of some type (Goldstein and Einhorn, 1987). During the first stage, the decision maker arrives at a psychological representation of the alternatives. Expression theory does not specify exactly how this psychological representation is arrived at by decision makers. Indeed, as Goldstein and Einhorn (1987) have noted, "This representation may take the form of encoding a subjective value for each dimension of the gamble (i.e., probabilities and outcomes are encoded as subjective probabilities and utilities) or the person may more or less directly perceive the gambles as having a particular amount of some psychologically relevant dimension ..." (p.240).

Further, the relation between expression theory and prospect theory (Kahneman and Tversky, 1979) should be noted in that the encoding stage in expression theory may involve both the editing as well as the framing operations specified by prospect theory. The evaluation stage involves the integration of the encoded elements into a measure of the overall or comparative worth of the alternative(s). This stage is nothing but the combination function or model form used to integrate the component dimensions or attributes of an alternative into an overall evaluation of the alternative. Once again, the parallel to prospect theory (Kahneman and Tversky, 1979) is apparent, for the second phase of the choice process in prospect theory is an evaluation phase.
The third stage of expression is the heart of the theory, for it is this stage which is suggested as being responsible for preference reversal. This stage involves the relation of the subjective worth or preference score to the response measure by which the decision maker responds. These three stages correspond to the \( H \), \( C \), and \( J \) functions in Figure 3. It is in this stage that expression theory moves beyond the scope of prospect theory, since prospect theory implicitly assumes that choice favors the alternative with the superior basic evaluation and therefore does not consider further stages in decision making. Indeed, Kahneman and Tversky (1979 p.284) have themselves noted that some modification of prospect theory is required to account of the phenomenon of preference reversal.

A fundamental assumption of expression theory is that the encoding and evaluation stages are the same for all response modes. Therefore, regardless of the particular response mode, the subjective worth or preference score of an alternative is the same (\( \Psi' \)). It is in the expression stage that different response modes exert their influence, and that is how the theory gets its name. When choices are used as the response mode, it is assumed that the alternative with the highest subjective worth or preference score is chosen. When judgments are used as the response mode, decision makers use a proportional matching strategy between the subjective evaluation dimension and the external response scale. Specifically, the judgmental response is determined by a process of matching the proportional distance from the subjective worth or preference score of the best outcome (\( \Psi_{\text{best}} \)) to \( \Psi' \) relative to the distance from \( \Psi_{\text{best}} \) to the preference score of the worst outcome (\( \Psi_{\text{worst}} \)), with the distance from the top of the
response scale \((T)\) to the response \((R_y)\), relative to the distance from the top to the bottom of the scale \((B)\). Thus,

\[
\frac{(\Psi_{\text{best}} - \Psi_y)}{(\Psi_{\text{best}} - \Psi_{\text{ave}})} = \frac{(T - R_y)}{(T - B)} \tag{4.11}
\]

This is what Goldstein and Einhorn (1987) refer to as a ‘proportional adjustment’ that reflects the subjective worth of the alternative relative to the best and worst outcomes. Respondents are theorized to map this proportional adjustment to a comparable point having the same proportional adjustment for each response mode. This mapping occurs by means of what Goldstein and Einhorn (1987) call a ‘subjective interpolation process.’

In giving pricing judgments, the top of the scale is the best outcome (i.e., \textit{best}) and the bottom is the worst outcome (i.e., \textit{worst}), whereas in giving rating-scale judgments the top and bottom are the best and worst possible ratings, respectively. Disparities between the ranking of evaluations and responses are due to differences in the curvature of the preference and response scales. In other words, it is the expression of the preference, rather than the way it is determined, that may lead to reversals (Goldstein and Einhorn, 1987; Schkade and Johnson, 1989). Therefore, expression theory attributes preference reversals to the differences in output processes that transform subjective worths of or preferences for alternatives into a response. Therefore, this theory attributes response mode effects to the function \(f\) in Figure 3.

The second major theoretical exposition of response mode effects was by Tversky, Sattath, and Slovic (1988), who proposed \textit{Contingent Weighting}
Theory to account for discrepancies in the preferential ordering of alternatives obtained using different response modes. In this regard, Tversky, Sattath, and Slovic (1988) proposed a hierarchy of contingent tradeoff models, which get their name from the fact that they were intended to describe the tradeoffs among inputs of the alternatives under evaluation which are contingent on the nature of the output, i.e., the response mode. A special case of these tradeoff models is the contingent weighting model, according to which the tradeoff between attributes or dimensions of an alternative depends on the nature of the response, and the weight associated with each attribute or dimension varies as a function of its compatibility with the output or response mode. In developing these contingent tradeoff models, Tversky, Sattath, and Slovic (1988) incorporated previously suggested explanations of response mode effects, i.e., the anchoring and adjustment hypothesis, the prominence hypothesis as well as the compatibility hypothesis into an overall framework to explain response mode effects in general.

The anchoring and adjustment hypothesis was the first explanation forwarded for the phenomenon of preference reversal (Lichtenstein and Slovic, 1971, 1973). In their pioneering studies on preference reversal (see Slovic and Lichtenstein, 1968; Lichtenstein and Slovic, 1971, 1973), these researchers attributed the pricing judgment results to a starting point (i.e., anchoring) and adjustment procedure -- subjects setting a price on an attractive gamble appeared to start with the amount to win and adjust it downward to take into account the probability of winning and losing as well as the amount that could be lost. Lichtenstein and Slovic (1971) conjectured
that the adjustment process was relatively imprecise on account of human information-processing constraints — a conjecture which was subsequently supported by Slovic (1972) and Tversky and Kahneman (1974) — leaving the pricing judgments greatly influenced by the starting point payoff. Although this explanation was initially forwarded with respect to a particular type of judgmental response (i.e., pricing judgments), it can easily be extended to other judgmental response modes.

As noted earlier, although Lichtenstein and Slovic (1971, 1973) had suggested the anchoring and adjustment hypothesis as an explanation of the decision making process in giving judgments, they did not forward a corresponding hypothesis for choices. It was Slovic (1975) who suggested a possible explanation for the decision making processes underlying choices and a hypothesis for the discrepancy between judgment and choice. Slovic (1975) named this hypothesis the "more important dimension" hypothesis, although it has now come to be known as the prominence hypothesis (Tversky, Sattath, and Slovic, 1988). This hypothesis simply states that in choosing between or among alternatives, respondents base their choices on the attribute or dimension which is most important. However, in judging alternatives, each attribute or dimension is given equal consideration. Therefore, the more prominent attribute or dimension of an alternative will loom larger (i.e., receive a higher preference score) when choices are used as the response mode as opposed to when judgments are used as the response mode. Initially forwarded by Slovic (1975), this hypothesis was supported by Tversky, Sattath, and Slovic (1988) in an experiment which controlled for some of the potential confounding variables in Slovic's (1975) original study.
A noteworthy aspect of the prominence hypothesis is the aspect of justification or accountability. As Slovic, Fischhoff, and Lichtenstein (1982) have noted, one issue that drives much of the deliberation prior to choice is the issue of justifying why one particular option was selected over the others. It is easy to justify a choice of a particular alternative on the basis that it was the best of all alternatives in the choice set on the most important dimension or attribute from the decision maker's perspective. For instance, consider Tversky's (1972) Elimination By Aspects model. According to this model, alternatives are viewed as sets of aspects [Tversky (1972) uses the term aspects to refer to what have been referred to as attributes or dimensions in this dissertation]. At each stage of the choice process, one aspect is selected, with probability proportional to its importance. The alternatives which are not adequate for each selected aspect are eliminated from the set of alternatives considered at the following stage of the choice process. Tversky (1972) argued that this model is an appealing process because it is easy to apply and justify.

Tversky, Sattath, and Slovic (1988) have drawn the distinction between qualitative and quantitative arguments for choice. Qualitative or ordinal arguments are based on the ordering of levels within each dimension or attribute or on the prominence ordering of the dimensions or attributes. Quantitative or cardinal arguments, on the other hand, are based on the comparison of value differences along the primary and secondary dimensions or attributes. Thus, dominance and lexicographic decision making models are qualitative arguments, while most other models of multiattribute decision making make extensive use of quantitative
arguments. The prominence hypothesis indicates that qualitative considerations loom larger in the ordinal procedure of choice rather than in the cardinal procedures of judgment. Therefore, Tversky, Sattath, and Slovic (1988) reason that the prominence hypothesis may be viewed as an example of what they term as a more general principle of compatibility.

According to the compatibility principle, the weight of any input component is enhanced by its compatibility with the output (Tversky, Sattath, and Slovic, 1988). The rationale for this principle is twofold (Tversky, Sattath, and Slovic, 1988; Slovic, Griffin, and Tversky, 1990). First, the characteristics of the task and the response scale prime the most compatible features of the stimulus. Further, noncompatibility between the input and output requires additional mental transformations, which increase effort and error, and reduce confidence and impact (Fitts and Seeger, 1953; Wickens, 1984). Along these lines, Slovic (1972) has noted that decision makers appear to use only the information that is explicitly displayed in the formulation of the problem and that information that has to be inferred form the display or created by some mental transformation tends to be ignored. Empirical evidence for this proposition has been presented in Slovic and Lichtenstein (1968) and Payne and Braunstein (1971). This would support the compatibility principle, since decision makers would not need to initiate any mental transformations to convert the input information into a format suitable for the required output.

The general principle of compatibility subsumes three types of compatibility -- strategy compatibility, semantic compatibility, and scale compatibility. Strategy compatibility was introduced earlier, in the case of
qualitative versus quantitative arguments. Briefly, the particular decision making strategy used by respondents will be determined by the nature of the response mode. More specifically, judgmental response modes will lead to more compensatory decision making strategies, while choice-based response modes will lead to more noncompensatory decision making strategies.

The issue of semantic compatibility is related to research in similarity and dissimilarity judgments (Tversky, 1977, Tversky and Cati, 1978). Tversky (1977) defined the similarity between objects $a$ and $b$ in terms of feature sets denoted by $A$ and $B$ respectively, in a similarity measure $S(a,b)$ given by the following equation:

$$S(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$$

(4.12)

where

- $(A \cap B) =$ features that $a$ and $b$ have in common
- $(A - B) =$ features that are distinctive to $a$
- $(B - A) =$ features that are distinctive to $b$
- $\theta, \alpha, \beta =$ parameters that affect the salience ($f$) of the feature sets

Tversky (1977) argued that with judgments of similarity, the focus is on the set $(A \cap B)$. On the other hand, with judgments of dissimilarity, the focus is on the distinctive sets $(A - B)$ and $(B - A)$. The compatibility principle suggests that common features loom larger in judgments of similarity than in judgments of dissimilarity, whereas distinctive features loom larger in
judgments of dissimilarity than in judgments of similarity. Therefore, the two judgments are not mirror images of one another. Tversky (1977) noted that a pair of objects with many common and many distinctive features could be judged as more similar as well as more dissimilar than another pair of objects with fewer common and fewer dissimilar features. This proposition was supported empirically by Tversky and Gati (1978), thus lending support to the compatibility principle since these results were explained on the basis that the relative weight of the common features is greater in similarity judgments than in dissimilarity judgments. Payne (1982) extended this conceptual argument to the process of choosing from a set of alternatives, when he noted that choice would seem to be more related to a dissimilarity response, i.e., what determined a choice between \(a\) and \(b\) was the distinctive features of \(a\) and \(b\) [i.e., \((A - B)\) and \((B - A)\)], not the features held in common [i.e., \((A \cap B)\)]. Indeed, some theories of decision making, such as prospect theory (Kahneman and Tversky, 1979) suggest that in the case of risky decision making, probability-outcome combinations held in common by two prospects will be edited out of the decision problem. On the other hand, in the typical judgmental response, all the features of an alternative are likely to be considered. In this way, the issue of semantic compatibility may explain the judgment versus choice distinction.

The issue of scale compatibility simply relates to the scales in which the inputs and outputs are expressed. For instance, pricing judgments for lotteries may be more likely to emphasize payoffs rather than probabilities because both the response and the payoffs are expressed in the same measurement unit. Scale compatibility has been shown to be a potential
causative factor in the phenomenon of preference reversal since the early studies on preference reversal. For instance, Slovic and Lichtenstein (1968), in their pioneering study on preference reversal, found that when subjects evaluated the attractiveness of a lottery using rating-scale judgments, the probability of winning was the most important attribute or dimension in their evaluation. However, when subjects evaluated the same lotteries in terms of pricing judgments, the payoff of the lottery was the most important attribute or dimension in their evaluation. The payoffs, being expressed in monetary units (i.e., dollars), were readily compatible with the units of the responses (also dollars). Slovic and Lichtenstein (1968) interpreted this compatibility as leading subjects to use one of the payoffs as a starting point for the response (i.e., an anchor). These starting points became the primary determinants of the responses due to insufficient adjustment (see Tversky and Kahneman, 1974). On the other hand, probabilities had to be transformed by subjects into values commensurate with dollars before they could be integrated with these other cues. Slovic and Lichtenstein (1968) hypothesized that the cognitive effort required to make this sort of transformation greatly detracted from the influence of probability cues in the pricing judgments. Indeed, in their 1973 article on preference reversals Lichtenstein and Slovic have noted: “The overdependence on payoff cues in pricing a gamble suggests a general hypothesis that the compatibility or commensurability between a cue dimension and the required response will affect the importance of the cue in determining the response” (Lichtenstein and Slovic, 1973 p.20). It should be noted that what Lichtenstein and Slovic
(1973) refer to as cues have been referred to as dimensions or attributes in this dissertation.

One of the earliest studies designed expressly to test the hypothesis of scale compatibility was by Slovic and MacPhillamy (1974). These researchers conducted five experiments in which subjects compared pairs of students with respect to potential college GPA. Both students had scores on one common dimension and one unique dimension. Slovic and MacPhillamy (1974) hypothesized that cue dimensions would have greater influence on comparative judgments when they were common to both alternatives than when they were unique to a particular alternative. The rationale for this hypothesis was as follows: the main source of cognitive effort arose from the fact that the cues may vary in importance and therefore must be weighted differentially. A comparison between the two stimuli along the same dimension allows the importance weight factor to cancel out and therefore should be easier and cognitively less demanding than a comparison between dimensions. This ease of use could lead to greater reliance on the common dimension. This hypothesis was supported by Slovic and MacPhillamy's (1974) data. These effects persisted even after cautioning subjects not to increase the weight of the common dimension and offering incentives for 'correct answers.' Further, the effect was substantial whether or not the common and unique dimensions had equal means and standard deviations. Finally, it should be noted that after the experiments, most subjects indicated that they had not wanted to give more weight to the common dimension and that they were unaware of having done so (Slovic and MacPhillamy, 1974). Since there is no common dimension when alternatives are judged one at a
time, one would expect a dimension that was common in choice to be given less weight in a judgment response mode. Evidence for this was also found by Slovic and MacPhillamy (1974).

Slovic, Griffin, and Tversky (1990) have noted that the compatibility principle may be mediated by a process of anchoring and adjustment -- subjects may use the score on the compatible variable (i.e., the attribute that matches the criterion) as an anchor and then adjust this number upward or downward according to the value of the noncompatible variable. Since such adjustments are usually insufficient (Tversky and Kahneman, 1974), the compatible attribute may be overweighted. Therefore, an anchoring and adjustment process may provide a natural mechanism for generating compatibility effects (Slovic, Griffin, and Tversky, 1990).

Slovic, Griffin, and Tversky (1990) conducted a series of experiments to empirically test the compatibility principle. They first demonstrated the compatibility effect in predictions of market value and course grades, where in each case the weight of a stimulus dimension or attribute was greater when it matched the response scale than when it did not. Next, they applied the compatibility principle to the study of choice and investigated the hypothesis that preference reversals were caused by the fact that payoffs of lotteries were weighted more heavily when subjects gave pricing judgments as opposed to when subjects used choice as the response mode. Such reversals were demonstrated in risky as well as riskless situations. They interpreted the collective results of their experiments as supportive of the role of the compatibility principle in causing response mode effects, although they acknowledged that it need not be the only factor accounting
for response mode effects in a given situation (Slovic, Griffin, and Tversky, 1990).

Further support for the compatibility principle was provided by Delquie and de Neufville (1988). These researchers employed a double-matching procedure, devised by Hershey and Schoemaker (1985), in which subjects first determined the missing value of an option that would make it equivalent to a second option. Later, the subjects were presented with the option they constructed and they now had to determine the missing value of the second option that would make the two options equally attractive. In this case, it should be noted that if procedure invariance holds, the latter match should coincide with the given value of the second option. Using both risky and riskless situations, Delquie and de Neufville (1988) found systematic violations of procedural invariance, which implied that the matched attribute was weighted more heavily than the other attribute, as predicted by the compatibility hypothesis.

Therefore, the compatibility principle is general enough to subsume the anchoring and adjustment and prominence hypotheses. In order to accommodate the compatibility effects observed in studies of preference, prediction, and judgment, Tversky, Sattath, and Slovic (1988) developed contingent weighting theory, which generally posits that the tradeoffs between attributes or dimensions of an alternative are dependent or contingent on the nature of the response mode. They applied this theory to the choice-indifference judgment discrepancy as well as the phenomenon of preference reversal, and found that the theory gave an adequate explanation of these phenomena.
Contingent weighting theory assumes that shifts in rank orders are attributable to changes in the subjective weights of attributes or dimensions of the alternatives under evaluation. Since the stimulus parameters depend on the method of elicitation, this account can be interpreted as attributing preference reversals to specific changes in the psychophysical function $H$ in Figure 3. It should be noted that Tversky, Sattath, and Slovic (1988) do present a more general contingent tradeoff model in which scales and/or processes (i.e., $H$ and/or $C$ in Figure 3) could change, depending on the task. However, they do not offer any further theoretical development of this model in terms of predictions about the conditions under which scales or processes would vary. Still, it should be acknowledged that such a model could conceivably bridge two response mode effect theories -- contingent weighting theory and change of process theory.

Before moving on to the third major theoretical exposition (i.e., change of process theory), another explanation which attributes response effects to changes in the subjective weights of attributes or dimensions of alternatives needs to be discussed. Initially proposed by Montgomery (1983; Montgomery and Svenson, 1989), the dominance structuring hypothesis refers to a decision maker's attempts to arrive at the indisputably best choice by modifying his or her beliefs and/or evaluations related to different attributes or dimensions in such a way that a favored alternative is: (a) not seen as worse than any of the other alternatives on any relevant attribute or dimension and (b) better than all the other alternatives in at least one relevant attribute or dimension. The relationship between this hypothesis and the issue of justification or accountability in decision making is
especially noteworthy, since the purpose of dominance structuring is to arrive at the 'indisputably best' alternative from a given choice set of alternatives.

Two published studies which have compared the dominance structuring hypothesis with two other explanations of response mode effects consistent with contingent weighting theory, i.e., the prominence hypothesis and the compatibility principle, will be reviewed here. The first of these studies, by Lindberg, Garling, and Montgomery (1989), compared the dominance structuring and prominence hypotheses in the context of explaining the differential predictability of preference ratings and choices by algebraic models of decision making. These researchers reasoned that if changes in subjective belief-value structures posited by the dominance structuring hypothesis take place when a choice is made from a set of alternatives, this should make it difficult to predict that choice by combining previously stated evaluations of individual attributes or dimensions in the way predicted by a MAP model. On the other hand, the prominence hypothesis could also explain the poor predictive accuracy of algebraic models of decision making, since this hypothesis suggests that the most important or prominent attribute or dimension will be given the highest weight in decision making with choice as the response mode, to the point of ignoring other attributes or dimensions in decision making. Therefore, if MAP models include all these other ignored attributes or dimensions, they would have poor accuracy in predicting choice (Lindberg, Garling, and Montgomery, 1989).
Accordingly, these researchers conducted two experiments aimed at investigating the differential predictability of judgmental ratings and choices by algebraic models of decision making. Specifically, these researchers were interested in understanding which of the two hypotheses -- dominance structuring or prominence -- could better account for this differential predictability. The specific algebraic model they used was a variant of a simple type of MAP model with equal or unit weighting of attributes (Edwards and Newman, 1982), with the variant being that the evaluation of each attribute or dimension was determined by the beliefs held by the subject concerning the attribute vis-à-vis the ability of the attribute in helping the individual to attain certain goals and the evaluation of those goals. Thus the variant was nothing but an expectancy-value model (see Ajzen and Fishbein, 1980; Feather, 1982).

In order to test for changes in belief-value structures (as posited by the dominance structuring hypothesis), they measured the beliefs and evaluations of attributes (i.e., the expectancy-value part of the modified MAP model) both before as well as after judgmental ratings and choices were made by subjects. According to the dominance structuring hypothesis, if changes in subjective belief-value structures explained the higher difficulty of predicting choices as opposed to preference ratings then the subjective beliefs and evaluations should change more following a choice task than a judgmental rating task. Further, the prediction of choices should also be expected to show greater improvement than that of preference ratings when subjective beliefs and evaluations are measured after the choice
or judgmental rating task, rather than before the tasks (Lindberg, Garling, and Montgomery, 1989).

If, on the other hand, the prominence hypothesis was the reason for the differential predictability of preference judgments and choices, subjective beliefs and evaluations need not be expected to change following a choice task. Instead, the source of the poor accuracy of the MAP model would lie in the combination rule used by the model since many attributes or dimensions are disregarded in making choices but not necessarily judgments. This was tested by constructing the MAP model with different subsets of attributes or dimensions -- if one or another of these subsets yielded better predictions of choices than the full set of attributes or dimensions, this would support the prominence hypothesis (Lindberg, Garling, and Montgomery, 1989).

In the first experiment, subjective belief-value structures of subjects were measured on two occasions, separated by 5-10 days. Before the second measurement, subjects performed either a judgmental rating or choice task (i.e., response mode was a between-subjects factor). Therefore, it was expected that changes in belief-value structures would have to be more or less permanent in order to be detected by this experiment. In the second experiment, an attempt was made to measure belief-value structures in connection with each choice or judgmental rating in order to find out whether temporary changes in such structures might account for the differential predictability of preferential ratings and choices (Lindberg, Garling, and Montgomery, 1989).
In both experiments choices were more difficult to predict than judgmental ratings with the modified MAP model. In the first experiment, there were no differences between the subjective belief-value structures in the two sessions -- therefore the dominance structuring hypothesis was not supported. However, the prominence hypothesis was supported by the first experiment, since a reduced MAP model with a subset of the attributes or dimensions of the original modified MAP model yielded better predictions of choices but not preferential ratings than the original modified MAP model. The results of the second experiment indicated that even temporary changes in belief-value structures could not explain the differential predictability of preferential ratings and choices via the dominance structuring hypothesis (Lindberg, Garling, and Montgomery, 1989). Taken together, the results of these two experiments support the prominence hypothesis over the dominance structuring hypothesis as a potential explanation for the differential predictability of preferential ratings and choices.

Subsequent studies have taken the prominence effect as a given, and tried to account for it using various explanations. For instance, in a pilot study, Montgomery, Garling, Lindberg, et al. (1990) investigated the role of dominance structuring in accounting for the prominence effect. Subjects were asked to: (a) give judgmental ratings for a set of alternatives, (b) choose one alternative in pairs of alternatives made equally attractive through a matching procedure (see Tversky, Sattath, and Slovic, 1988), and -- in connection with each of these tasks -- (c) give judgmental ratings of the attractiveness of the levels used to operationalize the attributes or
dimensions. It was found that the ratio of the difference between the attractiveness judgments of the levels of the prominent attribute or dimension on the one hand and the difference between the attractiveness judgments of the levels of the nonprominent attribute on the other hand was larger for when choice as opposed to judgmental ratings was used as the response mode. Further, there was a large discrepancy between judgmental ratings and choices. These results support the role of the dominance structuring hypothesis in accounting for the prominence effect.

However, Montgomery, Garling, Lindberg, et al. (1990) did not find a difference between judgmental ratings and choices when both were made on simultaneously presented alternatives. This was inconsistent with both, the dominance structuring hypothesis as well as the compatibility principle. On a related note, Birnbaum (1992) noted that simultaneous presentation of alternatives induces subjects to make implicit choices even though they may be requested to perform judgmental tasks. Therefore, apart from response mode, presentation mode may account for the discrepancy between preferences inferred from judgmental versus choice data.

Prompted by the above pilot study, Montgomery, Selart, Garling, et al., (1994) compared the relative merits of the dominance structuring hypothesis and the compatibility principle in the explanation of the prominence effect. Therefore, this study took the prominence effect as given (this was subsequently supported in the study) and investigated whether the dominance structuring hypothesis or the compatibility principle could better account for the prominence effect. A noteworthy extension of this study is that the number of alternatives in each choice set under consideration was
four -- most of the other studies reviewed here used only pairs of alternatives in each choice set. Therefore, this study permitted the extension of the explanation of response mode effects to situations in which more than two alternatives were under evaluation simultaneously. Further, on account of the presentation mode bias found in the pilot study, this study used a presentation mode where all alternatives were shown to the subject simultaneously (Montgomery, Selart, Garling, et al., 1994).

Subjects were divided into six groups. Response mode was a between-subjects factor in this study. One-third of the subjects performed only judgmental ratings or choices, another third gave judgmental ratings of attractiveness of attribute levels before performing judgmental ratings or choices, while the final third performed judgmental ratings or choices accompanied by verbal protocols or think-aloud reports (Montgomery, Selart, Garling, et al., 1994). Verbal protocols are helpful in determining whether subjects followed dominance structuring or not, since inferences can be drawn about how subjects both attend to, as well as evaluate, attributes and attribute levels (Montgomery and Svenson, 1989). Further, subjects may provide arguments in their think-aloud reports on the basis of which the true explanation of the prominence effect may be inferred.

The verbal protocols collected supported the dominance structuring hypothesis as an explanation for the prominence effect observed when choice was used as the response mode. The most striking result of this study was that a prominence effect was found for both choices as well as judgmental ratings. In this regard, the results replicated those of Montgomery Garling, Lindberg, et al. (1990). It should be noted that both
these studies found such a result when a presentation mode was used in which the alternatives were presented simultaneously to respondents. Therefore, this result did not seem to be confined to the two-alternative case studied in Montgomery Garling, Lindberg, et al. (1990). However, this result was inconsistent with both the dominance structuring hypothesis as well as the compatibility principle.

In order to explain this result, Montgomery, Selart, Garling, et al., (1994) proposed a more general form of the compatibility principle. According to Slovic, Griffin, and Tversky (1990), compatibility concerns characteristics such as structure, type, and amount of information in input and output. Montgomery, Selart, Garling, et al.’s (1994) generalization posited that it was the structure of information required from subjects that needed to be compatible with the input. They argued that whereas the choice task was qualitative, both judgmental rating as well as indifference judgment tasks are quantitative. However, the difference between these judgmental response modes is that in judgmental ratings subjects process alternatives one at a time, while in indifference judgments they process multiple alternatives together. Therefore, they argued, one might expect a prominence effect while using judgmental ratings as opposed to using indifference judgments. Further, the role of presentation mode is also important. In this regard, a simultaneous presentation mode would increase the similarity between choices and judgmental ratings (of simultaneously presented options). This is because even though judgmental ratings by themselves do not entail choices, they may involve implicit comparisons if
the alternatives are presented simultaneously (Montgomery, Selart, Garling, et al., 1994).

Change of Process Theory, the third major theoretical exposition of response mode effects, was proposed by Mellers, Ordonez, and Birnbaum (1992). This theory assumed that the subjective worths for or preference scores of the stimulus parameters (i.e., dimensions or attributes used to describe the alternative under evaluation) remain constant over response modes, but the process used to combine this information varies as a function of the response mode. It should be noted that the possibility of changing decision strategies is the central tenet of the theory of an individual as an adaptive decision maker, which is propounded by the research of Payne, Bettman, and Johnson over the course of the past fifteen years (for a review, see Payne, Bettman, Coupey, et al., 1992; Payne, Bettman, and Johnson, 1990, 1992, 1993). The theory of adaptive decision making suggests that the particular decision process used by an individual at any given opportunity is contingent upon the demands of the situation, and is affected by the response mode, the effort required, the accuracy needed, time pressures, etc. In other words, decision making strategies are constructed on the spot at the time of decision making. Change of process theory complements this theory by postulating specific models for judgmental response modes for certain types of alternatives (Mellers, Ordonez, and Birnbaum, 1992). For example, change of process theory postulates that probabilities and payoffs combine multiplicatively in pricing judgments, but additively in judgmental ratings. Further, change of process theory has the added premise of scale convergence (Birnbaum, 1974; Birnbaum and Veit, 1974) -- this means that
preferences measured using different measurement methodologies are functionally related to one another, since they are representations of the same underlying subjective scale of preference. According to the change of process theory, it is the combination function that varies with the response mode. Therefore, in terms of Figure 3, the attention is focused on the function C.

In the earliest published application of change of process theory, Mellers, Ordonez, and Birnbaum (1992) conducted three experiments to investigate contextual effects and response mode effects in risky decision making. Subjects provided attractiveness judgmental ratings and pricing judgments (more specifically, buying prices or willingness to pay judgments) for binary gambles using two different contexts -- positively and negatively skewed distributions of expected values. Consistent with previous research, preference reversals were obtained between attractiveness judgmental ratings and willingness to pay judgments. However, changes in context due to variations in skewing influenced the metric properties of the judgments but did not lead to preference reversal (Mellers, Ordonez, and Birnbaum, 1992).

In order to investigate which one of the major theoretical expositions of response mode effects -- i.e., expression theory, contingent weighting theory, or change of process theory -- could explain the data best, Mellers, Ordonez, and Birnbaum (1992) developed direct tests among the three theories in question. Expression theory and contingent weighting theory were inconsistent with the pattern of the data, but change of process theory could describe the changes in rank order across tasks as being the result of
the effect of the response mode on the manner in which respondents combined information and arrived at judgments. Specifically, under certain conditions, attractiveness judgmental ratings could be described by an additive combination of subjective probability and preference, whereas willingness to pay judgments were accounted for by a multiplicative function with the same scales of subjective probabilities and preferences in both tasks. Further, when the range of outcome values of the gambles included zero and negative values, preference orders for the attractiveness judgmental ratings of the gambles changed. This observation again could not be explained by either expression theory or contingent weighting theory, but could be explained by change of process theory by conjecturing that the inclusion of these levels caused more subjects to use a multiplicative combination function when giving rating-scale attractiveness judgments of the gambles in question (Mellers, Ordonez, and Birnbaum, 1992).

Shortly after this first empirical demonstration of change of process theory, Mellers, Chang, Birnbaum, et al. (1992) successfully extended change of process theory to other response modes, including selling price (or willingness to accept) judgments, risk judgmental ratings, unattractiveness judgmental ratings, avoidance pricing judgments, and strength of preference judgments. Different preference orderings were obtained when subjects used different response modes to judge gambles, even in the face of financial motivation to be consistent. Like the first study, these preference reversals could not be explained by either expression theory or contingent weighting theory, but could be described by change of process theory.
Specifically, risk and attractiveness judgments could be described by an additive combination of subjective worth or preference and subjective probability, avoidance price and selling price or willingness to accept judgments by a multiplicative combination rule, while strength of preference judgments could be described by a contrast-weighting model in which a stimulus dimension or attribute is weighted according to the difference between gambles along that dimension or attribute and the weighted products are compared by subtraction (Mellers, Chang, Birnbaum, et al., 1992).

These two studies present compelling evidence for the validity of change of process theory. Especially noteworthy is the fact that efforts were made to directly test the different response mode theoretical expositions on the same set of data and see which theoretical exposition explained the data the best. The major drawback of change of process theory is that it is focused mainly on various forms of judgments as the response mode -- there have been no published attempts to extend the theory to choice-based response modes.

As Mellers, Ordonez, and Birnbaum (1992) have noted, there are some interesting linkages between change of process theory and contingent weighting theory. When the alternatives under evaluation are binary gambles with some probability of receiving an amount (otherwise zero) and the amounts are either all positive or all negative, the change of process theory is a special case of a particular type of contingent tradeoff model (Tversky, Sattath, and Slovic, 1988 equation 4). This contingent tradeoff model asserts that probabilities and amounts are decomposable to an
additive structure in both tasks. This form of contingent tradeoffs could imply changing processes -- which is the view adopted by change of process theory -- since both the additive and multiplicative model forms that constitute change of process theory can be transformed to additivity (Mellers, Ordonez, and Birnbaum, 1992).

Having reviewed the three basic theories forwarded to explain response mode effects, it is now necessary to review the empirical literature which has tested these three theories. This review will only consider the studies which have not already been reviewed in the course of explaining the different theoretical expositions of response mode effects (as reviewed earlier in this section). Probably the first published empirical investigation on response mode effects (more specifically, the judgment versus choice issue) was a study by Slovic (1975). Slovic's research was prompted by the simple question: "What happens when a decision maker is faced with a choice between two alternatives that are equal value to him?" (Slovic, 1975 p.280). Slovic noted that most (then) existing theories of decision making implied that in such cases either alternative was equally likely to be selected by the decision maker. This was because the (then) existing theories typically assumed that choice was a probabilistic phenomenon where the probability of choosing object \( x \) over object \( y \), denoted as \( P(x,y) \), was some function of the scale scores of the preferences for \( x \) and \( y \). Under this assumption, \( P(x,y) = 0.5 \) implies that the preference score for \( x \) is equal to the preference score for \( y \).
Slovic's (1975) investigation was also influenced by the research in behavioral decision making of the late 1960s and early 1970s, which depicted the decision maker as continuously searching for heuristics in order to produce quick and reasonably satisfactory decisions (see Slovic, 1972; Slovic, Lichtenstein, and Edwards, 1965; Tversky, 1969; Tversky, 1972). This body of empirical research suggested that choice between equally valued alternatives was not random, as implied by the (then) existing theories on decision making, but systematic in that decision makers systematically determined their choices in such situations by selecting the alternative that was superior on the more important dimension or attribute. Slovic labeled this as the "more important dimension" hypothesis (Slovic, 1975), but it has now come to be known as the "prominence" hypothesis (Tversky, Sattath, and Slovic, 1988). Slovic's (1975) investigation was designed to test this hypothesis.

Slovic conducted four experiments. In each of these four experiments, subjects were asked to choose between pairs of alternatives that they had previously equated in value. Within each pair, one alternative was superior on an important dimension but so inferior on a less important dimension that this disadvantage nullified its advantage on the more important dimension. Slovic found that the majority of subjects resolved these choices by consistently selecting the alternative that was superior on the more important dimension, thus supporting the prominence hypothesis. He explained these results by conjecturing that the prominence hypothesis is operative in practice because it helps decision makers *justify* their choices (Slovic, 1975).
The effect of justification on the nature of the process of decision making was investigated by Hagafors and Brehmer (1983). Although this study did not study the effect of response mode on the degree of justification -- i.e., it used only judgment as the response mode -- it serves the current purpose to illustrate what having to justify one's decision does to the process by which one makes a decision. These researchers found that the act of having to justify one's judgment led to more consistency in decision making over time. They explained their findings by proposing that the act of justification led to a more analytical process of thinking in making judgments (Hagafors and Brehmer, 1983).

The importance of the Hagafors and Brehmer (1983) study to the present discussion is to provide empirical evidence that the act of justifying one's decisions plays a role in the process of decision making. It is especially noteworthy that these researchers found these results using only judgments as a response mode. This is because, as currently used in the health status preference measurement literature, respondents are not asked to justify their judgmental responses. Indeed, judgmental responses -- whether by ranking, rating, indifference, or pricing -- do not inherently require justification. Choice responses, on the other hand, require justification, if only to oneself -- i.e., "Why would I choose this particular alternative and not the others?"

As noted earlier, there were not many published empirical studies on the response mode issue in decision making in the late 1970s or early 1980s. There seemed to be an ignorance on part of researchers as to the potential effects of response modes on decision making. Indeed, some researchers went so far as to require subjects to "... reconcile any deviation between ...
ratings and choices" (Russo and Rosen, 1975). The Russo and Rosen study is particularly noteworthy in that subjects provided rating judgment and choice data on two separate occasions -- on both occasions they were made to 'reconcile' their evaluations in order to proceed with the experiment (Russo and Rosen, 1975). The few published articles in this period which did address the issue presented theoretical and/or logical arguments for the difference between judgment and choice (Einhorn and Hogarth, 1981; Payne, 1982). However, a study by Rosen and Rosenkoetter (1976) is a noteworthy exception, since it was one of the earliest published studies which tried to get into the 'black box' or respondents' minds -- in other words, Rosen and Rosenkoetter (1976) attempted to study the differential effects of response modes on actual cognitive processes involved in decision making.

Before discussing the specifics of the Rosen and Rosenkoetter (1976) study, it is necessary to introduce the terminology which will be adhered to in the following discussion on how response modes exert their influence on decision making. Individuals can follow either of two basic types of decision making strategies -- compensatory and noncompensatory. Compensatory decision making strategies assume that any one attribute or dimension of an alternative can be traded off for another. On the other hand, noncompensatory decision making strategies assume that there are some attributes or dimensions of alternatives which are of such importance that they must be present in an alternative in order for the alternative to be considered for selection. It is also possible that individuals may follow a combination of the two basic types of strategies in decision making. Within each of these basic types of decision making strategies there are several
decision making models, which represent the process by which individuals integrate information about the alternatives under evaluation and make a decision. For instance, the linear model is the most commonly studied model under compensatory decision making strategies. On the other hand, the conjunctive and Elimination By Aspects models have received empirical attention as models representing noncompensatory decision making strategies (Einhorn, 1970, 1971; Tversky, 1972). Whatever decision making strategy or model is used, individuals may process information by following a particular type of information search pattern. In this regard, there are two types of information search patterns, interdimensional and intradimensional. Interdimensional information search patterns have also been referred to as alternative-based information processing patterns, since they involve searching for information across dimensions or attributes within the same alternative. Intradimensional information search patterns, on the other hand are also referred to as dimension- or attribute-based information processing patterns, since they involve searching for information across alternatives within the same dimension or attribute.

Rosen and Rosenkoetter (1976) studied the actual decision making processes of respondents, as affected by two types of response modes -- judgmental ratings and choices. They were mainly interested in the particular type of information search pattern adopted by subjects in response to having to give judgmental ratings of stimuli versus having to choose between pairs of these stimuli. It should be noted that Rosen and Rosenkoetter (1976) used the terminology of 'evaluation strategy' to describe what has been referred to here as 'information search pattern.' For the
purpose of this discussion, the terminology introduced above will be followed.

Rosen and Rosenkoetter (1976) hypothesized that using judgment as a response mode encouraged respondents to use alternative-based information processing patterns, while using choice as the response mode encouraged respondents to use dimension- or attribute-based information processing patterns (Rosen and Rosenkoetter, 1976). It should be noted that this hypothesis was consistent with the anchoring and adjustment explanation offered by Lichtenstein and Slovic (1971, 1973) for the phenomenon of preference reversal.

Further, Rosen and Rosenkoetter (1976) were also interested in the moderating effect of stimulus environment on the type of information search pattern adopted by respondents. They operationally defined stimulus environment as the degree of interdependency among the dimensions or attributes comprising an alternative. They hypothesized that with stimuli whose attributes or dimensions were either interdependent or highly dissimilar, the selection of an information search pattern might be restricted. Specifically, they hypothesized that if the attributes or dimensions of stimuli were interdependent, an alternative-based information processing pattern may be used by subjects regardless of the nature of the response mode; if the attributes or dimensions of stimuli were dissimilar, a dimension- or attribute-based information processing pattern may be used by subjects regardless of the nature of the response mode. However, when the stimulus environment presented less extreme dimension or attribute configurations, they hypothesized that respondents would use a dimension- or attribute-
based information processing pattern with choice as the response mode but an alternative-based information processing pattern with judgmental ratings as the response mode. The degrees of dimensional dependency were embodied in the three types of stimuli they constructed -- gambles, gift packages, and vacation packages, representing the most, least (i.e., dissimilar) and intermediate levels of dimensional interdependence (Rosen and Rosenkoetter, 1976).

Using eye fixation analysis with six subjects comparing a choice response with a judgmental response -- where response mode was a within-subjects factor -- involving a 100-point scale (i.e., rating scale judgment), they found that using choice as the response mode did lead to more dimension- or attribute-based information processing patterns as opposed to using judgmental ratings as the response mode, which led to more alternative-based information processing patterns. Regarding the moderating role of stimulus environment, their hypotheses were partially supported. In accordance with their hypotheses, their respondents used an alternative-based information processing pattern for interdependent stimuli (i.e., gambles) regardless of the nature of the response mode and a dimension- or attribute-based information processing pattern for stimuli with dissimilar attributes or dimensions (i.e., gift packages) regardless of the nature of the response mode. However, although they found that their respondents did use dimension- or attribute-based information processing patterns more with choice as opposed to judgment as the response mode when evaluating stimuli with less extreme dimension or attribute configurations (i.e., vacation packages), they were unable to support their hypothesis that using judgment
as a response mode led to more alternative-based information processing patterns with such stimuli (Rosen and Rosenkoetter, 1976).

The major drawback of this study was its small sample size -- only six subjects. Further, the results they reported held for only five out of the six subjects. Another drawback is that the choice sets had only two alternatives, and therefore the results cannot be generalized to situations where respondents are faced with more alternatives to evaluate.

Still, the contribution of this study to decision making processes in judgment and choice should not be denied, since it was the first published study to specifically study the effects of response mode on decision making processes -- specifically the effects of response mode on the information search patterns used in decision making processes. Further, it introduced an important moderating variable -- stimulus environment -- as a point to consider in constructing stimuli for research in decision making. Specifically as applied to the measurement of health status preferences, as noted previously, there is reason to believe that health status dimensions or attributes might interact with one another. In other words, health state stimuli might be expected to be interdependent. Therefore, it is possible that respondents evaluating health state stimuli would be expected to use holistic strategies in decision making. It should be noted that the above line of reasoning depends on the specific operational definition of health status used, since this has a direct bearing on whether the health status dimensions or attributes are constructed to be interdependent or dissimilar.

It was not until a decade after Rosen and Rosenkoetter's (1976) study that the next empirical investigation of response mode effects on
decision making processes was published. This study, by Billings and Scherer (1988), also used judgmental ratings and choices as the response modes under investigation. However, Billings and Scherer (1988) extended the Rosen and Rosenkoetter (1976) study by investigating the effects of response modes on not just the information search patterns used by respondents but also other aspects relevant to the information search stage of decision making (as explained below). They hypothesized that compared to choices, judgmental ratings as a response mode would lead to more information searched, a less variable pattern of search, and a greater amount of alternative-based information search patterns. It should be noted that these hypotheses imply that judgmental ratings as response modes prompt the use of compensatory decision making strategies, while choices as response modes prompt the use of noncompensatory decision making strategies (Billings and Scherer, 1988).

Further, Billings and Scherer (1988) were also interested in the moderating effects of decision importance on the differential effects of judgmental ratings and choices on decision making processes. Specifically, they hypothesized that decisions considered to be of high importance would be characterized by systematic, careful decision processes — i.e., more compensatory decision making strategies, meaning more information searched, less variability of information searched across alternatives, and more alternative-based information search patterns. Moreover, they hypothesized an interaction between decision importance and response mode, wherein using judgments as the response mode would lead to compensatory decision making strategies regardless of the level of
importance because the more careful, analytical information processing would already be produced on account of the judgmental response mode. However, they hypothesized, when choices were used as the response mode decision importance would make a significant difference -- under low importance noncompensatory decision making strategies and dimension- or attribute-based information search patterns would predominate, while under high importance compensatory decision making strategies and alternative-based information search patterns would predominate even though they are not demanded by the response mode. In other words, they hypothesized that decision importance would produce analytical judgment processes prior to choice, even if the response mode does not require explicit judgment (Billings and Scherer, 1988).

Using information display boards, they tested these hypotheses in an experiment with response mode (judgmental ratings and choices) and importance (high and low) as between-subjects factors. In support of their hypotheses, their data showed that response mode had strong and consistent effects on all dependent variables. Judgmental response modes led to more information searched, more alternative-based information search patterns, and less variability of information searched across the alternatives. Decision importance had less of an effect -- although there was an effect of importance on the information search pattern (with more alternative-based information processing patterns for decisions of high importance), there was no effect of importance on variability of search or amount of information searched (Billings and Scherer, 1988).
Following Payne (1976), these researchers combined the search pattern and variability of search variables to get an indication of the type of decision making models used by respondents. They found that an Elimination By Aspects model of decision making was more likely using choice as opposed to judgment as a response mode and for decisions of low as opposed to high importance (Billings and Scherer, 1988).

Regarding the hypothesized interaction, they found mixed results. The hypothesized interaction was supported by the data for the amount of information searched, but not for any other dependent variable. Essentially, importance did not moderate the effects of response mode on the decision making processes used by their respondents -- choice as a response mode led to more variable and dimension- or attribute-based information processing patterns under both high and low decision importance. However, as noted earlier, importance did reduce the impact of response mode on the amount of information searched, since choice and judgment differed less on amount of information searched for decisions of high as opposed to low importance. These nonsignificant effects of decision importance were explained by the researchers on account of failure in operationalization of high versus low levels of decision importance (Billings and Scherer, 1988).

It should be noted that although Billings and Scherer (1988) were unable to show that decision importance affected the decision making processes used by individuals in their study, other researchers have found the opposite. Other researchers have operationalized decision importance in terms of personal involvement (e.g., Chaiken, 1980) and well as need of justification of either the choice itself (e.g., Chaiken, 1980) or the process
through which the choice was made (e.g., Hagafors and Brehmer, 1983). It is possible that the variable of decision importance is sensitive to particular operationalization adopted in any given study.

The Billings and Scherer (1988) study did, however, demonstrate the effect of judgmental versus choice on decision making strategies and models used by individuals. By doing so, this study established the logical chain of events, which Billings and Scherer (1988) aptly summarized: “Therefore, if choice leads to Elimination By Aspects and judgment leads to compensatory processing, different final decisions could result” (p.17). This is nothing but the phenomenon of preference reversal.

To further investigate the nature of the particular cognitive mechanisms involved in preference reversal, Schkade and Johnson (1989) conducted three process tracing experiments. The interesting aspect of this study is that it demonstrated response mode effects on decision making processes not only by using judgmental versus choice responses (as in most other studies), but also with judgmental ratings versus pricing judgments. Further, and more significant from a theoretical point of view, it seemed to be motivated by the need to empirically test the various explanations forwarded for the phenomenon of preference reversal (Goldstein and Einhorn, 1987; Lichtenstein and Slovic, 1971, 1973; Tversky, Sattath, and Slovic, 1988).

Schkade and Johnson’s (1989) first experiment investigated the differential effects of pricing judgments (more specifically, selling prices or willingness to accept judgments) and choices on decision making processes. Their second experiment was very similar to the first — the only difference
was that it compared pricing judgments (more specifically, selling prices or willingness to accept judgments) and rating-scale judgments as alternative response modes. In a rather detailed data collection procedure, these researchers collected information about total time taken per response, attention to stimulus features, patterns of information acquisition, and response activity using the Mouselab software package (see Johnson, Payne, Schkade, et al., 1988). This package is a computer-driven data collection procedure specifically designed for process tracing studies. The purpose of this amount of detail was to isolate the sources of preference reversals obtained using willingness to accept judgments versus choices as response modes.

The selection of these four measures as indications of the cognitive processes underlying decision making deserves further comment. Total time provides an approximate measure of overall effort and also permits the examination of specific issues about judgment and choice response modes. In this regard, some theories of choice (see, for e.g., Abelson and Levi, 1985; Schoemaker, 1982) and specific theories of preference reversal (see, for e.g., Goldstein and Einhorn, 1987) suggest that choice between two alternatives consists mainly of two evaluation judgments, one for each alternative -- if this is true, one might expect the time required to make a choice to be larger than the time required to make a single judgment.

Regarding measures of attention to stimulus features, it has been suggested that the amount of attention given to a stimulus feature may be related to its importance (e.g., Fiske, 1980; Taylor and Fiske, 1981) and the way it is processed (e.g., Just and Carpenter, 1976; Russo and Dosher, 1983).
For instance, in their eye fixation analysis Russo and Rosen (1975) found that alternatives with higher preference scores received greater attention. If the attention given to an item is related to its importance, measures such as the proportion of total time spent on various problem elements should vary across response modes.

Measures of information search pattern were the earliest used process tracing measures, simply because different decision making strategies are characterized by different information search patterns. For instance, the linear model of compensatory decision making is characterized by alternative-based information search patterns while an Elimination By Aspects model (see Tversky, 1972) of noncompensatory decision making is characterized by dimension- or attribute-based information search patterns.

Unlike measures of information search pattern, response measures have not been widely used in process tracing research. These measures are concerned with the process by which subjects generate their responses. While not very useful for choice data, these measures can potentially offer much insight into the response mechanisms for judgmental data -- especially anchoring and adjustment mechanisms.

The results of their first experiment provide insight into the cognitive processes underlying preference reversal between willingness to accept judgments and choices. On the basis of their results, they concluded that the cognitive processes underlying decision making using choices and willingness to accept judgments appeared to be quite different. Choice was a quicker process, involving a significant number of comparisons between alternatives. In fact, choosing between the two alternative lotteries used in
the study took less time than even pricing a single lottery. Willingness to accept judgments were characterized by greater attention to the payoff information about the lotteries, while choices were characterized by greater attention to probabilities of winning or losing the lotteries. Finally, a distinguishing feature of willingness to accept judgments was a fairly long period of response generation in which subjects seemed to use the scale to simulate various responses (Schkade and Johnson, 1989).

Schkade and Johnson (1989) had two interesting extensions to standard methodology in preference reversal studies. The first of these extensions was the differences between process tracing measures for reversals versus consistent responses. In this regard: (a) there were no significant differences for the total time taken for both types of responses, (b) the amount of attention given to probabilities relative to payoffs was strongly related to reversals -- further, the locus of this effect appeared to be in the choice mode [it should be noted that this is consistent with the Elimination By Aspects (see Tversky, 1972) model of noncompensatory decision making], (c) the information search pattern was related to reversals -- specifically, in the case of choice as the response mode search is more outcome-oriented for reversals than for consistent responses, and (d) in willingness to accept judgments, starting points are more extreme for reversals than for consistent responses (Schkade and Johnson, 1989).

The second interesting extension of this study was a regression analysis to predict preference reversals from process measures. In this analysis, five independent variables were entered -- difference between response modes in the proportion of time spent looking at probabilities,
proportion of time in choice spent looking at the probability of winning the lottery, search pattern in choice, relative position of starting points for the lotteries, and whether adjustments away from starting points were in the same direction for both lotteries. A dichotomous dependent variable (1 if reversal; 0 if consistent) was regressed on these five independent variables. The overall model was highly significant (p<0.0001; R² = 0.266), indicating that the processes that occur in providing choice as well as those that occur in providing willingness to accept judgmental data were highly related to preference reversals (Schkade and Johnson, 1989).

The results of this first experiment, taken together, suggest that using choice as a response mode leads to a different decision making process by respondents than using willingness to pay judgments as a response mode. While the two types of processes are related, the process measures used in this study indicated that they differ in several respects, including the amount of time to respond, relative attention to probabilities and payoffs, and response activity (Schkade and Johnson, 1989).

Schkade and Johnson (1989) attributed these results to at least two potential causes of preference reversal -- an anchoring and adjustment mechanism for willingness to accept judgments and the fact that noncompensatory strategies adopted when choice is used as a response mode may be related to the frequency of preference reversals.

The second experiment of Schkade and Johnson (1989) was intended to extend the scope of the first experiment to the case of preference reversals between willingness to accept and rating-scale judgments. The methodology used in this second experiment was identical to the first, except for the use of
a rating scale to evaluate individual lotteries instead of direct choices between lotteries. They found that the two judgment modes showed clear similarities but also important differences in terms of decision making processes of respondents. The most striking similarities between the decision making processes used as a result of the two modes was the large amount of time spent in and around the response scale and the frequent use of what was interpreted as an anchoring and adjustment mechanism. However, the starting points and direction of adjustment differed between the two judgment modes; other differences included greater attention to payoffs than probabilities in giving willingness to accept judgments but the opposite in rating-scale judgments, more dimension- or attribute-based information search patterns in giving willingness to accept judgments, and less total time spent per lottery for rating-scale judgments (Schkade and Johnson, 1989).

Regarding the differences between process tracing measures for reversals versus consistent responses: (a) in terms of total time taken for both types of responses, consistent responses took less time than reversals, (b) there was no significant relationship between reversals and the proportion of time spent examining probabilities, despite the large difference between the two judgment modes on this measure, (c) there was no significant relationship between frequency of reversals and information search patterns, and (d) starting points and the direction of adjustments were strongly related to reversal. These results were interpreted as providing support for the compatibility hypothesis (Schkade and Johnson, 1989).
Regarding the regression analysis to predict preference reversals from process measures, they used eight independent variables -- total time in rating, proportion of time in and around the response scale, relative positions of the starting points in the two bets when using willingness to accept judgments as the response mode, relative positions of the starting points in the two bets when using rating-scale judgments as the response mode, and the amount of adjustment for each combination of lottery type and response mode (there were four such combinations). The dependent variable was dichotomous (1 if reversal; 0 otherwise). The time in and around the response scale and amount of adjustments for $ bets in ratings were dropped due to lack of significance. The overall model fit was highly significant (p<0.0001, \( R^2 = 0.433 \)). Therefore, the occurrence of reversals appeared to be predictable -- at least in part -- by decision making process tracing measures (Schkade and Johnson, 1989).

The results of this second experiment, taken together, indicated that willingness to accept and rating-scale judgments shared largely similar strategies of anchoring and adjustment mechanisms. Although there were some differences (i.e., greater attention to payoffs of lotteries in willingness to accept judgments but to probabilities of lotteries in rating-scale judgments, longer time for willingness to accept judgments, and more dimension- or attribute-based information search patterns in willingness to accept judgments), these were not related to preference reversals (Schkade and Johnson, 1989).

The process differences between the two response modes which were related to reversals concerned the way subjects used information to generate
responses. Specifically, the roles of the elements of the lotteries seemed to be quite different for the response modes -- starting points in willingness to accept pricing judgments were heavily influenced by the greater payoff as opposed to the probability of winning in the case of rating-scale judgments. Moreover, this tendency was stronger for reversals than consistent responses -- thus, although the two types of response modes might share the same strategy subroutine, they give it different inputs to process (Schkade and Johnson, 1989).

In order to probe deeper into the role of the starting point in determining preference reversals between willingness to accept pricing and rating-scale judgments, Schkade and Johnson (1989) conducted a third experiment using a starting point manipulation. They hypothesized that such a manipulation would decrease the incidence of preference reversals. Although the starting point manipulation they explored did not totally eliminate preference reversals between willingness to accept pricing and rating-scale judgments, it did significantly reduce the incidence of such reversals in their subjects (Schkade and Johnson, 1989).

The Schkade and Johnson (1989) study was significant in that it provided a rather detailed analysis of the cognitive processes underlying decision making prompted by different response modes. This study did show that process measures of decision making bear strong relationships to preference reversals and -- at least in part -- were connected to fundamental process differences between judgment and choice. Further, this study also demonstrated the differences in cognitive processes in decision making prompted by differences in two types of judgmental response modes, thus
providing process-related information about a phenomenon first observed by Slovic and Lichtenstein (1968) and subsequently replicated by Goldstein and Einhorn (1987). Therefore, it seems that the response mode issue is more complex than merely a distinction between judgment and choice. In other words, one needs to specify which particular type of judgmental or choice response is used in any given study.

Overall, Schkade and Johnson's (1989) results were consistent with an explanation for preference reversal which emphasized starting points, anchoring and adjustment, and compatibility. In other words, the results supported Tversky, Sattath, and Slovic's (1988) contingent weighting response mode model. Moreover, the results also supported the focus on generating responses suggested by Expression Theory (Goldstein and Einhorn, 1987). Therefore, it appears that more than one response mode model may account for the phenomenon of preference reversal.

Finally, the research by Westenberg and Koele (1990, 1992) needs to be recognized for its contribution to the understanding of response mode effects. Westenberg and Koele (1990) conducted two experiments in which several response modes were examined. The first experiment adopted a process tracing approach and compared four response modes -- selecting, rejecting, classifying into two classes, and classifying into four classes. Selecting and rejecting induced more noncompensatory decision making strategies than classifying into two or four classes. The second experiment followed a regression-based modeling approach and compared ranking with classifying into two classes. Ranking induced more compensatory decision
making strategies than classifying into two classes (Westenberg and Koele, 1990).

In both experiments, there was more compensatory decision making as the response mode involved a greater number of scale points. Subjects used one explicit scale point when selecting or rejecting, two and four while classifying into two and four classes respectively, and as many scale points as the number of alternatives under evaluation under ranking. On the basis of these findings and in view of the fact that choice-based response modes generally induce more noncompensatory decision making strategies as compared to judgmental response modes, Westenberg and Koele (1990) proposed a response mode continuum, running from choice (one scale point) to judgment (many scale points) which in turn corresponded to decision making strategy continuum ranging from noncompensatory to compensatory decision making strategies. It was noted that the finer the response scale, the more differentiation needed in the evaluation, the more tradeoffs between attributes or dimensions (Westenberg and Koele, 1990).

Further, Westenberg and Koele (1990) also found an effect of response mode on the outcome of decisions -- subjects accepted significantly less alternatives when they were selecting as compared to when they were rejecting or classifying. Thus, selecting appeared to be a response mode quite different from the other three response modes studied by Westenberg and Koele (1990). This was also indicated by the finding that while selecting, respondents used dimension- or attribute-based information search patterns, while when rejecting or classifying respondents mainly used alternative-based information search patterns. Westenberg and Koele (1990) singled out
Westenberg and Koele (1990) attributed the discrepancy to the different frames prompted by the response modes. It should be noted that — although not recognized by Westenberg and Koele (1990, 1992) -- a similar discrepancy between selecting and rejecting alternatives was over a decade earlier by Coombs, Donnell, and Kirk (1978), who attributed the discrepancy due to changes in saliency of the various attributes or dimensions of the alternatives due to response mode effects.

This study was noteworthy in that it expanded the scope of research on response mode effects in three ways: (a) it drew attention to the effects on decision making processes and outcomes of the response modes of rejecting and classifying -- most of the previously published research on response mode effects used selecting as a choice-based response mode and pricing judgments, indifference judgments, or judgmental ratings as judgmental response modes and (b) it introduced the notion of scale points as a factor to consider in evaluating the effects of response modes on decision making processes and outcomes, and (c) it moved beyond preference reversals as the only relevant decision outcome affected by response mode effects, by studying the effects of response modes on the number of alternatives accepted using different response modes. Further, the response mode
continuum is helpful as a visual classification and representation of the effects of response modes on decision making processes.

Extending this line of research, Westenberg and Koele (1992) studied the effects of ranking, classifying into four classes, selecting, rejecting on decision processes and outcomes as applied to multiple alternatives (i.e., greater than two) each described by more than two attributes or dimensions. Further, they were interested in the moderating effects of task complexity or information load on the effects of these response modes on the process of decision making. Previous research had generally shown that compensatory decision making strategies were used only when the number of alternatives and/or attributes is small; otherwise a noncompensatory decision making strategy is more common (Ford, Schmitt, Schechtman, et al., 1989). Accordingly, they operationalized information load as a two-level variable (high and low) by using two choice sets of 15 and six alternatives per choice set respectively (Westenberg and Koele, 1992).

Using an information display board (Payne, 1976), Westenberg and Koele (1992) conducted a two-way analysis of variance with response mode as a between-subjects factor and information load as a within-subjects factor. The dependent variables were the same as the Billings and Scherer (1988) study reviewed earlier -- depth of search, variability of search, and information search pattern. Apart from these dependent variables, Westenberg and Koele (1992) also looked at another process variable -- depth of search per attribute -- in order to test the compatibility principle (as explained below) and a measure of decision outcome -- percentage of alternatives accepted.
In an explicit test of the compatibility principle (more specifically, scale compatibility), Westenberg and Koele (1992) represented each description of each attribute or dimension on three different types of scales, each compatible with at least one of the response modes investigated in the study: (a) a dichotomous scale, compatible with the selecting and rejecting response modes, (b) a grade scale with four grades, compatible with the classifying into four classes response mode, and (c) a rank scale, compatible with the ranking response mode (Westenberg and Koele, 1992).

Their results indicated that the response mode influenced the decision process as well as the decision outcome. Specifically, they found that ranking was clearly different from the other response modes since it was the only response mode which induced mainly dimension- or attribute-based information search patterns and compensatory decision making strategies. Classifying into four classes induced mainly alternative-based information search patterns and equal amounts of compensatory and noncompensatory decision making strategies. Rejecting induced only alternative-based information search patterns that consisted mainly of noncompensatory decision making strategies. Finally, selecting induced mainly alternative-based information search patterns, with mainly noncompensatory decision making strategies under high information load and with approximately equal amounts of compensatory and noncompensatory decision making strategies under low information load (Westenberg and Koele, 1992).

Further, the results of this study did not support the compatibility principle, since the compatible attributes or dimensions did not receive greater attention in terms of amount of information searched for than the
noncompatible attributes or dimensions in any of the response mode/information load combinations. Westenberg and Koele (1992) conjectured that this might have been due to the fact that their study -- unlike most other studies on response mode effects, including the studies on which the compatibility principle was founded upon -- used more than two attributes or dimensions to describe each alternative.

However, the results of this study were in agreement with the response mode continuum forwarded by Westenberg and Koele (1990), since there was more compensatory decision making as the response scale involved more scale points. However, this effect only held under conditions of high information load, with the order of response modes being (in order of increased compensatory decision making): rejecting, selecting, classifying into four classes, and ranking. Under conditions of low information load, the expected trend occurred only when the response mode selecting was omitted, since selecting induced an unexpectedly large amount of compensatory decision making strategies under conditions of low information load. Moreover, the discrepancy between rejecting and selecting that was observed in their previous study (Westenberg and Koele, 1990) was replicated in this study. Finally, with respect to decision outcomes, the response modes of selecting and rejecting accepted less alternatives than the response modes of ranking and classifying into four classes (Westenberg and Koele, 1992).
Summary

To summarize the preceding discussion, the literature on the effects of response mode effects on decision making has moved from a stage of denial to one of active problem solving and theoretical exposition. As noted, there are currently three major theoretical expositions of response mode effects. These three theoretical expositions of the response mode effect have linked the effects of response modes to either the psychophysical function ($H$ in Figure 3), the combination function ($C$ in Figure 3), or the judgment function ($J$ in Figure 3) within the framework of subjective measurement. However, as Goldstein and Einhorn (1987) have suggested, the veracity of the existence of procedural invariance and the phenomenon of preference reversal may be so robust because there are multiple underlying potential causes, each of which may be in operation in a given situation either alone or in combination with the others. Whatever the true explanation for preference reversal and response mode effects in general, as Payne, Bettman, and Johnson have observed, “... it is now clear that the answer to how much you like a decision option can depend greatly on how you ask the question” (1992, p.95).

In spite of the above discussion on response mode effects in general and judgment versus choice in particular, it should be recognized that there may be some instances in which the distinction between judgments and choices may be quite tenuous. As Bettman (1981) has argued, these are the cases in which judgment about an alternative is an integral part of the decision making process. Bettman (1981), Billings and Scherer (1988), and
Einhorn, Kleinmuntz, and Kleinmuntz (1979), have provided an account of some of these potential instances.

For instance, Bettman (1981) has identified four general classes of conditions in which choice proceeds with evaluative judgment of alternatives: (a) mode of presentation and type of information -- i.e., when the information required for decision making is memorized, information about each alternative is presented sequentially, and information is not easily comparable within attributes across brands; (b) importance and involvement -- i.e., when the decision situation is highly important and involving, (c) characteristics of the set of alternatives -- i.e., when the set of alternatives are relatively uncomplicated in terms of number and understandability, and (d) characteristics of the decision situation -- when the decision maker has ample time, little distraction, and when the decision requires a quantitative judgment (Bettman, 1981). Billings and Scherer (1988) have added two more instances when the difference between judgment and choice might be tenuous -- when the decision maker has thorough knowledge of the decision task and when decision situations involve sorting alternatives into more than one accept/reject categories.

**The Issue of Respondent Evaluations**

In order to be considered useful in terms of fulfilling their intended purpose, health status preference measurement methodologies must be economical to administer as well as acceptable to respondents. Economy of administration is an aspect which sponsors and/or investigators need to
determine for themselves on a case-by-case basis, since what might be economical in one situation might not be in another. On the other hand, acceptability to respondents is something which may be quite consistent across situations. However, there is a paucity of reliable and valid information about respondent evaluations of different preference measurement methodologies in the published literature. Given the importance of respondent evaluations, it is important to actually measure respondent evaluations for the different preference measurement methodologies before making a decision as to which particular methodology to adopt in any given study.

Indeed, Huber (1974b) has recommended that "... management scientists should generally make the choice ...[among different preference measurement methodologies] ... using acceptability of the model and the method ... as the major choice criterion" (p.1401). The logic behind this is simple -- if subjects respond unfavorably toward preference measurements, this has negative implications for the quality of the data and hence the quality of the inferences based on these data. This is especially so if respondents are unfavorably inclined toward one particular methodology as opposed to another. Along these lines, Fischer (1977) has specifically recommended, "If ... two [preference measurement] methods generate essentially equivalent values, then so much the better. But if the methods conflict, the decision maker should be asked to determine which method best reflects his true preferences" (p.314).

This recommendation is especially important to the present study, which investigated the construct validity of discrete choice conjoint
methodology by way of the principle of convergent validation. In this regard, even if different preference measurement methodologies are found to be reliable and valid as measures of health status preferences, there is nothing in the process of convergent validation that recommends one methodology over another, i.e., they are all valid as measures of health status preferences. Therefore, it is possible that respondent evaluations of the preference measurement methodologies may prove useful in deciding which one(s) of several alternative reliable and valid preference measurement methodologies should be selected for use in any particular application.

Since respondent evaluations of preference measurement methodologies was an important factor considered in this dissertation, it is necessary to evaluate the published literature which has compared different preference measurement methodologies in terms of respondent evaluations. This review will first focus on reviews comparing different measurement strategies and then move onto reviews comparing different scaling methods.

Comparing Measurement Strategies

The comparisons between the top-down and bottom-up strategies reported by Nutt (1980, 1981) had been reviewed earlier in this chapter. Both these studies also collected respondent evaluations of the different preference measurement strategies used in the studies. In his first study -- in which the ARS procedure was compared with a modification of SCJ analysis proposed by Gustafson, Pai, and Kramer (1971) -- Nutt (1980) collected respondent evaluations on the perceived accuracy and ease of use of the two strategies. Both these attributes were operationalized on a five-point Likert-
type scale, ranging from 1=best to 5=worst. It was found that the modified SCJ analysis was considered by respondents to be the most accurate but also the least easy to use. This finding is notable in that it suggests that ease of use and accuracy of methods need not go hand-in-hand -- in fact they may be tapping into distinct dimensions of overall respondent evaluations. Nutt (1980) also compared the different procedures in terms of overall time taken by respondents for completion. In this regard, the modified SCJ analysis took the longest time, while the ARS procedure took the shortest.

In his second study -- in which the ARS procedure was compared with regular SCJ analysis -- Nutt (1981) was more elaborate in the nature of the data collected for respondent evaluations of the different strategies. Three specific dimensions of respondent evaluations were measured for each strategy -- understandability, acceptance, and accuracy. Understandability was operationalized by two concepts -- clarity of instructions and understanding. Clarity of instructions was measured on a 0-100 point horizontal visual analog scale, anchored at 0=unclear to 100=clear-cut, while understanding was measured on a 0-100 point horizontal visual analog scale, anchored at 0=obscure to 100=clear-cut. In both cases, a middle anchor point was also provided, i.e., 50=mostly clear. Acceptance was operationalized by two concepts -- selection of procedures and acceptability. Selection of procedures was measured on a 0-100 point horizontal visual analog scale, anchored at 0=reject and 100=adopt, while acceptability was measured on a 0-100 point horizontal visual analog scale, anchored at 0=unacceptable to 100=fully acceptable. In both cases, a middle anchor point was also provided, i.e., 50=should try and 50=might try respectively. Finally,
accuracy was also operationalized by two concepts -- subjective opinions in predicting the criterion and relation between a particular input and the criterion. Subjective opinions in predicting the criterion were measured on a 0-100 point horizontal visual analog scale, anchored at 0=inaccurate and 100=precise, while relation between the particular input and the criterion was measured on a 0-100 point horizontal visual analog scale, anchored at 0=unclear relationship to 100=clear relationship. In both cases, a middle anchor point was also provided, i.e., 50=usable and 50=somewhat clear respectively (Nutt, 1981).

Respondent evaluations showed that study participants believed that SCJ analysis was more understandable (p<0.05), more acceptable (p<0.10), and more accurate (p<0.10) at both the overall (i.e., pooling both questions) as well as the individual question level. Further, in terms of time taken by respondents to complete the tasks, the ARS procedure required 1.33 to 2.00 hours, while SCJ analysis required between 0.42 and 0.58 hours. The time differences were also statistically significant (p<0.01). These results strongly favor SCJ analysis over the ARS procedure. Indeed, this was exactly what Nutt (1981) recommended.

Lyness and Cornelius (1982) reported a comparison between the holistic preference measurement strategy, the ARS procedure, and combination of these two procedures. This study was not reviewed earlier in this chapter since it did not include any statistically inferred preference measurement strategy as one of the comparators. After completing both tasks, subjects were asked to indicate the degree of confidence they had with each task on a seven-point rating scale. It was found that subjects had greater
confidence in the holistic judgments compared to both the ARS procedure as well as the combination procedure. Further, subjects also provided judgments using two response modes -- judgmental ratings and pricing judgments. The evaluations of these two response modes in terms of confidence (on the same scale as described above) indicated that subjects had more confidence in the judgmental ratings than in the pricing judgments. All these results were statistically significant. These results run counter to the divide and conquer principle of explicitly decomposed preference measurement strategies, since respondents had more confidence in the overall (or holistic) judgments as opposed to the decomposed judgments. Further, these results demonstrate that response mode effects also influence respondent evaluations of preference measurement methodologies.

Adelman, Sticha, and Donnell's (1984) comparison of the relative effectiveness of multiattribute weighting procedures as a function of the number of attributes used to describe the alternatives and the distribution of correct attribute preference scores was reviewed earlier in this chapter. As noted earlier, this was a comparison of the individual attribute preference scores obtained via SCJ analysis and the ARS procedure implemented using ratio weighting, paired comparisons, dividing 100 points among the attributes in relation to their importance, and the standard gamble scaling methods. Respondent evaluations of confidence and difficulty in using each of the procedures was obtained on seven-point Likert-type scales. There were significant differences found between the confidence scores for the five procedures (p<0.001) -- participants had the most confidence in dividing 100 points, followed by the ARS procedure implemented using ratio weighting,
SCJ analysis, paired comparisons, and finally the standard gamble in that order. In terms of difficulty in using each procedure, once again significant differences were found between the five procedures (p<0.001), with dividing 100 points being judged the easiest, followed by the ARS procedure implemented using ratio weighting, paired comparisons, SCJ analysis, and finally the standard gamble in that order. These results generally support the bottom-up strategies over the top-down strategies in terms of these particular dimensions of overall respondent evaluations.

Usher's (1991) study comparing the ARS procedure to conjoint methodology was discussed earlier in this chapter. In addition to a comparison of the two methodologies in terms of predictive ability, Usher (1991) also compared them in terms of respondent-assessed difficulty in performing the tasks required of the methodologies and confidence in the fact that the preference measurement methodologies had accurately captured their preferences. Difficulty in performing the tasks required of the methodologies was assessed with a five-point Likert-type scale, ranging from 1=very difficult to 5=very easy, while confidence in the fact that the preference measurement methodologies had accurately captured their preferences was also assessed using a five-point Likert-type scale, ranging from 1=very unconfident to 5=very confident. The results of this analysis favored the ARS procedure — subjects rated conjoint methodology as significantly more difficult (p<0.001) than the ARS procedure. Further, respondents were significantly less confident in the results of conjoint methodology as compared to the ARS procedure (p<0.001).
On the basis of the review of the above studies which have considered the issue of respondent evaluations of preference measurement methodologies, there does not appear to be any clear consensus on any dimension of overall respondent evaluation, at least in terms of measurement strategies. Some studies have found the top-down strategies to be more accurate and easy to understand than bottom-up strategies (for e.g., see Nutt, 1980, 1981), while others have found just the opposite result (for e.g., see Usher, 1991). The same lack of consensus holds for the attribute of confidence in measurement strategy, as shown by the results of Lyness and Cornelius (1982) whose respondents favored top-down strategies in this regard and Adelman, Sticha, and Donnell (1984), whose respondents favored bottom-up strategies in this regard. Since these studies used a wide variety of respondent groups, not even any tentative linkages can be suggested between evaluations and respondent groups. Naturally, this is an area which requires further research.

It is, however, interesting to note the results of Nutt’s (1980) study, wherein it was found that ease of use and accuracy do not necessarily go hand-in-hand. This suggests that the construct of overall evaluation may comprise of at least two different dimensions, one relating to the specific purpose of the methodology (i.e., how does the methodology fare in terms of fulfilling its purpose?) and the other relating to operational aspects of the methodology (i.e., how easy is it to actually perform the methodology?).
Comparing Scaling Methods

In his empirical comparison between the RS, SG, and TTO scaling methods, Torrance (1976a) compared the feasibility of these three methods. Feasibility was defined in terms of acceptability to subjects, ease of use for interviewers, and cost. Regarding acceptability to subjects, it was found that the percentage of respondents who broke-off interviews was quite low (at 2% overall -- no breakdown by scaling method was provided). However, subjects found the TTO questions to be the easiest to answer, the SG questions to be slightly more difficult to answer, and the RS questions the most difficult to answer by a substantial margin. However, no details were provided regarding how the respondent evaluations of difficulty were operationalized. Further, it should be remembered that Torrance (1976a) only administered the SG scaling method to highly educated non-elderly respondents. Indeed, he specifically noted that elderly, poorly educated subjects might experience difficulties with the SG scaling method. Regarding ease of use for interviewers, it was found that the professional interviewers used in this study found all three scaling methods easy to learn and straightforward to administer in the field, and had no clear preference for any one scaling method. Regarding cost, the SG and TTO scaling methods were the most expensive, since they required elaborate visual aids to help respondents understand the concepts that were fundamental to the scaling task, e.g., probabilities (Torrance, 1976a).

In their comparison between the RS and SG scaling methods, Bass, Steinberg, Pitt, et al., (1994) compared the total time taken by respondents to complete each scaling method. Although the RS scaling method took slightly
longer time to complete compared to the SG scaling method, this difference was not statistically significant.

Busschbach, Hessing, and de Charro (1993a) also collected respondent rankings of the difficulty in doing each of the four methods compared in their study -- the RS, SG, TTO, and RKI methods. Overall, it was found that the RS was the easiest, followed by the SG, then by the TTO, and finally by the RKI method. These researchers also measured the total amount of time taken by respondents to complete each of these methods. It was found that the SG and TTO methods were the shortest to complete, followed by the RS and finally by the RKI method. However, these researchers urged caution in interpreting these results in absolute terms, since the RS scaling method -- as implemented in this study -- also included additional questions about social and demographic characteristics of the respondents as well as information about respondents' own health states. This additional information was not requested in the other methods (Busschbach, Hessing, and de Charro, 1993a).

As part of their process tracing study on the cognitive processes underlying the phenomenon of preference reversal, Johnson, Payne, and Bettman (1988) collected information about respondent evaluations of the difficulty in doing the judgmental rating and pricing judgment tasks required of them. Respondent evaluations were collected on a 11-point scale (1=very easy, 11=very difficult). No significant differences were found in the respondent evaluations. However, it should be noted that these respondents took significantly longer to give pricing judgments as opposed to judgmental ratings (p<0.001). Thus, combining these results with those of
Lyness and Cornelius (1982), it appears that subjects might find pricing judgments and judgmental ratings equally easy or difficult to do, but may have greater confidence in the results of judgmental ratings as opposed to pricing judgments.

In another process tracing study on the cognitive processes underlying the phenomenon of preference reversal, Schkade and Johnson (1989) compared the total time taken by respondents to complete pricing judgments versus choices in one experiment and pricing judgments versus judgmental ratings in another experiment. In the first experiment, it was found that choosing between alternatives took significantly less time than offering pricing judgments for alternatives (p<0.01). These results are in keeping with the judgment versus choice distinction, wherein respondents usually make choices on the basis of considering fewer attributes or dimensions of alternatives than they do when making judgments about the same alternatives. In the second experiment, it was found that judgmental ratings took significantly less time to be elicited from respondents as opposed to pricing judgments (p<0.001). These results replicate the earlier ones of Johnson, Payne, and Bettman (1988) reviewed above.

The results of the studies reviewed in this section show a little more consistency than the results of studies comparing respondent evaluations of preference measurement strategies. Although there might appear to be some discrepancy between the results of Torrance (1976a) and Bass, Steinberg, Pitt, et al., (1994) on the one hand and Busschbach, Hessing, and de Charro (1993a) on the other hand, this can be explained in terms of the respondents of each of these studies. The respondents of Torrance's (1976a) and Bass,
Steinberg, Pitt, et al.'s (1994) studies were older than the university students who served as respondents for the Busschbach, Hessing, and de Charro (1993a) study. In this regard, university students may be more familiar with rating scales (simply because this is something they might be expected to learn about in school and also use in faculty evaluations) as opposed to older subjects. Therefore, it is possible that the reason for the discrepancy between the results of these studies is due to the nature of the respondents used for each study. The evaluation of rating scales by older patients can be tested in the present study, which used veterans as subjects.

Summary

In summary, it should be noted that respondent evaluations have not been given much attention in the preference measurement literature. In the few instances where they have been considered, they have only been operationalized by a few questions, which do not seem to do justice to the construct of overall evaluation. In this regard, the focus seems to have been placed on total time taken by respondents, with the implicit assumption that the longer time a particular methodology takes the worse it will be evaluated by respondents. Further, there does not seem to be a systematic attempt to infer the structure of overall respondent evaluations in the context of preference measurement methodology.
Putting It All Together

On the basis of this literature review and in view of the nature of the problem to be investigated in this dissertation (as presented in Chapter II), the specific results in the published literature can be linked to the research questions that drove this study; accordingly, empirical support for the research hypotheses can be provided.

Once again, in order to facilitate the comparison among the three preference measurement methodologies investigated in this study, it is instructive to view each methodology in terms of its component measurement strategies and scaling approaches. The conditional utility function-based strategy with a rating scale scaling method (CUF-RS) is a “bottom-up” measurement strategy with a rating scale scaling approach, the conditional utility function-based strategy with a standard gamble scaling method (CUF-SG) is a “bottom-up” measurement strategy with a standard gamble scaling approach while discrete choice conjoint methodology (DCCM) is a “top-down” measurement strategy with a pick-one scaling method.

Further, it is also instructive to differentiate between the components of the different scaling methods in terms of response modes and incorporation of attitudes toward risk. In this regard, the rating scale scaling method uses judgmental ratings as the response mode and measures preferences under certainty (i.e., measures values, not utilities). The standard gamble scaling method uses indifference judgments as the response mode and measures preferences under risk (i.e., measures utilities, not values). Finally, the pick-one scaling method uses choices as the response
mode and measures preferences under certainty (i.e., measures values, not utilities). These distinctions were presented in Table 1.

First, considering the CUF-RS and the CUF-SG, both adopt the same measurement strategy (i.e., bottom-up). The only difference between these two methodologies is in terms of the scaling method. In this regard, CUF-RS uses judgmental ratings and provides value-based preference scores, while CUF-SG uses indifference judgments and provides utility-based preference scores. Given the paucity of published information on response mode effects of indifference judgments versus judgmental ratings, no specific hypothesis can be forwarded. However, the published literature does support the hypothesis that utilities will have a higher score than values and that both scores will be highly correlated. Further, the published literature supports the existence of a power curve to describe the relationship between the two preference functions. Since the literature on concordance coefficients between these two measures is limited, no specific hypotheses can be forwarded in this regard. Finally, in terms of respondent evaluations, given the fact that veterans were used as subjects in this study, the literature supports the hypothesis that the SG scaling method will be considered easier to perform than the RS scaling method.

Next, considering the CUF-RS and the DCCM, the CUF-RS adopts a bottom-up measurement strategy, while the DCCM adopts a top-down measurement strategy. In terms of top-down versus bottom-up strategies, the literature supports the hypothesis that the DCCM preference scores for
multiattribute alternatives will be higher than those obtained using the CUF-RS methodology, although linear correlations between the two scores may be high. Since the two methodologies use different response modes, no particular hypothesis can be offered about their relative predictive abilities.

In terms of the scaling method, the CUF-RS methodology uses judgmental ratings while DCCM uses the pick-one method. Both methodologies yield value-based preferences. Since both methodologies use different response modes, one might expect to see some instances of preference reversal -- this would be reflected in a lower value Kendall's Tau coefficient for these two measures. In terms of functional relationships between these two value functions, on the basis of Fischer's (1977) a linear relationship would be expected. Finally, in terms of respondent evaluations, no specific hypotheses can be forwarded on account of the ambiguity in the literature comparing different preference measurement strategies. However, it is expected that a choice-based response mode will take less time to complete than a judgmental response mode.

Finally, with respect to the CUF-SG and the DCCM, the CUF-SG adopts a bottom-up measurement strategy, while the DCCM adopts a top-down measurement strategy. In terms of top-down versus bottom-up strategies, the literature supports the hypothesis that the DCCM preference scores for multiattribute alternatives will be higher than those obtained using the CUF-SG methodology, although linear correlations between the two scores may be high. However, since the CUF-SG incorporates the respondents' attitude toward risk, this would nullify this particular effect.
No particular hypothesis can be forwarded regarding which score would be higher, since there is not a large enough body of literature to predict which effect would prevail. Further, since the two methodologies use different response modes, no particular hypothesis can be offered about their relative predictive abilities.

In terms of the scaling method, the CUF-SG methodology uses indifference judgments while DCCM uses the pick-one method. The CUF-SG methodology yields utility-based preferences, while DCCM yields value-based preferences. Since both methodologies use different response modes, one might expect to see some instances of preference reversal -- this would be reflected in a lower value Kendall's Tau coefficient for these two measures. In terms of functional relationships between these two value functions, a power curve would be expected to describe the relationship between these two functions. Finally, in terms of respondent evaluations, no specific hypotheses can be forwarded on account of the ambiguity in the literature comparing different preference measurement strategies. However, it is expected that a choice-based response mode will take less time to complete than a judgmental response mode.

**Plan of the Remaining Chapters**

This chapter has reviewed the relevant published literature that pertains to the specific research questions which were investigated in this dissertation. The next chapter will detail the methodology followed in this study in order to address each of these questions. The chapter after that will
present the results of the study. Finally, the last chapter will discuss these results, study limitations, and the potential implications of this dissertation to health status preference measurement, health status measurement, as well as health-care decision making in general in an era of outcomes management.
CHAPTER V

METHODOLOGY

Chapter Overview

This chapter details the methodology followed in order to address the research questions which were presented in Chapter II. The chapter begins with a discussion of the conceptualization and operationalization of health status that was adopted for this study. This is followed by a discussion of the implementation of the three health status preference measurement methodologies which were used in the study, along with the instrumentation that was developed for data collection purposes for the three methodologies as well as for measuring respondent evaluations of the three methodologies. The implementation of the methodology is then discussed in terms of the university Human Subjects Committee approval procedures, clinic Research and Development Committee approval procedures, instrumentation development procedures, sample characteristics and sampling procedures, and data collection procedures. The chapter concludes with a discussion of the data analysis procedures used to address each of the research questions which was investigated in this study.
Conceptualization and Operationalization of Health Status

The conceptualization and operationalization of health status used in any study is a function of whether one uses an existing health status measurement instrument or develops such an instrument according to the specifications of the study. This decision will also address some of the outstanding conceptual and methodological issues in health status measurement outlined in Chapter I, such as the identification of dimensions of health status, the expression of different levels of each dimension in terms of functional capacity or performance, and the evaluation time frame. This is because these issues may already have been taken care of in existing instruments. If, however, the instrument is developed for the purpose of the study, these issues will have to be addressed in the process of development. For the purpose of this study, health status was conceptualized and operationalized in accordance with the EuroQol. This section will first discuss the rationale behind the selection of the EuroQol as the measure of health status and then discuss how some of the conceptual and methodological issues introduced in Chapter I were addressed for the purpose of this study.

Rationale Behind the Selection of the EuroQol

The first question that was addressed was whether an existing health status measurement instrument should be used or should a new conceptualization and operationalization of health status be developed for the purpose of this study. On the basis of a review of the most commonly used generic health status measurement instruments (see Chapter I), it was
decided to opt for an existing health status measurement instrument. This was because it was felt that most of the currently used instruments had content validity as health status measures -- at least according to Figure 1 -- and had satisfactory psychometric properties. This is in keeping with the advice of experienced health status measurement researchers, who have urged researchers to make an attempt to use the existing health status measurement instruments instead of re-inventing the wheel for every study (for e.g., see Bergner and Rothman, 1987).

Further, within the existing health status measurement instruments, it was decided to opt for a generic health status measure, as opposed to a disease-specific measure. The reasons for this are twofold: (a) more experience has been gained with generic as opposed to disease-specific measures in that the psychometric properties of generic measures have been subjected to more stringent investigation in the published literature and (b) it would then be possible -- in future applications of discrete choice conjoint methodology -- to compare preference functions of respondents having more than one specific disease or condition, in an attempt to understand differential valuation of health status dimensions as a function of disease state.

Although it was finally decided to conceptualize and operationalize health status in accordance with the EuroQol, it should be noted that other health status measurement instruments were also considered. It is instructive to briefly review the reasons why it was decided to not use these other health status measurement instruments so that the strengths of the EuroQol vis-à-vis the needs of this study can be appreciated.
Initially, it was considered using the SF-36 as a measure of health status for this study was because the SF-36 has been validated as a health status measure in a variety of practice settings with a variety of patients. Further, being a health status profile, it provides rich detail in terms of the effects of health-care interventions on the different dimensions or attributes of health status. However, health status profiles are less amenable to be used as-is for the purposes of constructing preference functions for health status. This is because they would need to be operationalized in terms of specific levels of functioning or performance for each dimension or attribute of health status in order to be used as measures of health status preferences. In this regard, health status indices are much better suited for the purposes of constructing preference functions for health status.

Although it is possible to "index" health status profiles by reconstructing them in terms of specific levels of health status dimensions, it was felt that such a reconstruction would result in serious deviation from the central purpose behind this dissertation, i.e., developing and validating discrete choice conjoint methodology as a measure of health status preferences. Therefore, it was decided to use an existing health status index as a measure of health status for the purpose of this dissertation. This should not be interpreted as a negation of the potential use of the SF-36 as a measure of health status preferences, since such an application of the SF-36 is certainly possible in the future -- albeit with some reconstruction in the current wording of the SF-36. Indeed, research is currently being conducted at the University of Sheffield on this very issue (see Brazier, 1995).
After having decided to use a health status index as the measure of health status for this study, the next question was which one of the three most prominent generic health status indices should be used — the Quality of Well-being (QWB) Scale, the McMaster Health Utilities Index (HUI), or the EuroQol? It was decided against selecting the QWB on account of two reasons: (a) the QWB does not explicitly account for the effects of mental functioning as a separate dimension of health status — the published literature on the choice of health status dimensions which need to be included in any measure of health status in hypertension (the condition investigated in this study) has identified mental functioning as an important dimension to be considered (for e.g., see Bulpitt and Fletcher, 1994) and (b) the QWB was constructed so as to support an additive model form, i.e., it specifically did not allow for interactions among health status dimensions. The importance of accounting for interactions among health status dimensions in any measure of health status preference was discussed in Chapter II, and will not be repeated here.

Unlike the QWB, both the Mark II and Mark III versions of the HUI account for the effects of mental functioning as a separate dimension of health status. Further, in empirical tests, the Mark II version of the HUI has been shown to support a multiplicative model form (Torrance, Zhang, Feeny, et al., 1992). Results of modeling the data collected using the Mark III version of the HUI have not yet been published. Still, at least the Mark II version of the HUI can account for interactions among health status dimensions.
The major problem (at least for the purpose of this study) with both versions of the McMaster HUI lies in the number of dimensions in the health status classification system adopted by each version. In this regard, the Mark II version had seven dimensions, while the Mark III version had eight dimensions. The problem with these number of dimensions for the present study was the specific computer package (i.e., the CBC System by Sawtooth Software) used to design and implement the double conditional design required by discrete choice conjoint methodology could deal with a maximum of six dimensions. Although this is not an insurmountable problem by itself (one could always combine different dimensions into a "mega-dimension"), there were some other drawbacks of the Mark II and III versions of the HUI which argued against their use in this study.

With respect to the other drawbacks of the two versions of the HUI, the Mark II version was developed for use in pediatric oncology patients. Although the developers claim that the Mark II version of the HUI is general in that it can be used in other patient populations as well, this claim has no empirical support. The Mark III version of the HUI, on the other hand, was developed for general population health surveys. However, given the paucity of published information regarding the psychometric properties of this instrument at the time this study was being designed (it should be noted that some information has been published since -- see, for e.g., Boyle, Furlong, Feeny, et al., 1995; Feeny, Furlong, Boyle, et al., 1995; Torrance, Furlong, Feeny, et al., 1995), it was decided against using this instrument in this study. Therefore, health status was conceptualized and operationalized
in this study according to the health status classification system of the EuroQol.

Conceptual and Methodological Issues in Health Status Measurement

Since the details about the conceptualization and operationalization of health status in the EuroQol have already been provided in Chapter I, they will not be repeated here. The health status classification system of the EuroQol is presented in Table 3. Apart from these particular levels of the health status dimensions, there is also a question pertaining to general health perceptions, comparing health status on the day of the interview to general level of health over the past 12 months (the response categories are “better,” “much the same,” and “worse”).

In discussing how the particular conceptual and methodological issues raised in Chapter I were addressed for the purpose of this study, the conceptual issues will first be addressed, followed by the methodological issues. It had been noted in Chapter I that there is no agreement among health status researchers as to what specific dimensions need to be included in any health status measurement instrument. Therefore, as suggested by Kirshner and Guyatt (1985), the adequacy of the specific health status dimensions used in any given study should be determined by the objectives of the study. Given the methodological focus of this study, it was felt that the major concern in deciding the adequacy of dimensions of the generic health status measure used in this study should be relevance to respondents. In this regard, there it is important that there should be some correspondence between the health status dimensions of the EuroQol and disease-specific
health status measures used in hypertension (the condition investigated in this study).

It can be seen from Table 3 that there is representation of physical functioning (mobility, self-care, pain/discomfort), mental functioning (anxiety/depression), social functioning (usual activities), and role functioning (usual activities) in the health status classification system of the EuroQol. This representation is consistent with the definitions of health status and health-related quality of life provided in Chapter I and illustrated in Figure 1.

Regarding hypertension, the health status dimensions mostly considered in most health status measures specifically developed for hypertension have been summarized by Bulpitt and Fletcher (1992) as measuring "... symptomatic well-being, psychological well-being and activity" (p.94-96). In terms of the terminology adopted in this dissertation, these descriptors roughly correspond to physiological status, mental functioning, and a catch-all category including physical functioning and regular day-to-day activities (e.g., role functioning and social functioning) respectively.

Thus, it can be seen that the health status dimensions of the EuroQol are consistent with those usually included in disease-specific health status measures for hypertension. However, as a check, it was decided to ask the respondents of this study whether the health status dimensions included in the study (i.e., the EuroQol) were sufficient to describe their health (as explained later in this chapter).
Table 3: Health Status Classification System of the EuroQol

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility</td>
<td>no problems in walking about</td>
</tr>
<tr>
<td></td>
<td>some problems in walking about</td>
</tr>
<tr>
<td></td>
<td>confined to bed</td>
</tr>
<tr>
<td>Self-Care</td>
<td>no problems with self-care</td>
</tr>
<tr>
<td></td>
<td>some problems with self-care</td>
</tr>
<tr>
<td></td>
<td>unable to wash or dress self</td>
</tr>
<tr>
<td>Usual Activities</td>
<td>no problems with performing usual activities (e.g., work, study,</td>
</tr>
<tr>
<td></td>
<td>housework, family or leisure activities)</td>
</tr>
<tr>
<td></td>
<td>some problems with performing usual activities</td>
</tr>
<tr>
<td></td>
<td>unable to perform usual activities</td>
</tr>
<tr>
<td>Pain/Discomfort</td>
<td>no pain or discomfort</td>
</tr>
<tr>
<td></td>
<td>moderate pain or discomfort</td>
</tr>
<tr>
<td></td>
<td>extreme pain or discomfort</td>
</tr>
<tr>
<td>Anxiety/Depression</td>
<td>not anxious or depressed</td>
</tr>
<tr>
<td></td>
<td>moderately anxious or depressed</td>
</tr>
<tr>
<td></td>
<td>extremely anxious or depressed</td>
</tr>
</tbody>
</table>
Regarding the methodological issues relevant to the design stage which were raised in Chapter I, the handling of most of these issues was largely determined by the choice of the health status classification system used in the study, i.e., the EuroQol. For instance, in keeping with the operationalization of health status in the EuroQol, the levels of each health status dimension were expressed in terms of functional capacity. Further, the evaluation time frame of the EuroQol is a period of one year -- this was the same evaluation time frame used in this study, in order to be consistent with the EuroQol and to enable comparisons of preference functions determined on the basis of the data collected in this study with the published literature on the EuroQol. Further, in keeping with the spirit of generic health status measures, no particular disease state or condition was specified in the question stems. This would also facilitate comparisons with future studies using discrete choice conjoint methodology in different disease states and conditions. In agreement with Torrance, Boyle, and Horwood's (1982) position on the issue, it was decided to aggregate individual preferences using the arithmetic mean as the measure of central tendency for reporting purposes. It should be noted that given the current state of the art of discrete choice conjoint methodology, it is not recommended to estimate parameters and preference functions at the individual level, as noted in Chapter III. Regarding the methodological issues relevant to the implementation stage which were raised in Chapter I (e.g., whom the respondent should be and the site of administration of the instrument), these will be addressed in the sections of this chapter which detail the sampling procedures and the data collection procedures respectively.
The data collection form which was used to operationalize the EuroQol for the purpose of this study is presented in Appendix A. The major change from the original EuroQol instrument that needs to be noted is in the instructions to respondents -- the word "tick" (to indicate how to respond to an item) was replaced with the word "check." This was because the word "tick" is more commonly used in this context in the UK as opposed to the US -- therefore, it was decided to replace this word with "check," which is more commonly used in this context in the US in order to avoid any possible ambiguity in instructions to respondents. Appendix A also contains the correspondence with the EuroQol group regarding permission to use the EuroQol health status measurement instrument for this study.

**Health Status Preference Measurement Strategies and Scaling Approaches**

The purpose of this study was to develop and validate discrete choice conjoint methodology as a measurement methodology for health status preference measurement. The principle of convergent validation as a means to support construct validity was adopted for this purpose. This study assessed construct validity by means of convergence of the health status preferences obtained as a result of the discrete choice conjoint methodology with the preference scores obtained by two other commonly used methodologies in health status preference research -- a conditional utility function-based strategy/standard gamble scaling method and a conditional utility function-based strategy/rating scale scaling method. The choice of these two particular methodologies reflects their popularity in the published
literature on health status preference measurement, and has been explained in Chapter II. Therefore, it will not be repeated here. This section will provide a detailed account of the specific procedures by which the selected methodologies were implemented in this study.

Two of the methodologies used in this study adopted the conditional utility function-based procedure as the measurement strategy. In discussing these two methodologies, the implementation of the conditional utility function-based procedure they shared will first be described, followed by the specifics of the rating scale scaling method and finally by the specifics of the standard gamble scaling method. After this, the implementation of the third methodology, discrete choice conjoint methodology, will be described.

Implementation of the Conditional Utility Function-based Procedures

As noted earlier, the conditional utility function-based procedure was developed by Keeney and Raiffa (1976). The general steps required to implement this procedure as Keeney and Raiffa (1976) described them were presented in Chapter III. The implementation of this strategy in this study was based on Torrance, Boyle, and Horwood's (1982) modification of Keeney and Raiffa's (1976) original implementation procedure. It is this modification which will be described in this section.

The first step in Keeney and Raiffa's (1976) original implementation involved checking independence assumptions in order to determine which model form -- additive, multiplicative, or multilinear -- is appropriate. As described by Keeney and Raiffa (1976), this step requires considerable interviewer-respondent interaction and is quite long. Indeed, in a published
demonstration of the actual dialogue between the interviewer and the respondent, it took about 8 hours to complete (Keeney, 1977)! Naturally, such a time commitment is quite prohibitive for use in large samples of patients in busy clinic settings.

However, Torrance and his colleagues at McMaster University have modified Keeney and Raiffa's (1976) original implementation by not carrying out this first step through personal interviews of the kind described by Keeney and Raiffa (Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992). Instead, they determined the appropriate model form on the basis of a few multiattribute evaluations which are also used to determine scaling constants for the models. This modification was also adopted for the purpose of this study. Details of this modification are provided below.

Further, Torrance and his colleagues at McMaster University have also modified Keeney and Raiffa's (1976) original implementation in terms of the determination of scaling constants (step 3 in the original implementation). The original implementation called for assessing preference functions directly in terms of values and utilities. This was done by determining scaling constants by having respondents holistically assess single attribute "corner states" (i.e., the five health states of the type where one health status dimension is at its most preferred level, while the four remaining dimensions are at their least preferred levels) for the construction of value and utility functions (depending on whether the scaling method encodes attitude toward risk or not). It should be noted that although other approaches are available to determine the scaling constants, the corner state
approach is the simplest mathematically, since with these states the scaling constant for any attribute is simply the preference score for the corner state with that particular attribute at its most preferred level (with the other attributes at their least preferred levels).

The additive model form is represented by:

$$ p(x) = \sum k_i p_i(x_i) ; $$  \hspace{1cm} (5.1)

and the multiplicative model form by:

$$ 1 + k p(x) = \prod \left[ 1 + k k_i p_i(x_i) \right] ; $$  \hspace{1cm} (5.2)

Rearranging the terms in Equation 5.2, one gets:

$$ p(x) = \left( \frac{1}{k} \right) \left\{ \prod \left[ 1 + k k_i p_i(x_i) \right] - 1 \right\} ; $$  \hspace{1cm} (5.3)

where $i = 1 \ldots 5$ (i.e., five health status dimensions in the EuroQol)

$x \in X = \text{multiattribute alternative}$

$k = \text{interaction parameter (-1 < k < 0 or k > 0)}$

$1 + k = \prod \left[ 1 + k k_i \right]$

$k_i = \text{scaling constant (0 \leq k_i \leq 1)}$

$x_i = \text{level of x on the i-th attribute}$

$p_i(x_i) = \text{i-th single attribute preference function}$

$p(x) = \text{overall preference function for multiatribute alternative}$
However, Torrance and his colleagues at McMaster University (see Torrance, Boyle, and Horwood, 1982) have argued that in health status preference measurement it is more meaningful to respondents to assess preference functions in terms of disvalues and disutilities -- instead of values and utilities -- where disvalues and disutilities are mathematically defined as the complement of values and utilities respectively. For instance, if value and utility are measured on a 0-1 scale, disvalue and disutility are simply 1 minus value and 1 minus utility respectively. The disvalue and disutility functions can then be converted into value and utility functions by taking their respective complements. In this regard, it has been argued that the disvalue and disutility formulations of health status preference functions are more natural and easier for respondents to understand since they use health states which are more feasible than those used by the value and utility formulations of health status preference functions. Specifically, the disvalue and disutility formulations of the corner states involve health states with only one attribute at its lowest level and all the other attributes at their highest levels (as opposed to only one attribute at its highest level and all the other attributes at their lowest levels in the value and utility formulations).

As implemented in this study, the conditional utility function-based procedure began with the assessment of single attribute preference functions using the appropriate scaling methods. Since the details of the implementation of the different scaling methods are provided in the sections following the present one, they will not be discussed here.

After assessing the single attribute preference functions, scaling constants were determined by having respondents holistically assess single
attribute corner states. In keeping with Torrance, Boyle, and Horwood's (1982) advice, it was decided to estimate the disvalue and disutility formulations of health status preference functions. In either case (i.e., using value/utility or disvalue/disutility formulations), the preference scores for the multiattribute corner states serve as the scaling constants -- the only difference would be that for the value/utility formulation, the scaling constants for any attribute would be the preference scores for the corner states in which the attribute in question was at its highest or most preferred level and the other attributes were at their lowest or least preferred levels, while for the disvalue/disutility formulation, the scaling constants for any attribute would be the preference scores for the corner states in which the attribute in question was at its lowest level and the other attributes were at their highest levels. Therefore, the mathematical convenience of using corner states remains unaffected whether the value/utility or disvalue/disutility formulation is employed.

Moreover, in order to assess the predictive ability of the methodology, respondent holistic evaluations of three multiattribute or multidimensional health states were also collected using both scaling methods. As opposed to the corner states used to assess the scaling constants, these multidimensional health states were not restricted to having only one attribute level at its worst and all others at their best -- individual attributes could take on any level of the classification system. In this regard, these multidimensional health states could be described as "interior states" (in contrast to corner states). Two of these interior states were selected at random from the 16 multidimensional health states used in the EuroQol instrument (states E3
and E8), while the remaining interior state was constructed to be a “backed off” corner state to allow for confounding effects between the dimensions of mobility and self-care (state BO).

The reason for the selection of this backed off corner state requires further explanation. As Torrance, Zhang, Feeny, et al. (1992) have noted, one of the problems with the use of pure corner states (i.e., where only one attribute takes on its best/worst level and all the other attributes take on their worst/best levels respectively) is that in some cases, these pure corner states may seem implausible to respondents being asked to provide preference evaluations. For example, it may be difficult for respondents to imagine being in a state of health of being in the best level of self-care (i.e., no problems with self-care) and worst level of mobility (i.e., confined to bed) as defined by the EuroQol. In order to correct for this possible problem, the backed off state BO was also used. The health state BO was operationalized as having the best levels of functioning on usual activities, pain/discomfort, and anxiety/depression, the middle level of functioning on self-care, and the worst level of functioning on mobility. When defined in such a manner, the pure corner state preferences for the two backed off attributes (which are needed for the purposes of determining the scaling constants in modeling) can be inferred by a process of extrapolation described by Torrance, Zhang, Feeny, et al. (1992). However, it was decided to use the process of extrapolation in this study only if a significant number of respondents indicated that they had difficulty in imagining the pure corner state for mobility. This would base the decision of whether to extrapolate or not on empirical evidence for the need to do so.
Specifically, the steps involved in this modified implementation of the conditional utility function-based procedure were:

(a) single attribute preference functions were first assessed using the appropriate scaling method (as detailed in the subsequent sections). The order of presentation of the five attributes was varied at random. It should be noted that these preferences were collected on value and utility scales, but were converted to disvalues and disutilities for the purposes of calculations by taking their respective complements,

(b) preferences for the five pure corner states and three interior states were then assessed using the appropriate scaling method. The order of presentation of these states was varied at random. These preferences were also collected on value and utility scales and converted to disvalues and disutilities for the purposes of calculations by taking their respective complements. It was necessary to convert the scores to disvalue and disutility scales since the corner states given to respondents were constructed to be those for the disvalue and disutility formulation of the preference functions. It should be noted that the three interior states (which served as the holdout sample) were interspersed with the pure corner states in order to avoid the possibility of position bias in the holdout samples, i.e., respondent fatigue needs to be considered if all the holdout states are placed at the end of the task,

(c) the appropriate model form -- additive or multiplicative -- was determined on the basis of an examination of the scaling constants. A multilinear model form was not considered on the basis of Keeney and Raiffa’s (1976) observation that such model forms occur rarely in practice;
further, the number of additional multiattribute health states that would need to be evaluated -- 30 in all, or 25 more than required for estimating additive and multiplicative model forms -- in order to estimate such model forms was considered excessive. For the purpose of estimating whether the additive or multiplicative model form was appropriate, the additive model form was considered appropriate if the sum of all the individual attribute scaling constants (i.e., $\sum k$) was equal to one; otherwise the multiplicative model form was considered appropriate,

(d) in cases where the multiplicative model form was considered appropriate, the interaction parameter $k$ also needs to be assessed. This parameter can be determined from the scaling constants ($k_i$) by an iterative solution of the following formula (Keeney, 1974):

$$1 + k = \prod [1 + k_i]; i = 1 \ldots 5$$  \hspace{1cm} (5.4)

The sum of the five scaling constants indicates the search interval for the iterative solution -- if this sum is greater than one, then the interaction parameter has a value between -1 and 0; if this sum is less than one, then the interaction parameter has a value greater than 0 (if this sum is equal to one then the interaction parameter has a value of 0, indicating an additive model form). Solving this equation iteratively for 140 respondents is quite time-consuming and also prone to human error if done manually. Therefore, it was decided to use a computer package to get the value of the interaction parameter. The particular computer package used was SOLVE (Mundt, 1989), which is a program written in Turbo Pascal (version 3.0) and run on
IBM-compatible computers using CGA graphics. This program is specifically designed to derive multiplicative preference functions and can estimate the interaction parameter $k$ in a few seconds using the scaling constants as input data.

(e) preference scores were calculated for each interior health state using the appropriate model forms determined as above,

(f) the calculated preference scores from (f) above were compared to the assessed preference scores for these interior states from (b) above to determine the predictive ability of the methodologies.

NOTE: steps (a) and (b) are the only ones which differed in implementation in case of the rating scale and standard gamble scaling methods; all the other steps were the same for the rating scale and standard gamble scaling methods.

The conditional utility function-based procedure described above was implemented with two different scaling methods in this study -- a rating scale scaling method and a standard gamble scaling method. Extensive use of visual aids was employed in the implementation of the rating scale and standard gamble scaling methods in this study. Therefore, before describing the implementation of the rating scale and standard gamble scaling methods in this study, it is necessary to describe the visual aids used in their respective implementations.
Visual Aids used with Rating Scale and Standard Gamble Scaling Methods

As noted earlier, extensive use of visual aids was employed in the implementation of the rating scale and standard gamble scaling methods in this study. In this regard, visual aids in the form of (3” x 4.5”) index cards were developed for the single and multiattribute health state descriptions and pointers (or arrows) representing these single and multiattribute health states, while existing visual aids were used as the actual instruments for operationalizing the rating scale and standard gamble scaling methods. Each of these types of visual aids will be described in turn.

Regarding the (3” x 4.5”) index cards describing single and multiattribute health states, these cards were used in the operationalization of both, the rating scale and standard gamble scaling methods. Separate cards were made for each level of each single attribute and for each multiattribute health state. Therefore, there were a total of 15 cards for the levels of each single attribute (5 attributes, each with three levels) and 11 cards for the multiattribute health states (1 card for the all-best or perfect health state, 1 card for the all-worst health state, 1 card for immediate death, 5 cards for each of the corner states and 3 cards for each of the interior states). It should be noted that the all-best health state is the state which has the highest or most preferred levels on all the five attributes, while the all-worst health state is the state which has the lowest or least preferred levels on all the five attributes. Each of the 26 cards generated is presented in Appendix B. A sample card is shown in Figure 6.
some problems walking about
no problems with self-care
some problems performing usual activities
extreme pain or discomfort
moderately anxious or depressed

Figure 6: Sample Card Describing a Multiattribute Health State

A few points about these cards need to be noted. First, their size -- i.e., (3” x 4.5”) -- was dictated by logistical considerations, i.e., they needed to fit within the pockets of the chance board used to operationalize the standard gamble scaling method, as described later. As it turned out, this was not an issue since -- on the basis of input from subjects who participated in the pre-test -- it was decided to place the cards on not in each pocket (details are provided in a subsequent section of this chapter). Second, the code in the top right-hand corner of the card (E3 in the sample card shown in Figure 6 above) was used as an identification label for the arrows which were made to correspond to each of the cards. Finally, the order in which the five
attributes appeared in each multiattribute health state cards was fixed as follows -- mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. It should be noted that the order of appearance of these attributes could have been randomized, but it was decided against doing so in order to be consistent with the manner in which multiattribute health states were presented by the EuroQol group of researchers in their published applications of health status measurement using the EuroQol. Indeed, all the groups whose research on health status preference measurement has been reviewed have used a fixed presentation of attributes in the descriptions of multiattribute health states.

Still, the issue of whether fixed versus randomized presentation of attributes would significantly affect preference scores for multiattribute health states is an interesting one and could be a factor to consider if respondents, due to information overload considerations, focus their attention on the first few attributes to the detriment of the last few attributes in making their evaluations. Unfortunately, there is little empirical research on this issue in the published health status measurement literature.

Regarding the **pointers (or arrows)** used to represent the single and multiattribute health state descriptions, these were only used in the operationalization of the rating scale scaling method. These were shaped as regular one-headed arrows, about 5” in length, about 0.6” in width at the shaft of the arrow and about 1” in width at the head of the arrow. The middle of the shaft of the arrows was clearly labeled with the identification corresponding to the card which the arrows were supposed to represent (E3 for the sample card shown in Figure 6 above). Three short velcro strips were
stuck at the back of each arrow (one at the base, one in the middle of the
shaft, and the other at the top of the shaft) in order to help the arrows stick to
the felt surface of the feeling thermometer (i.e., the actual instrument which
was used to operationalize the rating scale method).

Regarding the existing visual aid used as the actual instrument to
operationalize the rating scale scaling method, the **Feeling Thermometer**
developed by the McMaster group of researchers was used for this purpose
(Furlong, Feeny, Torrance, et al., 1990). Furlong, Feeny, Torrance, et al.,
(1990) have noted that this visual aid was called a Feeling Thermometer
because "... it helps subjects to measure their feelings ... for different health
states" (p.24). The Feeling Thermometer consists of a simple interval scale
ranging from 0 to 100, as a high contrast photographic line print on resin-
coated paper. A non-slip felt background is laid out on one side of the scale
in order to increase friction and prevent the pointers (or arrows) from
accidentally sliding down the scale (Furlong, Feeny, Torrance, et al., 1990).
The felt and scale are mounted on a stiff piece of reinforced foamboard (15”
x 26”). The top and the bottom of the scale were labeled "Best Imaginable
Health State" and "Worst Imaginable Health State" respectively, in
accordance with the anchors defined in the EuroQol health status
measurement instrument. A pictorial representation of the Feeling
Thermometer used in this study is provided in Figure 7. This representation
is adapted from Furlong, Feeny, Torrance, et al. (1990), in that the anchors
are labeled "Best Imaginable Health State" and "Worst Imaginable Health
State" respectively, as explained above [Furlong, Feeny, Torrance, et al.
(1990) used "Most Desirable" and "Least Desirable" as the anchors].
NOTE: (a) top of scale was labeled BEST IMAGINABLE HEALTH STATE
(b) bottom of scale was labeled WORST IMAGINABLE HEALTH STATE
(c) scale was more finely divided than shown (see text for details)

Figure 7: Feeling Thermometer Used in Present Study
[adapted from Furlong, Feeny, Torrance, et al. (1990)]
It should be noted that the Feeling Thermometer used in this study was a slightly different operationalization of a rating scale scaling method compared to the operationalization by the EuroQol group of researchers. In this regard, the EuroQol operationalization was a 0-100 point rating scale with clearly marked divisions of 0, 10, 20, 30, ... 80, 90, 100 (i.e., every 10th point). On the other hand, the Feeling Thermometer used in this study was a 0-100 point rating scale with clearly marked divisions of 0, 1, 2, 3, ... 97, 98, 99, 100 (i.e., every single point).

Regarding the existing visual aid used as the actual instrument to operationalize the standard gamble scaling method, the Chance Board developed by the McMaster group of researchers was used for this purpose (Furlong, Feeny, Torrance, et al., 1990). This board elicits utilities based on the probability equivalence standard gamble scaling method. The chance board is a square-shaped board (13" x 13") which is divided into two halves. The top half is labeled “Choice A,” and contains information relevant to the risky option. The bottom half is labeled “Choice B,” and contains information relevant to the sure-thing option. There is also an adjustable wheel, which changes the probabilities of the occurrence of each component of the risky option.

The chance board simultaneously presents the probabilities of the two risky alternatives and the one certain alternative associated with the two options presented to respondents. Presentation of both risky alternatives may appear to be redundant, but it reduces the potential for the occurrence of the framing bias that may be associated with using only one state that may have a positive or negative connotation to the respondent (Read, Quinn,

The developers of the Chance Board have noted that the layout of the sequence of probabilities -- in a converging ping-pong (i.e., alternating back-and-forth between high and low probabilities) manner -- was dictated by the need to reduce the possibility of: (a) an anchoring bias, whereby a strategy of constantly increasing/decreasing probabilities may encourage respondents to overestimate or underestimate their true indifference point and (b) a framing effect associated with consistently increasing or decreasing probabilities which are often interpreted by respondents as gains or losses relative to a reference point (Furlong, Feeny, Torrance, et al., 1990). A pictorial representation of the Chance Board used in this study is provided in Figure 8.

Implementation of the Rating Scale Scaling Method

As noted, the conditional utility function-based procedure was implemented using visual aids with two different scaling methods in this study -- a rating scale scaling method and a standard gamble scaling method. This section will detail the implementation of the rating scale scaling method in this study. The visual aids used in the implementation of the rating scale scaling method in this study included the (3” x 4.5”) descriptive cards, the arrows corresponding to these cards, and the Feeling Thermometer.
NOTE: (a) top anchor = M1, S1, U1, P1, A1, or AB as appropriate
(b) bottom anchor = M3, S3, U3, P3, A3, or ID as appropriate
(c) see Appendix B for detailed descriptions of levels and states

Figure 8: Chance Board Used in Present Study
[adapted from Patrick and Erickson (1993)]
The first task in the implementation of the rating scale scaling method was explaining what the Feeling Thermometer was and how it worked (i.e., what the anchors represented and how the scale needed to be used). After this, respondents were made to undergo a trial run, with trial (3” x 4.5”) descriptive cards and trial arrows corresponding to these trial cards, in order to make sure they understood how to use the Feeling Thermometer. Such trial runs have been shown to increase the reliability of the data collected (Anderson, 1981, 1982). The trial set consisted of three cards (with corresponding arrows). Respondents were given the three cards and arrows together and asked: (a) to place the arrow corresponding to the card they thought was the “best imaginable” at the top of the scale at the 100-mark, (b) to place the arrow corresponding to the card they thought was the “worst imaginable” at the bottom of the scale at the 0-mark, and (c) to place the remaining arrow at any point along the scale so that it reflected their evaluation of that level with respect to the levels already placed on the scale (i.e., with respect to the two anchors of the best and worst imaginable levels). This last step was necessary to obtain interval-level data for the valuations. At the end of this trial run, all the arrows were removed from the Feeling Thermometer in preparation for the study levels and states to be evaluated. Scores were not recorded for this trial run.

The respondents were then given the five single attribute sets of (3” x 4.5”) descriptive cards and arrows corresponding to these cards. Each single attribute set was given one at a time. For each set, respondents were given the three cards and arrows together and asked: (a) to place the level they thought was the “best imaginable” at the top of the scale at the 100-mark, (b)
to place the level they thought was the "worst imaginable" at the bottom of the scale at the 0-mark, and (c) to place the remaining level at any point along the scale so that it reflected their evaluation of that level with respect to the levels already placed on the scale (i.e., with respect to the two anchors of the best and worst imaginable levels). As above, this last step was necessary to obtain interval-level data for the valuations. As noted earlier, the order in which each single attribute set of cards and arrows was given to respondents was varied at random. Scores were recorded in the data collection form after each individual set of cards was evaluated by respondents. Further, the Feeling Thermometer was cleared after each set of evaluations, so that each new set began on a clean (i.e., without any arrows) board.

After the respondents evaluated each of the five single attribute sets on the Feeling Thermometer, they were given the cards representing the all-best, all-worst, and immediate death states (along with their corresponding arrows). They were asked to place the all-best state at the top of the scale (i.e., at the "best imaginable health state" or 100-mark). Regarding the other two states, patients were asked to place the "worst imaginable" of these two states at the bottom of the scale (i.e., at the 0-mark), and the remaining state at any point along the scale so that it reflected their evaluation of that state with respect to the levels already placed on the scale (i.e., with respect to the two anchors of the best and worst imaginable states). These arrows were not removed from the board while the respondents completed the subsequent evaluations -- indeed, they served as the reference states for the subsequent evaluations.
After this, respondents were given the sets of five corner states and three interior states to evaluate. The order in which these states was given to them was varied at random. Therefore, the interior states (which served as the holdout sample) were interspersed with the corner states. The valuations of these multiattribute health states was recorded in the data collection forms. It should be noted that for all evaluations, respondents were instructed to rate the multiattribute health states in relation to the other health states already on the Feeling Thermometer -- as before, this was necessary to get interval-level data for the multiattribute health state valuations.

A detailed account of the instructions for the implementation of the rating scale scaling method is provided in the interviewer schedule presented in Appendix C. A copy of the data collection form used for the implementation of the rating scale method is presented in Appendix D. It should also be noted that after completing the rating scale scaling method, respondents were asked to provide an evaluation of this methodology. Since the evaluation form was the same for all the three methodologies, the specifics of this evaluation form will be discussed after the implementation of the three methodologies is described.

In summary, the specific steps involved in the implementation of the rating scale scaling method were:
(a) explaining what the Feeling Thermometer was and how it worked,
(b) doing a trial run with respondents,
(c) rating the all-best, all-worst, and immediate death states,
(d) rating the five single attributes in random order,
(e) rating the five corner states and three interior states in random order.

Implementation of the Standard Gamble Scaling Method

As noted earlier, the conditional utility function-based procedure was also implemented using a standard gamble scaling method. This section will detail the implementation of the standard gamble scaling method in this study. The visual aids used in the implementation of the standard gamble scaling method in this study included the (3" x 4.5") descriptive cards and the Chance Board.

The first task in the implementation of the standard gamble scaling method was explaining what the Chance Board was and how it worked (i.e., explaining the concept of probability and the nature of the gamble). As for the implementation of the rating scale scaling method, respondents were also made to undergo a trial run for the standard gamble scaling method -- with the same trial (3" x 4.5") descriptive cards used for the trial run of the rating scale scaling method -- in order to make sure they understood how to use the Chance Board. As noted, such trial runs have been shown to increase the reliability of the data collected (Anderson, 1981, 1982). Once again, scores were not recorded for the trial run.

After the trial run, utilities for the individual attributes were determined. As noted earlier, these attributes were presented to respondents in random order. For each set (where a set consisted of three cards denoting the three levels of each attribute), the highest or most preferred level which was identified in the rating scale scaling method (i.e., the level which was
given a score of 100 in the rating scale method) was scaled as 1.0 and the lowest or least preferred level which was identified in the rating scale method (i.e., the level which was given a score of 0 in the rating scale method) was scaled at 0.0. In other words, these were the two levels which formed the alternatives in "Choice A" of the gamble. The remaining level of the set of three levels formed the alternative which comprised "Choice B" of the gamble (i.e., the sure-thing). After each set, scores were recorded in the data collection form.

After assessing the single attribute utilities, respondents were made to evaluate the multiattribute health states. For this evaluation, the all-best state was scaled at 1.0 and the immediate death state was scaled at 0.0. In other words, these were the two states which formed the alternatives in "Choice A" of the gamble. The remaining states were randomly used as the alternatives which comprised "Choice B" of the gamble (i.e., the sure-thing). As in the case of the rating scale scaling method, the interior states were interspersed with the five pure corner states. Scores were recorded in the data collection form after each individual state was evaluated.

It should be noted that no attempt was made to measure utilities for health states considered worse than immediate death. This decision was made on the basis of several factors. First, in Torrance, Boyle, and Horwood's (1982) study which measured utilities for health states worse than death, the unbounded nature of the negative utilities (since they had no lower bound) compared to the bounded nature of the positive utilities (since they had an upper bound of 1.0) caused some concern and required ad hoc adjustments. Second, Torrance, Zhang, Feeny, et al. (1992) have noted that
from a clinical policy perspective, it is possible that such health states may be considered to be qualitatively different from those preferred to immediate death and thus need to be treated in a special way instead of simply extending the scale below immediate death. Thus, in keeping with Torrance, Zhang, Feeny, et al.'s (1992) position, it was decided not to measure negative utilities for health states worse than immediate death. Still, it should be noted that there were several instances in which respondents expressed a preference for immediate death as compared to some other health states. This issue will be further discussed in Chapter VII.

A detailed account of the instructions for the implementation of the standard gamble scaling method is provided in the interviewer schedule presented in Appendix C. A copy of the data collection form used for the implementation of the standard gamble method is presented in Appendix E. It should also be noted that after completing the standard gamble scaling method, respondents were asked to provide an evaluation of this methodology. Since the evaluation form was the same for all the three methodologies, the specifics of this evaluation form will be discussed after the implementation of the three methodologies is described.

In summary, the specific steps involved in the implementation of the standard gamble scaling method were:
(a) explaining what the Chance Board was and how it worked,
(b) doing a trial run with respondents,
(c) evaluating the five single attributes in random order,
(d) evaluating the five corner states and three interior states in random order.

Implementation of Discrete Choice Conjoint Methodology

As noted earlier, this study implemented discrete choice conjoint methodology using the CBC System developed by Sawtooth Software, Inc. (Mete grano, 1993). The acronym CBC stands for Choice-Based Conjoint. The CBC System is relatively new, having been introduced in December 1993. Its principles are based on the work of Louviere and his colleagues (for e.g., see Louviere and Woodworth, 1983), who pioneered the design and implementation of discrete choice conjoint strategies in transportation economics and marketing research, as explained in Chapter III.

The CBC System handles the design, data collection, and data analysis steps in the implementation of discrete choice conjoint methodology. In terms of the design, it allows the user a choice between two options: (a) a fixed orthogonal design, which has the advantage of maximum efficiency of measuring main effects and the particular interactions for which it was designed and (b) a randomly constructed design, which is slightly less efficient than (a) above but has the offsetting advantage that all interactions can be measured. For the purpose of this study, the choice between these two options was based on theoretical justification, user requirements, and specific needs of the study.

In case of theoretical justification, there is no clear favorite between these two designs since optimally efficient designs for discrete choice conjoint strategies remain elusive (Batsell and Louviere, 1991) in spite of
much research on this issue. Indeed, a state-of-the-art review on discrete choice conjoint modeling concluded, "... despite the progress made in understanding how to design choice experiments, in many cases the estimation efficiency of the resulting designs is unknown" (Batsell and Louviere, 1991 p.207).

Further, the same review continued that in the case of the simple multinomial logit (MNL) choice model, several designs were consistent with the statistical assumptions. In this regard, one could use (a) balanced incomplete block designs (including the 2^N designs discussed earlier) to assign a set of designed alternatives to choice sets, (b) random assignment (without replacement) of statistically equivalent sets of profiles to sets, or (c) orthogonal, fractional factorial arrays to design alternatives and choice sets simultaneously. This review concluded that the statistical efficiency properties of designing discrete choice conjoint strategies in these ways was not yet fully understood (Batsell and Louviere, 1991).

In case of user requirements, the fixed design needs to be designed by the user and programmed into the CBC System; on the other hand, the random design is automatically constructed by the CBC System. Therefore, in terms of this criterion the random design would be favored over the fixed design.

Finally, in case of the specific needs of the study, the importance of accounting for the effects of interactions among health status dimensions while measuring health status preferences was discussed in Chapter II. In this regard also, the random design would be favored over the fixed design. Further, it is preferable to have as many unique alternatives (i.e., health
states) evaluated by respondents as possible, so that when the analysis is conducted at the aggregate level it is more representative of the alternatives under evaluation. Once again, the random design is favored over the fixed design for this particular issue. On the basis of these reasons, it was decided to take advantage of the CBC System's capability of randomly constructed designs for designing the choice sets and the alternatives placed within each choice set used in this study.

Further, within the randomly constructed design, there are two types of choice set construction strategies available in the CBC System -- complete enumeration or a simpler shortcut strategy. The complete enumeration strategy considers all possible alternatives and chooses each one so as to produce the most nearly orthogonal design for each respondent, in terms of main effects and all two-way interactions. Further, the alternatives within each choice set are kept as different as possible. The shortcut strategy attempts to construct each alternative by choosing attribute levels used least frequently in previous alternatives for each particular respondent. However, unlike the complete enumeration strategy which keeps track of co-occurrences of all pairs of attribute levels, the shortcut strategy considers attributes one-at-a-time. On account of this, it was decided to use the complete enumeration strategy for construction of the choice sets. It should be noted that the complete enumeration strategy is recommended over the shortcut strategy by the CBC manual (Metegrano, 1993).

In terms of data collection, the CBC System offers the flexibility of printing out the choice sets so that they can be used in regular paper-pencil written surveys or administering the constructed choice sets in an interactive
computerized survey. Since it was decided to use the randomly constructed design strategy, it was considered prohibitive to administer a regular paper-pencil written survey simply because each respondent views a unique combination of choice sets -- therefore, it regular paper-pencil written surveys were to be used, a separate survey would have to be printed out for each respondent. Therefore, for the purpose of this study, it was decided to collect the data for the discrete choice conjoint methodology by means of an interactive computerized survey. It should be noted that since the implementation of the conditional utility function-based procedure did not use computers for the data collection process, the possibility of pro- or anti-computer bias could have affected the respondent evaluations of the different methodologies. This issue will be discussed in greater length in Chapter VII.

In terms of data analysis, the CBC System assumes that the MNL choice model represents the underlying decision making processes of respondents and uses maximum likelihood as the estimation procedure to estimate the parameters of the MNL choice model. This is consistent with the advice offered by Bunch and Batsell (1989), which was reviewed in Chapter III. It should be noted that the CBC System also offers a much simpler way of analyzing the data which involves only frequency counts of the occasions when every single level of each attribute and every two-way combination of levels was chosen by respondents. These two counts would correspond to the main effects and two-way interactions in any model. However, on account of the fact that this estimation method lacks the theoretical rigor of the MNL choice model (which was -- as noted earlier -- one of the major
attractions of the discrete choice conjoint methodology), it was decided against analyzing the data by this simpler method.

A few limitations and delimitations of the CBC System need to be noted. In using the CBC System, the researcher is required to decide on the attributes or dimensions and their levels. The CBC System cannot handle more than six attributes or dimensions and more than nine levels per dimension (this, however, is not a problem with the health status classification of the EuroQol, which has five attributes or dimensions, each with three levels). Each choice set can contain between two and nine alternatives (or eight alternatives plus a base alternative like a “none of the above” option). It is permissible to use choice sets of different sizes, so as to allow for the effect of varying choice set size. Each interview can contain upto 50 choice sets. The respondent's task is to select one alternative from each choice set. Therefore, the scaling approach is indirect, yielding nominal-level data which are subsequently transformed in the analysis. The specific scaling method used is the pick-one scaling method.

Since it was decided to use the randomly constructed designs, three major decisions needed to be made -- how many alternatives to show per choice set, should there be a “none” alternative, and how many choice sets to show per interview. Each of these three issues will be discussed in turn. The overriding concern in making these decisions was balancing the need to have sufficient information from each respondent so that the sample size needed for estimation of model parameters would not be excessively prohibitive against the threat of respondent fatigue due to information
overload on account of too many alternatives per choice set as well as too many choice sets per interview.

Regarding the number of alternatives to show per choice set and the related question of whether there should be a "none" alternative, the published literature has not reached a consensus on recommendations about these issues. As noted earlier, the CBC System allows a maximum of eight alternatives per choice set. The number of alternatives per choice set in published studies has ranged from two (in binary choice modeling) to up to eight (see Billings and Scherer, 1988), although most published studies have used between three and five alternatives per choice set in order to avoid information overload effects. In accordance with the majority of studies, it was decided to use three multiattribute alternatives per choice set for the purpose of this study. The decision to use three multiattribute alternatives per choice set was also prompted by the fact that the random design adopted would ensure that each level of each attribute appeared once in each choice set if three multiattribute alternatives were shown in each choice set.

On a related issue, it was decided to include a "none" alternative in each choice set. The major reason for this was in order to allow for the inclusion of immediate death in the health status classification system, so as to make meaningful comparisons between the results of the discrete choice conjoint strategy and the conditional utility function-based procedure. Therefore, each choice set included the alternative of "immediate death" in addition to three multiattribute alternatives.

However, in order to get some input from potential respondents about the number of alternatives to be shown per choice set, test subjects in
the pre-testing and pilot testing of the instrumentation were asked whether they felt the number of alternatives shown in each choice set was "too few," "about right," or "too many." Since all but one of the 18 test subjects felt that the alternatives per choice set shown in the pre-test and pilot tests were "about right," it was decided to keep the number of alternatives per choice set fixed at three multiattribute alternatives and "immediate death."

Finally, regarding how many choice sets to show per interview, as noted earlier, the CBC System allows a maximum of 50 choice sets per interview. However, this number was considered prohibitively large for any single interview. Although there is no consensus in the published literature on this issue, the CBC System manual has noted that the studies that were examined and conducted while writing the manual and developing the program used between 6 and 16 choice sets per interview (Metegrano, 1993). On the basis of this figure, it was decided to include 12 choice sets per interview, out of which 10 would serve as the estimation sample and the remaining two would serve as the holdout sample to determine the predictive ability of the estimated model.

This last decision brings up the issue of the construction of the holdout choice sets. As noted earlier, two holdout choice sets were included per interview. It was decided to intersperse these two holdout choice sets among the ten choice sets which formed the estimation sample (this was comparable to how the holdout multiattribute health state sample was handled in the case of the implementation of the conditional utility function-based procedure). Specifically, the third and the eighth choice sets were the two holdout choice sets. The multiattribute alternatives which were included
in these two holdout choice sets also need some explanation. The first holdout choice set (i.e., the third choice set shown in the interview) included the three multiattribute interior states used in the conditional utility function-based procedure -- states BO, E3, and E8. The second holdout choice set (i.e., the eighth choice set shown in the interview) included three multiattribute health states selected at random from the total universe of 243 possible health states according to the health status descriptive classification system of the EuroQol. It should be noted that each of the holdout choice sets also had “immediate death” as an alternative.

A detailed account of the instructions for the implementation of the discrete choice conjoint methodology is provided in the interviewer schedule presented in Appendix C. A sample interview including the information screens as well as the choice sets shown to respondents is presented in Appendix F. It should also be noted that after completing the discrete choice conjoint methodology, respondents were asked to provide an evaluation of this methodology. Since the evaluation form was the same for all the three methodologies, the specifics of this evaluation form will be discussed after the implementation of the three methodologies is described.

Specifically, the steps involved in the implementation of the discrete choice conjoint strategy in this study were:

(a) interview information screens were constructed, which explained the purpose of the study and the choice tasks to respondents. Further, these information screens also provided respondents with instructions on how to perform the choice tasks,
(b) twelve choice sets were constructed -- ten for the estimation sample and two for the holdout sample. The randomly constructed design was used for the ten choice sets of the estimation sample (the complete enumeration strategy was used to construct the alternatives to be placed within each choice set), while the design of the two choice sets of the holdout sample was fixed (as explained before),

(c) data were collected using interactive computerized interviews,

(d) respondents were first made to undergo a trial run, in order to familiarize them with the task. The particular scaling method adopted was the pick-one scaling method,

(e) data from the estimation sample were analyzed using maximum likelihood estimation to determine the parameters of the MNL choice model assumed to represent respondents' decision making processes. In order to determine the appropriate model form, stepwise logistic regression was used,

(f) descriptions of the alternatives included in the two holdout choice sets were input into a share simulator (which estimates the proportion of respondents who would choose each of the alternatives) constructed on the basis of the model form estimated in (e) above,

(g) the simulated shares from (f) above were compared with the actual shares of the two holdout choice sets in order to determine the predictive ability of the methodology.
Respondent Evaluations

The importance of respondent evaluations in comparing health status preference measurement methodologies was discussed in Chapter IV. As noted in that chapter, most published studies have operationalized respondent evaluations by only a few questions, which do not seem to do justice to the construct of overall evaluation. Further, it was noted that the emphasis in the published literature seemed to be on total time taken by respondents, with the implicit assumption that the longer time a respondent takes to complete the methodology the worse it will be evaluated by respondents. Finally, it was noted that there did not seem to be any systematic attempt to infer the structure of overall respondent evaluations in the context of preference measurement methodology. This section discusses the approach adopted in this study in order to construct an instrument to measure respondent evaluations and address some of the shortcomings identified in the published literature on this topic.

Apart from measuring the total time taken by respondents to complete each of the three methodologies, it was deemed necessary to obtain a more comprehensive evaluation of these methodologies in terms of specific attributes or dimensions which comprise respondents' overall evaluations of the preference measurement methodologies. In order to gain more information about what these specific attributes may be, Osgood, Suci, and Tannenbaum's (1957) pioneering research on the nature of semantic meaning was reviewed. Operating on the premise that most commonly used adjectives have considerable connotative overlap, these researchers estimated the underlying factorial structure of the meaning of several
commonly used adjectives using factor analysis. Although a detailed review of the work of these researchers is beyond the scope of this dissertation, it should be noted that the adjectives tested were selected on the basis of a review of Roget's *Thesaurus* (1941 edition) and the analysis was repeated in different subject populations, using different judgmental situations, and estimating the data collected with different estimation and rotation techniques as part of the factor analysis. In spite of these systematic modifications, the same three primary factors were extracted from the different sets of data -- an evaluative factor, a potency factor, and an activity factor (Osgood, Suci, and Tannenbaum, 1957). In other words, it was found that a large proportion of the meaning of commonly used adjectives could be accounted for by the three dimensions of evaluation, potency, and activity.

On the basis of their research, Osgood, Suci, and Tannenbaum (1957) introduced the *semantic differential* as an instrument for measuring the extent to which respondents attribute each of the different meaning dimensions to particular objects. Pairs of opposite adjectives that are representative of the dimension(s) to be measured serve as the items of the instrument. Respondents indicate the extent to which each adjective (or its paired opposite) describes the object (Osgood, Suci, and Tannenbaum, 1957).

It had been noted earlier that one of the unaddressed issues in the published literature on respondent evaluations was the lack of any systematic attempt to infer the structure of overall respondent evaluations in the context of preference measurement methodology. Osgood, Suci, and Tannenbaum's (1957) research on the three dimensions underlying semantic meaning offers some guidance toward a resolution of this issue. Specifically,
the bipolar adjective pairs which loaded primarily on the *evaluative* dimension in their factor analyses would be most helpful in this regard. These sets of bipolar adjectives could then be used as the starting point to investigate the factorial structure of respondent evaluations of preference measurement methodologies.

A review of Osgood, Suci, and Tannenbaum’s (1957) research yielded the following sets of bipolar adjectives which loaded primarily on the evaluative dimension and were applicable to the context of evaluation investigated in this study -- good-bad, valuable-worthless, clear-hazy, nice-awful, successful-unsuccessful, meaningful-meaningless, important-unimportant, and useful-useless. It should be noted that there were other sets of bipolar adjectives which loaded primarily on the evaluative dimension in Osgood, Suci, and Tannenbaum’s (1957) research, but these other sets were not considered for the purpose of this study because it was felt that they were not consistent with the context of evaluation, i.e., evaluation of preference measurement methodologies. Examples of such “rejected” sets of bipolar adjectives included beautiful-ugly, sweet-sour, pleasant-unpleasant, kind-cruel, etc.

From the eight sets of bipolar adjectives which were selected for further consideration from Osgood, Suci, and Tannenbaum’s (1957) research, the sets of good-bad and nice-awful were rejected upon deliberation, mainly because it was felt that these sets of adjectives were overly positive or negative in their evaluation. Further, these two sets were also considered to be too global for the purpose of this study -- they were more of an overall evaluation and would therefore be of little help in any attempt to infer a
structure to the construct of respondent evaluations of preference measurement methodologies.

Moreover, the sets of valuable-worthless, important-unimportant, and useful-useless are essentially interchangeable. Osgood, Suci, and Tannenbaum (1957) themselves noted this point in their original research (p.54). Therefore, it was decided to use only one of these three sets of bipolar adjectives for the purpose of this study. It was decided against using the valuable-worthless set in order to avoid any types of monetary connotations -- this was considered necessary because of the nature of the study population, who were largely of a lower income level. Further, the particular clinic at which the study interviews were conducted had recently introduced some patient cost-sharing mechanisms. Since there was the possibility of patients associating the valuable-worthless set with actual payment on their part (which would consequently bias the evaluations of the preference measurement methodologies), it was decided against using this set for the purpose of this study.

In deciding between the important-unimportant and useful-useless sets, it was decided to opt for the useful-useless set of bipolar adjectives. This is because the word “important” was used in the text to recruit patients for the study; therefore, it is possible that using the same adjective in an evaluation of any aspect of the study might bias the evaluation in favor of the importance of that aspect of the study. Further, the useful-useless set of bipolar adjectives also taps into respondents’ evaluations about whether the procedures are worthwhile or not, without getting into the monetary
connotations that accompany any other descriptions of value (as explained earlier).

Regarding the clear-hazy set of bipolar adjectives, on the basis of the advice of one of the experts who reviewed the instrument for content (i.e., inclusivity and exclusivity) it was decided to change the adjective "hazy" to "vague" because it was felt that the adjective "vague" was better suited to the present context of evaluation than the adjective "hazy." Further, on the basis of a recommendation of another expert, it was decided to change the "meaningful-meaningless" set of bipolar adjectives to "sensible-senseless." The remaining set of bipolar adjectives, i.e., successful-unsuccessful, was used unchanged.

Apart from these sets of bipolar adjectives, the set of long-short was also selected for the purpose of this study, although this set loaded primarily on the potency dimension in Osgood, Suci, and Tannenbaum's (1957) research. It is possible that the reason why this set loaded primarily on the potency and not the evaluative dimension was because of the nature of the concepts that were the subject of the analysis. In this regard, the concepts that were used were mainly people, objects, incidents e.g., father, statue, fire. For these concepts, it is possible that the long-short set would load mainly on the potency dimension, which had other sets such as heavy-light, large-small. This is because the "long" in this set could have been interpreted to mean "tall" for the concepts. Further, although this set loaded primarily on the potency dimension is Osgood, Suci, and Tannenbaum's (1957) research, it also had a fairly high loading on the evaluative dimension. Therefore, it is possible that it could tap into the evaluation dimension. Further, the
published literature on preference measurement methodology evaluation reviewed before noted the importance of time taken to complete the measurement methodologies. Although an "objective" measure of time taken was collected in this study (i.e., actual time taken by respondents, as measured by a watch), it was considered necessary to determine a more "subjective" measure of time, since it is possible that there may be a disparity between these objective and subjective measures of time, i.e., the I-didn't-realize-it-took-so-long effect. Therefore, it was decided to include the long-short set of bipolar adjectives for the purpose of this study.

Further, respondents' acceptance of different preference measurement methodologies might be a function of the degree of interest they have or exhibit in performing the tasks required to implement the methodologies. Therefore, it was decided to include a set of bipolar adjectives which would tap into the respondents' evaluations of how interesting they felt these tasks were. After searching through the dictionary and Roget's Thesaurus, it was decided to use the set "interesting-boring" for this purpose.

The published literature on respondent evaluations for preference measurement methodologies has noted that some methodologies are easier than others. Specifically, the standard gamble has been described as being quite difficult (Torrance, 1976a). Further, within the area of "difficulty of preference measurement methodologies," there are two different aspects which need to be considered -- difficulty in understanding how to do the tasks required of respondents and difficulty in actually doing the tasks required of respondents. Both these aspects need to be a part of any instrument which evaluates any preference measurement methodology.
Therefore, it was decided to include two additional sets of bipolar adjectives in the evaluation instrument -- difficult to understand-easy to understand and difficult to do-easy to do.

In summary, the eight sets of bipolar adjectives used as attributes to evaluate each of the preference measurement methodologies were: (a) interesting-boring, (b) sensible-senseless, (c) easy to do-difficult to do, (d) easy to understand-difficult to understand, (e) clear-vague, (f) useful-useless, (g) long-short, and (h) successful-unsuccessful. In keeping with Osgood, Suci, and Tannenbaum's (1957) original construction of the semantic differential, there were seven response categories for each set of bipolar adjectives (these response categories physically separate the two elements in a set of bipolar adjectives in the instrument). Osgood, Suci, and Tannenbaum (1957) have reported that they had experimented with several different numbers of response categories and found that seven to be the preferred alternative (p.85).

However, a modification was made regarding the specific descriptors of each of the seven response categories. As implemented by Osgood, Suci, and Tannenbaum (1957), the descriptors for the first, second, and third response categories (which are also the fifth, sixth, and seventh response categories since the first three categories apply to one element of the set of bipolar adjectives, while the last three categories apply to the other element of the set) were “very closely,” “quite closely,” and “only slightly.” The instructions in the original implementation called for checking the appropriate response category which was very closely/quite closely/only slightly related to the concept under evaluation. On the basis of feedback
from the panel of experts as well as subjects of the pilot test of the instrument, it was decided to use the descriptors "extremely," "moderately," and "slightly" for the first, second, and third response categories respectively. The instructions for the instrument used in the present study called for checking the appropriate response category which best described the concept under evaluation, i.e., extremely/moderately/slightly interesting or extremely/moderately/slightly boring for the interesting-boring set of bipolar adjectives (see Appendix G). The fourth response category used by Osgood, Suci, and Tannenbaum (1957) -- i.e., "neutral" -- was used unchanged in this study.

Finally, it should be noted that the order in which the eight sets of bipolar adjectives was printed was decided at random. Further, in an effort to avoid response sets, the relative positioning of positive and negative elements in each set of adjectives was varied, so that there was not an abundance of positive adjectives (e.g., interesting, sensible, successful) or negative adjectives (e.g., boring, senseless, unsuccessful) on any one side of the response categories.

Apart from the semantic differential with the eight sets of bipolar adjectives, respondents were also asked to rate their confidence with each of the three preference measurement methodologies on a 5-point Likert-type scale (with 1=very unconfident, 2=somewhat unconfident, 3=neither confident nor unconfident, 4=somewhat confident, and 5=very confident). This scale was also used by Usher (1991) in her comparative evaluation of the ARS procedure and conjoint methodology. The scale used in the Usher (1991) study was used unchanged in this study. It was decided to adopt this
scale for measuring confidence as opposed to using a bipolar set of adjectives -- i.e., confident-unconfident -- mainly in order to be able to compare the results to those of Usher (1991).

In addition to the respondent evaluation instruments for each individual preference measurement methodology, an overall comparative evaluation of the three preference measurement methodologies was also asked of respondents. The reason behind this was prompted by Slovic's (1975) study on choice between equally attractive options (which was reviewed in Chapter IV). Specifically, the interest was in observing which particular preference measurement methodology was chosen as the best on any attribute when there was a tie as indicated by respondent evaluations of the individual methodologies (e.g., if all three methodologies were rated as "extremely interesting," which one would be chosen as the most interesting?). It is important to consider such information in the decision about which particular preference measurement methodology to use in any given situation, since such a decision involves a choice between alternatives.

In order to operationalize the choice questions, respondents were asked to evaluate the three preference measurement methodologies as the "most," "middle," and "least" on each of the nine attributes identified earlier -- interest, sensibility, clarity, success, length, difficulty in understanding, difficulty in doing, usefulness, and confidence. The specific attributes were represented by one of the pair of bipolar adjectives used in their operationalization in the semantic differential (except for the attribute of confidence, which was represented by the word "confident"). A balance was maintained between positive and negative operationalizations of these
attributes. The last question in the overall comparison was an overall ranking of the three preference measurement methodologies. This overall ranking was included in order to understand what attributes were most instrumental in making an overall evaluation, if the evidence from the individual attributes was in conflict with each other.

Apart from these individual and comparative respondent evaluations of the three preference measurement methodologies, respondents were also asked to evaluate the content validity of the health status descriptive system used by the EuroQol. It was decided to ask this question for two major reasons: (a) there have been no published applications of the EuroQol in this particular population, i.e., veterans in the US and (b) to check whether the descriptive classification of the EuroQol was comprehensive in its generality to address all the dimensions of health which need to be considered in the case of hypertension (the condition of interest in this study). This question was operationalized by an open-ended question, asking respondents if they felt there were any other aspects of health that needed to be considered in future studies of this kind.

The individual evaluation instruments for each preference measurement methodology are presented in Appendix G. The instrument for overall comparative evaluation and other feedback from respondents is presented in Appendix H.
University Human Subjects Committee Approval Procedures

All research conducted at The Ohio State University involving human subjects needs to be reviewed by the Human Subjects Committee to ensure that the privacy, safety, health, and welfare of such subjects are adequately protected. There are two such committees at The Ohio State University, each with responsibilities for the different fields of research they are charged with reviewing -- the Biomedical Sciences and the Behavioral and Social Sciences Review Committees. Since the research conducted as part of this dissertation did not involve any invasive procedures (e.g., taking medications, surgery), it was decided to submit the application to review the procedures of this study to the Behavioral and Social Sciences Review Committee.

Since the research conducted in this study involved identifying the patients by means of computerized records (i.e., specific patients could be identified and their responses could be traced back to them -- data collection was not anonymous), the Behavioral and Social Sciences Review Committee rejected an initial application for exemption from full review. A copy of this application from exemption from full review is presented in Appendix I.

An application for full review was accordingly submitted to the Behavioral and Social Sciences Review Committee. The summary sheets and proposal submitted as part of this application are presented in Appendix J. This application was unanimously approved by the committee members, with a few conditions. The conditions pertained to changes in the text used to recruit patients and in the informed consent form (both of which will be discussed later in this chapter). There was also one change with respect to an item in the summary sheets and a few suggestions regarding the interviewer
schedule (the revised version of which is presented in Appendix C). This response from the Behavioral and Social Sciences Review Committee is presented in Appendix K.

The required changes and requested suggestions were implemented and the appropriate revised forms were submitted to the Behavioral and Social Sciences Review Committee, which then gave final approval to the study. The revised summary sheets submitted to the Behavioral and Social Sciences Review Committee are presented in Appendix L, while the final approval form (i.e., Action of the Review Committee) is presented in Appendix M. It should be noted that the entire application package that was submitted to the Behavioral and Social Sciences Review Committee consisted of other material, including the interviewer schedule and data collection forms. Since these other materials are discussed elsewhere and presented in other appendices, they are not addressed in this section.

Clinic Research and Development Committee Approval Procedures

Although the details of the sample selection procedures will be provided in a subsequent section of this chapter, for the purpose of this section it needs to be noted that it was decided to use veterans treated on an outpatient basis for hypertension at the Department of Veterans Affairs Outpatient Clinic (VAOPC) in Columbus, Ohio as subjects for the research conducted in this dissertation. The VAOPC has a Research and Development Committee, which evaluates all proposals to conduct research on veterans treated by the clinic. Accordingly, a proposal was submitted to this
committee. This proposal was similar to the one submitted to the Behavioral and Social Sciences Review Committee, but with a few minor changes. The proposal submitted to the clinic Research and Development Committee is presented in Appendix N.

The clinic Research and Development Committee approved the proposal pending approval from the university Human Subjects Committee and a rewrite of the university informed consent form using the VA Informed Consent Form as a template. The minutes of the committee meeting and the official response of the committee are presented in Appendix O.

The clinic Research and Development Committee was informed when approval from the Behavioral and Social Sciences Review Committee was received. The rewritten informed consent form was also submitted to the committee (see Appendix P). Appendix P also includes the research consent form of The Ohio State University for behavioral and social sciences research, which was also used in this study -- i.e., subjects were asked to sign two different informed consent forms: (a) the VA research consent form and (b) The Ohio State University Behavioral and Social Sciences research consent form. Final approval was then given by the committee. The final approved Research and Development Information System Project Data Sheet is presented in Appendix Q.
Instrumentation Development Procedures

The data collection forms for the implementation of the rating scale and standard gamble scaling methods as well as the data collection forms for respondent evaluations of preference measurement methodologies have already been discussed in previous sections of this chapter and will not be repeated here. The instrument used in this study to administer the EuroQol to patients has also been discussed in an earlier section of this chapter and therefore will not be repeated here. Apart from these instruments, there was one other instrument that was used for data collection purposes in this study -- an instrument used to collect background information about respondents.

Three specific categories of background information were collected from all respondents of this study -- demographic characteristics, severity of hypertension, and presence of comorbidities. It should be noted that this information was collected purely for descriptive purposes. No sub-group analyses were conducted on the basis of this information for two main reasons: (a) it was believed that such analyses would detract from the central purpose of this dissertation and (b) there was not sufficient statistical power to detect significant differences between groups of respondents classified according to these variables.

Three questions were asked to profile the demographic characteristics of the respondents -- age, highest education level, and main activity. These three variables were chosen because they have been shown in the published literature on the EuroQol to significantly affect valuations of health states (Badia, Fernandez, and Segura, 1994; Kind, 1990; Stouthard and Essink-Bot, 1992). Age was operationalized in years as an open-ended question. Highest
education level was operationalized as a multiple choice question with seven response categories, while main activity was operationalized as a multiple choice question with four response categories. Details of the response categories are provided in Appendix R.

Information about the severity of the respondent’s hypertension was obtained by asking them how long they had been diagnosed as having high blood pressure and noting down their most current systolic and diastolic readings of their blood pressure. The question of the diagnosis of high blood pressure was operationalized as a multiple choice question with eight response categories (details are provided in Appendix R). The most recent reading of the respondent’s systolic and diastolic blood pressure was obtained from their medical records. On the basis of these blood pressure readings, patients were classified according to the severity of their hypertension according to the classification provided in the Fifth Report of The Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure or JNC V (see The Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure, 1993).

Information on comorbidities was collected by asking patients to indicate if they had each of 28 different comorbidities. The list of these 28 comorbidities was prepared on the basis of a review of Facts and Comparisons (Reinert, Sery, and Threlkeld, 1994) and consultation with the expert panel and pharmacists at the VAOPC. This list is presented in Appendix R.

Finally, a summary sheet was prepared, denoting information on: (a) whether the interview was conducted before, after, or between appointments, (b) the beginning and ending times of the interview (to
calculate the total time taken to complete the interview), and (c) the order in which the preference measurement methodologies were presented to subjects. A copy of this summary sheet is presented in Appendix S.

An expert panel consisting of health-care researchers and measurement theoreticians checked the instrumentation for content and consistency with the purpose of the study. Since their comments on the specifics of the instrumentation have already been discussed in various sections of this chapter, they will not be repeated here. The original instrumentation developed for the purpose of this study was pre-tested on four students at The Ohio State University and three members of the general population in Columbus, Ohio. The purpose of selecting the four students was to give the interviewer some experience with conducting the interview, while the purpose of selecting the three members of the general population was to gauge the level of understanding of the questions and tasks required of respondents in another population than students, who might be expected to be more acceptable to the nature of questions in the study (since some were health-care professionals).

These interviews took a little over an hour to complete. On account of the small sample sizes, measures of central tendency will not be reported; however, individual time estimates for the total interview and each of the three preference measurement methodologies which were conducted per interview are given in Table 4.
Table 4: Time Estimates for Elements of the Pre-test

<table>
<thead>
<tr>
<th>Subject</th>
<th>Total</th>
<th>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</th>
<th>CUF-SG&lt;sup&gt;e&lt;/sup&gt;</th>
<th>DCCM&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>75</td>
<td>25</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>S2</td>
<td>42&lt;sup&gt;a&lt;/sup&gt;</td>
<td>25</td>
<td>---</td>
<td>12</td>
</tr>
<tr>
<td>S3</td>
<td>65</td>
<td>20</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td>S4</td>
<td>55</td>
<td>20</td>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>G1</td>
<td>75</td>
<td>25</td>
<td>30</td>
<td>12</td>
</tr>
<tr>
<td>G2</td>
<td>60</td>
<td>25</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>G3</td>
<td>70</td>
<td>25</td>
<td>25</td>
<td>15</td>
</tr>
</tbody>
</table>

Legend:
a = all times are in minutes  
b = subjects, where  
  S = students  
  G = general population  
c = time for the full interview  
d = conditional utility function-based procedure with rating scale  
e = conditional utility function-based procedure with standard gamble  
f = discrete choice conjoint methodology  
g = this respondent could not complete all three methodologies; the dashed lines in the body of the table represent the methodologies which were not completed by the respondent
There appeared to be no differences between the time taken by students and members of the general population to complete the interview. With the exception of the fourth student (i.e., S4), all subjects took between one hour and an hour and a quarter to complete the entire interview. There appeared to be no major differences between students and members of the general population in terms of time taken to complete the rating scale and standard gamble scaling methods. Students did, however, seem to complete the discrete choice conjoint methodology faster than members of the general population. It should be noted that these results only generalize to the small sample itself. Further, as noted in the legend to Table 4, one of the test subjects (i.e., S2) could not complete all three preference measurement methodologies on account of a prior meeting.

Informal feedback from test subjects indicated that the interview was very interesting and they themselves did not mind taking the time to participate in the interview. This observation is, of course, highly restricted to only these subjects since they were volunteers. None of the subjects indicated that they had difficulty in following the instructions. All but one of the test subjects indicated that the number of multiattribute alternatives shown in each choice set was "about right" -- the remaining respondent indicated that the number of alternatives should be increased to five. On account of the feedback from the majority of test subjects (as well as the patients who participated in the pilot testing of the instrumentation), it was decided to keep the number of multiattribute alternatives at three in each choice set.
As noted earlier, one of the major reasons for the pre-testing was to give the interviewer some practice in conducting the interviews -- in this regard, the pre-tests were very helpful, because informal feedback from test subjects was very useful in making modifications to the interviewing procedure. For instance, one of the test subjects (S2) commented on the importance of holding the Feeling Thermometer straight, so as to avoid a parallax effect. Another test subject (G3) commented on the amount of time wasted in putting each card into the pockets of the Chance Board. On the basis of this observation, it was decided to place the Chance Board on a flat surface (fortunately, a table was available in the data collection site of the study) and place the cards on (not in) the pockets, so that patients could still view the contents of the cards without worrying about any reflection from the plastic pockets; further no time was spent in removing and replacing each card from the pockets.

Pilot testing of the instrumentation was also conducted with eleven patients at the VAOPC. Since these pilot interviews were unscheduled in the sense that the respondents were not contacted via telephone to arrange a time for the interview in all cases, some pilot interviews only included a subset of the three preference measurement methodologies. However, five complete interviews (out of the eleven) were able to be conducted. As in the case of the pre-test, measures of central tendency for the time estimates will not be reported; however, individual time estimates for the total interview and each of the three preference measurement methodologies which were conducted per interview are provided in Table 5.
Table 5: Time Estimates for Elements of the Pilot-test*

<table>
<thead>
<tr>
<th>Subject</th>
<th>Totalb</th>
<th>CUF-RSc</th>
<th>CUF-SGd</th>
<th>DCCM*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19f</td>
<td>17</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2</td>
<td>11f</td>
<td>---</td>
<td>---</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>25f</td>
<td>21</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>4</td>
<td>20f</td>
<td>---</td>
<td>18</td>
<td>---</td>
</tr>
<tr>
<td>5</td>
<td>11f</td>
<td>---</td>
<td>---</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>65</td>
<td>20</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>56</td>
<td>10</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>35f</td>
<td>30</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>9</td>
<td>70</td>
<td>23</td>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>60</td>
<td>15</td>
<td>16</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>48</td>
<td>9</td>
<td>16</td>
<td>9</td>
</tr>
</tbody>
</table>

Legend:

- a = all times are in minutes
- b = time for the full interview
- c = conditional utility function-based procedure with rating scale
- d = conditional utility function-based procedure with standard gamble
- e = discrete choice conjoint methodology
- f = this respondent could not complete all three methodologies; the dashed lines in the body of the table represent the methodologies which were not completed by the respondent
Patients appeared generally receptive of -- a few even got quite involved in -- the interviews. They seemed very willing to take the time (about 70 minutes on an average) to be interviewed. They indicated that they understood the task required of them in each methodology, and were very accepting of the visual aids used to implement each methodology. When asked their opinion about the number of multiattribute alternatives shown in each choice set, they responded as a group that three multiattribute alternatives per choice set was "about right." The major difference between the patients who took part in the pilot test on the one hand and the students and members of the general population who took part in the pre-test on the other hand was in terms of the amount of time required to explain the purpose of the study and provide instructions. As illustrated by the discrepancies between the sum of the times taken for each of the three methodologies and the total time for the interview, it took longer to explain the purpose of the study and provide instructions to patients as opposed to students and members of the general population.

On the basis of the results of the pre-testing and pilot testing described above, it was concluded that the interview should be acceptable and understandable to patients. The final instrumentation consisted of the following: (a) VA informed consent form and The Ohio State University Behavioral and Social Sciences Research consent form (see Appendix P), (b) the summary sheet of patient interview information (see Appendix S), (c) the EuroQol instrument (see Appendix A), (d) data collection forms for the rating scale and standard gamble scaling methods (see Appendices D and E), (e) respondent evaluation forms of each individual preference
measurement methodology (see Appendix G), and (f) an overall respondent evaluation form (see Appendix H).

**Sample Characteristics and Sampling Procedures**

As noted earlier, the issue of whose preferences need to be assessed in health-care decision making is one of the unresolved methodological issues relevant to the implementation phase in health status measurement (as noted in Chapter I). Potential subjects could be patients, family members or friends of patients, health-care professionals, or members of society at large. The published empirical comparisons of the health status of patients as reported by patients themselves and by proxies (e.g., spouses, caregivers) supports the notion that there are systematic differences between the responses of these two groups of respondents. Specifically, although the health status scores reported by patients and proxies for patients are generally similar for the physical functioning dimensions of health status, there are significant differences in scores for mental functioning (Epstein, Hall, Tognetti, et al., 1989; Hays, Vickery, Hermann, et al., 1995; Rothman, Hedrick, Bulcroft, et al., 1991). Further, in terms of health status preference measurement, although there have not been many significant differences found between different sub-groups of the population on the basis of demographic characteristics (see Froberg and Kane, 1989c), there have been significant differences found between the preferences of patients and healthy subjects as well as between health-care professionals and laypeople (see, for e.g., Froberg and Kane, 1989c; Sackett and Torrance, 1978). These differences
have been explained on the basis of the role of experience with the disease or condition under evaluation, since patients and health-care professionals would be more experienced with the particular state of health resulting from a given disease or condition as opposed to healthy subjects and laypeople respectively (Boyd, Sutherland, Heasman, et al., 1990).

Given these difference between preferences of different groups of respondents, the question that arises, then, is whose preferences need to be assessed for the purpose of health-care decision making. Although there is no consensus in the health status preference measurement literature on this issue, the spirit of Kirshner and Guyatt's (1985) premise of letting the purpose of the analysis dictate the choice of instrumentation and methodology in general offers some guidance toward the resolution. For instance, if the objective of the study is to provide input for public policy decision making which would affect society at large, members of the general public might be used as subjects. The emphasis of this study was to provide input to clinicians to improve patient care decisions with the ultimate aim of improving patient outcomes such as compliance, satisfaction, and health-related quality of life. In such a situation, it was believed that patients would be the most appropriate subjects since it is their preferences which need to be brought to the clinicians' attention. Indeed, Patrick and Erickson (1993) have noted that patients are "ideal" respondents in that they are actually experiencing the states of health being evaluated. Therefore it was decided to select patients as the subjects for this study.

From a clinical decision making point of view, this decision is supported by a study published by Boyd, Sutherland, Heasman, et al. (1990),
who examined the clinical significance of the differences in preferences between patients and healthy subjects for choice of therapeutic intervention for carcinoma of the rectum with or without colostomy. It was found that the expected clinical value of the treatment alternatives was substantially influenced by the differences observed in the preferences for colostomy, and that different options would be chosen depending on whose preferences were used for the decision making process (Boyd, Sutherland, Heasman, et al., 1990).

The choice of hypertensive patients needs to be explained in some more detail. It was decided to sample patients from only one disease state in order to minimize potential variation in the resulting estimated parameters. From all the different disease states and conditions, it was decided to select hypertension as the condition of interest to this study for the following reasons: (a) hypertension is a chronic condition, where the emphasis of drug therapy is on controlling the condition (i.e., controlling blood pressure levels) and improving the health status of the patient -- hence health status is a very important outcome in hypertension. Indeed, interest in measuring the effects of antihypertensive drugs on health status and health-related quality of life has been consistently increasing (Bulpitt and Fletcher, 1992, 1994; Hollenberg, 1987) and the latest report of The Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure (1993) specifically recommends measuring such effects of antihypertensive drugs, (b) previous studies of antihypertensive drugs on health status have shown that these drugs have differential effects on health status dimensions (for e.g., see Croog, Levine, Testa et al., 1986; Testa and Simonson, 1988) -- hence,
the need to identify the relative importance of different health status dimensions to hypertensive patients is important (as discussed in Chapter II), (c) being a chronic condition, hypertension is most effectively controlled by active participation of patients in their therapy. As noted earlier, patients desire for participation extends to letting health-care professionals know what they want in terms of health outcomes. By determining antihypertensive patients' preferences for health states, it is possible to determine what specific health outcomes are considered preferential by these patients. Hence, clinical decision making considering these preferences may be associated with better patient outcomes, such as improved compliance, satisfaction with care, and health-related quality of life.

As noted in an earlier section of this chapter, it was decided to use veterans treated on an outpatient basis for hypertension at the Department of Veterans Affairs Outpatient Clinic (VAOPC) in Columbus, Ohio as subjects for the research conducted in this dissertation. The target population of this study was hypertensive patients taking one or more antihypertensive drugs. Initially, it was intended that the study sample be drawn from a sampling frame consisting of all hypertensive patients who were taking at least one antihypertensive drug and were being treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic (VAOPC) in Columbus, Ohio. However, since the computerized patient records at the VAOPC could not be sorted by diagnosis, it was decided to use a sampling frame consisting of all patients who were prescribed the most commonly used antihypertensive agent at the VAOPC -- Maxzide. Therefore, the study
sample was drawn from a sampling frame consisting of all hypertensive patients who were prescribed Maxzide and were being treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic (VAOPC) in Columbus, Ohio.

The specific selection criteria for inclusion of patients in the study were as follows: (a) patients should have been diagnosed as hypertensive, (b) patients should have been prescribed Maxzide, (c) patients should be treated on an outpatient basis for their hypertension at the VAOPC in Columbus, Ohio, and (d) patients should be able to read and write English.

Sample size calculations were based on a formula suggested by Taylor (1983) and recommended by Furlong, Feeny, Torrance, et al. (1990) for comparing related group means:

\[ n = \left( \frac{(z_a + z_\beta) \sigma_{\text{diff}}}{d} \right)^2 \]  

where

- \( n \) = number of subjects
- \( z_a \) = standard normal deviate corresponding to a probability \( \alpha \) of making a type I error
- \( z_\beta \) = standard normal deviate corresponding to a probability \( \beta \) of making a type I error
- \( \sigma_{\text{diff}} \) = standard deviation of differences between groups of observations
- \( d \) = minimum clinically significant difference
Further, $\sigma_{\text{diff}}$ may be estimated by (Furlong, Feeny, Torrance, et al., 1990; Taylor, 1983):

$$
\sigma_{\text{diff}} = \sqrt{\sigma_1^2 + \sigma_2^2 - 2 \sigma_1 \sigma_2 r}
$$

where

- $\sigma_1$ = standard deviation among subjects in group 1
- $\sigma_2$ = standard deviation among subjects in group 2
- $r$ = coefficient of correlation between the two groups

For the purpose of this study, the a priori alpha level (i.e., probability of accepting a type I error) was set at 0.05, which corresponded to a standard normal deviate (i.e., $z_{\alpha}$) of 1.96 for a two-tailed test of significance. Following Cohen (1977), a beta level (i.e., probability of accepting a type II error) of 0.2 -- which is equivalent to a power coefficient of 0.8 -- was considered acceptable, which corresponded to a standard normal deviate (i.e., $z_{\beta}$) of 1.282 for a two-tailed test of significance. Following Furlong, Feeny, Torrance, et al. (1990), a difference of 0.1 in preference scores was considered clinically significant (i.e., $d$).

Finally, regarding $\sigma_{\text{diff}}$, the above formula is only applicable to the case of two groups. As noted earlier, in this study there were three groups compared -- the conditional utility function-based procedure with the rating scale, the conditional utility function-based procedure with the standard gamble, and discrete choice conjoint methodology. Therefore, for the purpose of calculating sample sizes for this study, it was decided to adopt...
the conservative approach and use estimates of standard deviations and correlation coefficients which would produce the largest sample sizes. In this way, it was decided to opt for the risk of having more than enough power to detect significant differences between groups, as opposed to not having enough power to detect significant differences between groups.

A review of the published literature on preference measurement methodologies yielded the following estimates of standard deviations and correlation coefficients: (a) standard deviations of rating scale and standard gamble preference scores have usually ranged between 0.2 and 0.3 -- studies conducted on the general population generally yielded higher standard deviations than those conducted on patients (for e.g., see Furlong, Feeny, Torrance, et al., 1990; Llewellyn-Thomas, Sutherland, Ciampi, et al., 1984; Llewellyn-Thomas, Sutherland, Tibshirani, et al., 1982; Sackett and Torrance, 1978); standard deviations of discrete choice conjoint preference scores have usually ranged between 0.05 to 0.15 (for e.g., see Louviere and Woodworth, 1983) and (b) correlation coefficients between the rating scale and standard scaling methods have usually ranged between 0.5 and 0.9 (for e.g., see Read, Quinn, and Berwick, 1984; Richardson, Hall, and Salkeld, 1989); correlation coefficients between preference scores obtained via rating scale judgment conjoint and discrete choice conjoint strategies have ranged between 0.7 and 0.9 (for e.g., see Louviere and Gaeth, 1988); since there has been little published research comparing the standard gamble and pick-one scaling methods, an estimate of the correlation coefficient between these two methods had to be assumed -- this was assumed at 0.3, which is a very
conservative estimate, given the magnitude of correlation coefficients provided earlier.

For the purpose of estimating $\sigma_{\text{diff}}$, it was decided to use an estimate of the correlation coefficient between groups of 0.3 and an estimate of the standard deviations of each group of 0.3 -- it should be noted that these are the most conservative estimates from those provided above. Using these estimates, the calculated value of $\sigma_{\text{diff}}$ is 0.3549. Inputting all the values recorded above into equation 5.5 yielded an estimated sample size of about 130 patients (more specifically, a sample size of 128 patients if the value $\sigma_{\text{diff}}$ was taken to two decimal places and a sample size of 132 if the value $\sigma_{\text{diff}}$ was taken to four decimal places). The steps involved in this calculation are shown in Table 6. Assuming an 85% participation rate (which is consistent with the published literature in this area -- see, for e.g., Torrance, 1976a; Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992), it was decided to approach 153 patients to participate in the study.

Although it was initially intended to use systematic random sampling to select patients from the sampling frame, this was not considered feasible on account of time considerations. This was because the VAOPC was considering moving to another location three months after the beginning of this study. In order to avoid any potential location effects, it was necessary to complete the data collection process within three months. A strictly systematic random sampling technique would not be able to select the required number of patients (i.e., 130, as explained above) within this time frame (patients treated on an outpatient basis had scheduled appointments only every three to six months).
Table 6: Calculations for Sample Size Determination

(a) determining $\sigma_{\text{diff}}$:

\[ \sigma_{\text{diff}} = \sqrt{(\sigma_1^2 + \sigma_2^2 - 2 \sigma_1 \sigma_2 r)} \]

where
\[
\begin{align*}
\sigma_1 &= \text{standard deviation among subjects in group 1} \\
\sigma_2 &= \text{standard deviation among subjects in group 2} \\
r &= \text{coefficient of correlation between the two groups}
\end{align*}
\]

as explained in the text,
\[
\sigma_1 = 0.3; \sigma_2 = 0.3; r = 0.3
\]

Therefore,
\[
\sigma_{\text{diff}} = \sqrt{(0.3)^2 + (0.3)^2 - 2 (0.3) (0.3) (0.3)} = 0.3549
\]

(b) estimating $n$:

\[ n = \left[ \frac{(z_\alpha + z_\beta) \sigma_{\text{diff}}}{d} \right]^2 \]

where
\[
\begin{align*}
n &= \text{number of subjects} \\
z_\alpha &= \text{standard normal deviate corresponding to a probability } \alpha \text{ of making a type I error} \\
z_\beta &= \text{standard normal deviate corresponding to a probability } \beta \text{ of making a type I error} \\
\sigma_{\text{diff}} &= \text{standard deviation of differences between groups of observations} \\
d &= \text{minimum clinically significant difference}
\end{align*}
\]
Table 6 (continued)

as explained in the text,

\[ z_\alpha = 1.96; z_\beta = 1.282; d = 0.1 \]

Two sets of sample size estimates were calculated, varying with respect to the number of decimal places to which the value of \( \sigma_{\text{diff}} \) was taken:

(i) \( \sigma_{\text{diff}} \) upto two decimal places (i.e., \( \sigma_{\text{diff}} = 0.35 \)):

\[
n = \left[ \frac{(1.96 + 1.282) \times 0.35}{0.1} \right]^2 = 128 \text{ patients}
\]

(ii) \( \sigma_{\text{diff}} \) upto four decimal places (i.e., \( \sigma_{\text{diff}} = 0.3549 \)):

\[
n = \left[ \frac{(1.96 + 1.282) \times 0.3549}{0.1} \right]^2 = 132 \text{ patients}
\]

It was decided to take the mean of these two estimates, i.e., 130 patients

Therefore, calculated sample size = 130 patients.
Therefore, a modified systematic random sampling technique was followed. The sampling procedure used in this study can be conceptualized as a two-stage sampling process. In the first stage, regular systematic random sampling was carried out on the sampling frame. There were 1,205 patients in the sampling frame; therefore, every ninth patient was sampled as part of this first stage of the sampling process. Sampled patients’ appointment schedules were checked to see if they had an appointment at the VAOPC during the period the study was expected to be conducted at the VAOPC. If so, they were selected for study recruitment; if not, they were not considered for the purpose of the study.

After two runs through the sampling frame there will still not enough subjects selected for recruitment purposes. Therefore, it was decided to implement a second stage in the sampling process. In this stage, a day-by-day listing of appointments for the remaining patients on the sampling frame was constructed. Each day was then divided into four parts -- 8:00 a.m. to 10:00 a.m., 10:00 a.m. to 12:00 noon, 1 p.m. to 3 p.m., and 3 p.m. to 5 p.m. (the time between 12:00 noon and 1 p.m. was the lunch break at the clinic, and no appointments were scheduled during this time period). Patients were then placed into each of the four parts of the day according to the time of their scheduled appointments and sampled at random from within each part of the day. It should be noted that it was not always possible to select four patients in each day, since some patients who were selected had appointments at the beginning of their time slots and preferred to be interviewed for the study before their appointment, thereby not allowing any interviews to be scheduled for the time period before their
time period. Similarly, some patients who were selected had appointments at the end of their time slots and preferred to be interviewed for the study after their appointment, thereby not allowing any interviews to be scheduled for the time period after their time period. Further, it should also be noted that interviews were scheduled both before as well as after appointments.

Telephone numbers of the patients who were selected by this two-stage sampling process were obtained from the computerized patient records. Selected patients were telephoned one or two days before their scheduled appointments and recruited over the telephone to participate in the study. The text used for recruiting patients over the telephone was standardized. A sheet with standardized fall-back answers to commonly asked questions was kept in hand while speaking with patients over the telephone. A datasheet for the telephone recruits was also prepared. Copies of the text for recruiting patients over the telephone, fall-back answers to commonly asked questions, and the datasheet for telephone recruits are presented in Appendix T.

**Data Collection Procedures**

As noted in Chapter I, one of the unresolved methodological issues related to the implementation of health status measures is the method of data collection. For the purpose of this study, it was decided to use face-to-face interviews as the method of data collection. Face-to-face interviews have been the most commonly used methods of data collection in the health status preference measurement literature, as evidenced by the research of the
McMaster group of researchers (see Torrance, Boyle, Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992) and the San Diego group of researchers (see Kaplan and Anderson, 1988a, 1988b, 1990 for a review). The notable exception is the research of the EuroQol group of researchers, who have used mail surveys as their method of data collection (The EuroQol Group, 1990). However, the more current research of the EuroQol group involves experimenting with different methods of data collection, including face-to-face interviews (Rosser and Sintonen, 1993), perhaps on account of the low response rates obtained in their previous research.

Although not without their own drawbacks, face-to-face interviews offer significant advantages for collecting data on health status preferences. These advantages include answering respondents' questions, probing for adequate answers, multimethod data collection procedures, and easily allowing varying sequencing of questions (Fowler, 1988). Each of these potential advantages was influential in deciding upon face-to-face interviewing as the preferred method of data collection for the purpose of this study.

It was important to have the opportunity to answer respondents' questions because of two main reasons: (a) Torrance (1976a) has noted that the standard gamble scaling method might be difficult for elderly and less educated respondents to understand and (b) this was the first known application of discrete choice conjoint methodology to the measurement of health status preferences -- therefore, it was considered necessary to have an investigator on hand to address any queries subjects may have in order to offer recommendations for future implementations of such methodologies in
health status preference measurement. It was considered important to have the opportunity to probe for adequate answers because this is required for the consistency checks needed for the implementation of the conditional utility function-based procedure.

Probably the most important reason for selecting face-to-face interviews as a method of data collection for this study was the need to use different visual aids for each of the preference measurement methodologies - no other method of data collection allows such flexibility in dealing with different types of visual aids. As noted earlier, these visual aids were constructed in order to simplify the task required of respondents; therefore, it was necessary to have a method of data collection which would allow the visual aids to be used to their fullest potential. Along these lines, it was important to allow for some flexibility in the order of presentation of the different preference measurement methodologies, since different respondents received the three methodologies in different orders (details on the order of presentation are provided later in this section). Face-to-face interviews easily allow such differential sequencing of information presentation to respondents at little additional cost.

Further, there were logistical issues which contributed to the selection of face-to-face interviews as the method of data collection used for the purpose of this study. In this regard, since patients were already coming into the VAOPC for their appointments and usually had to wait between 30 minutes to an hour to pick up their prescriptions, it was considered logistically feasible to schedule interviews within this time slot. Moreover, some patients had multiple appointments in the same day at the VAOPC;
therefore, since these patients would be at the VAOPC for most of the day, it was thought that it might not be overly bothersome for them to be interviewed at the VAOPC.

It should be noted that the use of face-to-face interviews has a few drawbacks, the chief one being their high cost and possibility of interviewer effects. While the issue of high cost was certainly a concern of this study, it was believed that using other data collection methods might be associated with other costs, such as nonresponse -- for instance, Cobb, King, and Chen (1957) have reported that elderly and less-educated subjects are less likely to respond to mail surveys than younger and more-educated subjects. As noted earlier, a significant portion of the subjects of this study were elderly and less-educated. The issue of interviewer effects is an extremely important one, because subtle cues given (knowingly or unknowingly) by the interviewer could bias the data collected via face-to-face interviews. In an attempt to control for such interviewer effects, a standardized interviewer schedule was prepared (see Appendix C). This schedule was tested in the pre-test and pilot tests and was followed in the actual data collection process. Therefore, in spite of some potential shortcomings (which an attempt was made to control), it was felt that the benefits of face-to-face interviews outweighed the drawbacks of face-to-face interviews; thus, it was decided to use face-to-face interviews as the method of data collection for the purpose of this study.

The advantage of allowing differential sequencing of information using face-to-face interviews was noted earlier. It was important to have this property since the order of presentation of preference measurement methodologies needed to be partially randomized within respondents. It
was not desired to have complete randomization of preference measurement methodologies within respondents because the implementation of the standard gamble scaling method requires the establishment of a preferential ordering of all alternatives before measuring the utility of each alternative. The theoretical rationale behind this requirement of the standard gamble scaling method is given in Luce and Raiffa (1957), while empirical evidence for this is provided by Laskey and Fischer (1987).

Rather than have respondents rank order alternative health states as a separate step in the implementation of the standard gamble scaling method (which, it was felt, would add even more time to an already long interview), it was decided to implement the standard gamble after -- although not necessarily immediately after -- the rating scale scaling method (since respondents visually rank-ordered alternative health states as part of the rating scale scaling method). Therefore, for the purpose of this study, the order of presentation of the three preference measurement methodologies was partially randomized (since the standard gamble scaling method had to be after -- although not immediately after -- the rating scale scaling method). Specifically, three different presentation orders were used: (a) rating scale/standard gamble/pick-one, (b) rating scale/pick-one/standard gamble, and (c) pick-one/rating scale/standard gamble.

One of the outstanding methodological issues pertaining to the implementation of health status measures that was introduced in Chapter 1 is the particular site of data collection. The specific site of data collection was a small semi-private booth in the pharmacy area at the VAOPC. The reason for conducting the interviews at the VAOPC itself was explained earlier in
this section. The location of the booth in the pharmacy area was dictated by logistical considerations. The pharmacy area in the VAOPC had four semi-private counseling areas, staffed by full-time pharmacists. The booth in which all interviews were conducted was adjacent to these counseling areas on one side and next to the pick-up window (where patients received their filled prescriptions) on the other side. This location was quite ideal, since it was easy to give patients directions as to how to find it (all patients knew where the pick-up window was) and it afforded the opportunity to let patients be interviewed and not run the risk of checking to see if their prescriptions were ready or not by leaving the booth.

Within the booth, there was a small rectangular table (about 4 feet by 2 feet). During the interview, this table was at the left-hand side of the interviewer and at the right-hand side of the respondent -- the interviewer and respondent did not sit on different sides of the table. The table was, however, used to support the laptop computer which was used to collect the data for the discrete choice conjoint methodology. Although access through the booth was restricted, there were a few instances in which clinic personnel had to cross through the booth in order to get in and out of the pharmacy. In such instances, they did so well behind the interviewer's back, so as to not come physically between the interviewer and the respondent.

Moreover, although an attempt was made to interview patients alone, some patients were accompanied by their spouses for the interview. In such cases, the spouse was asked to sit at a distance away from the patient and requested to let the patient decide for himself or herself what answers to the
questions posed should be. Spouses were generally understanding of the need to let the patient represent himself or herself in the interview.

The interviewer schedule presented in Appendix C gives the detailed instructions provided to respondents. A few points need to be noted about these instructions. First, the scenario presented to each patient for each of the three preference measurement methodologies was that they had to imagine that a medication (not necessarily any one of the medications they were currently taking) had the particular effects described for a period of one year. Using this common scenario, appropriate further questions were asked depending on the specific demands of the different preference measurement methodologies. Second, the handling to the state of immediate death is sensitive, and every effort was made to convey the hypothetical nature of the situation (without making it sound so unrealistic so as to bias the responses due to hypothetical bias) and that this condition was in no way related to the current state of health of patients. The specific explanation given to respondents is presented as a part of Appendix C. Informal responses from patients indicated that the handling of this sensitive issue was appropriate, in that patients said they understood that the state of immediate death was another state of health which was always a possibility of the effects of medications. Finally, a fair amount of time was spent in explaining the visual aids to the respondents and making sure they understood how each visual aid was supposed to work. As noted above, details are provided in Appendix C.

It should be noted that the interviewer schedule presented in Appendix C is largely similar to that presented in Furlong, Feeny, Torrance,
et al. (1990); it was decided to follow the schedule of these researchers because of their vast experience in health status preference measurement and the fact that this schedule was successfully used in a large field survey (Torrance, Zhang, Feeny, et al., 1992). The modifications made in the schedule of Furlong, Feeny, Torrance, et al. (1990) were mainly on account of decreasing the amount of time taken for each task required of respondents, since three different preference measurement methodologies were conducted in this study as opposed to two in the field survey of those researchers.

The specific steps in the data collection process followed for the purpose of this study were as follows:
(a) patients were recruited over the telephone to participate in the study (see Appendix T). This was usually done a day or two before their scheduled appointments at the VAOPC,
(b) patients were greeted when they arrived at the booth and informed of the purpose of the study. After explaining the study purpose and procedures, questions were entertained; then, patients were requested to read and sign the VA and The Ohio State University research consent forms (see Appendix P),
(c) the interview commenced with the background information about demographic characteristics, severity of hypertension, and comorbidities (see Appendix R),
(d) the interview continued by asking patients about their state of health on the day of the interview, as measured by the EuroQol (see Appendix A),
(e) the first of the three preference measurement methodologies was implemented (see Appendix D, E, or F as appropriate),
(f) respondent evaluations of the first preference measurement methodology were collected (see Appendix G),
(g) the second of the three preference measurement methodologies was implemented (see Appendix D, E, or F as appropriate),
(h) respondent evaluations of the second preference measurement methodology were collected (see Appendix G),
(i) the last of the three preference measurement methodologies was implemented (see Appendix E or F as appropriate),
(j) respondent evaluations of the third preference measurement methodology were collected (see Appendix G),
(k) comparative respondent evaluations of the three preference measurement methodologies and the interview in general were finally collected (see Appendix H).

**Data Analysis Procedures**

The data were analyzed using a variety of statistical techniques, as appropriate for each of the research questions which was investigated in this study. This section discusses the data analysis procedures used to address each of the research questions that were presented in Chapter II. First, however, the data entry and coding procedures will be discussed.
Data Entry and Coding Procedures

A codebook was prepared for all data collected as part of the interview. The only data that was not coded was that collected as part of the implementation of the discrete choice conjoint methodology, since this was collected on the computer and was analyzed using the CBC System. All other analyses were carried out using SAS. A copy of the codebook is presented in Appendix U. After each line, the column number was noted. If there was any discrepancy between this column number and that expected from the codebook, it was resolved before going any further. After every twentieth questionnaire, a random check of the input data was conducted. Finally, after all the data had been entered, they were visually inspected to check for any noticeable discrepancies (e.g., scores greater than 1.0 or less than 0.0 for the rating scale and standard gamble scaling methods, respondent evaluations greater than 7 for any of the attributes of respondent evaluations). All such discrepancies (there were very few) were resolved before conducting the analyses.

It should be noted that descriptive analyses of respondent characteristics and response rates were also conducted; since these only involved regular descriptive statistics (e.g., means and frequencies) and were only of interest for reporting purposes, they will not be discussed in this chapter. They are, however, discussed in the next chapter of this dissertation.
Research Questions Pertaining to Modeling the Data

As noted in Chapter II, the first six research questions that were investigated pertained to modeling the data obtained using the three preference measurement methodologies. This section will discuss the analytical procedures followed for each of these six questions. The questions will be addressed two at a time, in recognition of the fact that there were two research questions which were concerned with modeling the data obtained using each preference measurement methodology. Further, since two of the methodologies shared the same modeling procedure (i.e., the conditional utility function-based procedure), they will be discussed together. Therefore, the data analysis procedures for four of the research questions will be discussed together (Research Questions # 1-4) and the procedures for the remaining two questions will also be discussed together (Research Questions # 5 and 6).

(a) Modeling the Data for the Conditional Utility Function-based Procedure

This section discusses the procedure used for modeling the data collected via the rating scale and standard gamble scaling methods used in the implementation of the conditional utility function-based procedure. Since the modeling procedure used for the data collected via both scaling methods (i.e., rating scale and standard gamble) was identical, the distinction between these two scaling methods will not be maintained in this section. The specific research questions for which the data analysis procedures will be described in this section are:
Research Question #1: What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?

Research Question #2: What is the predictive ability of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology as a measure of health status preferences, as indicated by the mean absolute residual?

Research Question #3: What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Research Question #4: What is the predictive ability of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology as a measure of health status preferences, as indicated by the mean absolute residual?

As noted earlier in this chapter, single attribute preferences were measured using value and utility scales, while the multiattribute corner states were constructed so as to be consistent with disvalue and disutility scales. The first task, then, was to get a “common” denominator for these
single and multiattribute preference functions. Following Torrance, Boyle, and Horwood (1982) and Torrance, Zhang, Feeny, et al. (1992), it was decided to use the disvalue and disutility scales as this common denominator (as explained earlier). Therefore, for the purpose of modeling the data for the conditional utility function-based procedure, the complements of the single attribute preferences were calculated (in order to convert them into disvalue and disutility scales). Further, although the multiattribute corner states were constructed so as to be consistent with the disvalue or disutility formulation in order to ease understanding by respondents (as explained earlier), they were actually measured on value and utility scales (again in order to not confuse respondents by providing different anchor points). Therefore, preferences for the multiattribute states were also converted to their disvalue and disutility mathematical equivalents by taking their respective complements.

The importance of determining the appropriate model form of preference functions was discussed in Chapter II. As noted in this chapter, measuring preferences for single attribute corner states allows easy determination of the appropriate model form of preference functions in the conditional utility function-based procedure (since the preference scores for these single attribute corner states are equal to the individual attribute scaling constants or $k$).

The appropriate model form -- additive or multiplicative -- was determined on the basis of an examination of these scaling constants. The reason why a multilinear model form was not considered was noted earlier in this chapter. For the purpose of estimating whether an additive or
multiplicative model form was appropriate; the additive model form was considered appropriate if the sum of all the individual attribute scaling constants (i.e., $\sum k$) was equal to one; otherwise the multiplicative model form was considered appropriate and the interaction parameter $k$ was determined using a computer package. The particular computer package used was SOLVE (Mundt, 1989).

After determining the appropriate model form, preference scores for multiattribute health states can be calculated using either Equation (5.1) or Equation (5.3), depending on whether an additive or multiplicative model form is considered appropriate. Using this equation, preference scores (in terms of disvalues and disutilities) were calculated for the three interior states (i.e., BO, E3, and E8). These calculated preference scores were compared to the measured preference scores for these same states by calculation of the residuals for each set of scores. Finally, the mean absolute residual was calculated in order to get an estimate of the predictive ability of these methodologies.

(b) Modeling the Data for the Discrete Choice Conjoint Methodology

This section describes the analysis procedures followed for the research questions pertaining to modeling the data obtained using the discrete choice conjoint methodology. The specific research questions for which the data analysis procedures will be described in this section are:
**Research Question # 5:** What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the discrete choice conjoint methodology or DCCM?

**Research Question # 6:** What is the predictive ability of the discrete choice conjoint methodology or DCCM as a measure of health status preferences, as indicated by the mean absolute residual?

The data collected using the pick-one scaling method in the choice tasks were analyzed using maximum likelihood estimation to determine the parameters of the MNL choice model assumed to represent respondents' decision making processes. Hosmer and Lemeshow (1989) have succinctly summarized the general principle behind maximum likelihood estimation:

"In a general sense, the method of maximum likelihood yields values for the unknown parameters which maximize the probability of obtaining the observed set of data. In order to apply this method we must first construct a function, called the **likelihood function**. This function expresses the probability of the observed data as a function of the unknown parameters. The **maximum likelihood estimators** of these parameters are chosen to be those values which maximize this function. Thus, the resulting estimators are those which agree most closely with the observed data" (p.8; emphasis in original).

For the purpose of this study, maximum likelihood estimation was performed using the CBC System. Since the data were collected directly on a diskette, all that was needed to be done was run the analysis on the CBC System using this diskette. As noted earlier in this chapter, only ten of the
twelve choice sets were used as the estimation sample for modeling purposes; the other two choice sets served as the holdout choice sets for estimating the predictive ability of the discrete choice conjoint methodology.

The CBC System permits estimation of all main effects and two-way interactions. In order to estimate higher order interactions, it is possible to save the data file in a format which can be analyzed using SYSTAT LOGIT. However, in keeping with the advice of researchers experienced in discrete choice conjoint modeling (for e.g., see Batsell and Louviere, 1991), it was decided to only estimate up to two-way interactions. Therefore, the capability of the CBC System to interface with SYSTAT LOGIT was not utilized in this study.

A stepwise multinomial logit procedure was utilized in modeling the data to determine the appropriate model form of the preference function as indicated by the discrete choice conjoint methodology. The guiding principle behind testing for the significance of a coefficient of a variable in any model under evaluation as part of this modeling procedure related to the following question (Hosmer and Lemeshow, 1989): "Does the model that includes the variable in question tell us more about the outcome (or response) variable than does a model that does not include that variable?" (p.12). This question was answered by comparing observed scores of the response variable to predicted scores obtained from each of two models -- the first with and the second without the variable in question being considered for inclusion in the model (Hosmer and Lemeshow, 1989).

The comparison between observed and predicted scores is based on the log likelihood function, which is nothing but the natural logarithm of the
likelihood function (i.e., the function which expresses the probability of the observed data as a function of the unknown parameters). The reason for expressing this function in log terms is largely due to mathematical convenience -- the actual likelihoods are very small numbers and therefore it is more convenient to work in log terms, since the interpretation of the function will remain unchanged whether actual or log terms are used. In mathematical terms, the comparison between the observed and predicted scores using the likelihood function may be represented by the following expression:

\[ D = -2 \ln \left( \frac{L_{\text{current model}}}{L_{\text{saturated model}}} \right) \] (5.7)

where

- \( L_{\text{current model}} \) = likelihood of current model (i.e., model with variable in question being considered for inclusion in model)
- \( L_{\text{saturated model}} \) = likelihood of saturated model (i.e., model containing as many parameters as there are data points)
- \( \ln \) = natural logarithm
- \( D \) = deviance

In Equation (5.7), \( L_{\text{current model}} \) corresponds to the predicted score and \( L_{\text{saturated model}} \) corresponds to the observed score. The quantity in the large square brackets [i.e., \( (L_{\text{current model}})/(L_{\text{saturated model}}) \)] in Equation (5.7) is called the likelihood ratio. The reason for using minus twice its natural logarithm is mathematical and is necessary to obtain a quantity whose distribution is
known and therefore can be used for hypothesis testing purposes. This test is called the likelihood ratio test.

Assessing the significance of an independent variable involves comparing the value of the deviance statistic ($D$) with and without the independent variable in question. The change in $D$ due to including the independent variable in the model is obtained using the likelihood ratio statistic $G$, which is defined as:

$$G = D_{\text{model with variable}} - D_{\text{model without variable}}$$  

where

$D_{\text{model with variable}}$ = deviance for model with variable in question

$D_{\text{model without variable}}$ = deviance for model without variable in question

$G$ = likelihood ratio statistic

Further, since the likelihood of the saturated model is common to both the values of $D$ being differenced to compute $G$, a simpler computational formula for $G$ can be used:

$$G = -2 \ln \left[ \frac{L_{\text{model without variable}}}{L_{\text{model with variable}}} \right]$$  

This is the statistic which was used to evaluate the significance of successive models (both univariate as well as multivariate) in the stepwise modeling procedure used to model the DCCM data. Under the null hypothesis that the coefficient of each of the added variables is equal to zero, the likelihood ratio statistic ($G$) follows a chi-square distribution with
degrees of freedom equal to the number of estimated parameters over and above those of the "base" model [i.e., the model without the added variable(s) in question]. Rejection of the null hypothesis can be interpreted to mean that at least one, and perhaps all, the coefficients of the model under evaluation are different from zero.

For the multivariate case, the significance of individual independent variables (i.e., health status dimensions or attributes) was evaluated using the Wald test, which was obtained by comparing the maximum likelihood estimate of the coefficients to a the maximum likelihood estimate of its standard error. Under the null hypothesis that the coefficient of each of the individual added variables is equal to zero, the resulting ratio follows a normal distribution. Thus, the value of the Wald statistic may give an indication of which of the individual independent variables in the estimated model may or may not be significant. A useful cutoff value of the Wald statistic in this regard is 2.00, which approximately corresponds to an alpha level of 0.05 (Hosmer and Lemeshow, 1989).

The computation method used by the CBC system is iterative. It starts with a "null model," consisting of coefficients all equal to zero, and computes the likelihood of the respondents' choices, given probabilities corresponding to that initial solution. During that computation, information is also accumulated about a "gradient vector," indicating how the initial solution should be changed for greatest improvement. Those changes are made, and a second iteration is begun. The iterative process continues until a pre-specified condition is reached, e.g., a set number of iterations, a fixed change in log likelihood from one iteration to the next, or convergence is
reached (Metegrano, 1993). It was decided to retain the default conditions of the CBC System for the purpose of this study, since these have been set down on the basis of the developers' experience with discrete choice conjoint modeling. These default conditions were a maximum of 20 iterations and a fixed change in log likelihood of less than 1 in the tenth decimal place.

Specifically, in modeling the DCCM data for the purpose of this study, a series of univariate models were first fit (five in all, corresponding to the five attributes) to the data in order to determine the need for including each individual attribute in the model. The significance of each of the individual attributes was determined by the likelihood ratio statistic ($G$), as described above. Following the observation that each of these univariate models was highly significant compared to a null model (as indicated by the likelihood ratio statistic), a main-effects model was fit to the data (i.e., with all five attributes, but no two-way interactions). This model was also found to be highly significant.

After this, a series of models with main-effects plus single two-way interactions were fit to the data, and compared to the main-effects only model (as indicated by the likelihood ratio statistic). It was found that only one two-way interaction -- between self-care and usual activities -- was significant. When a series of multiple two-way interaction models were fit (i.e., main-effects plus the self-care/usual activities interaction plus other two-way interactions in turn), no other two-way interaction was found to be significant. Therefore, the estimated model form of the preference function was main-effects plus the two-way interaction between self-care and usual activities.
The CBC System also offers a share simulation estimation, using the estimated model form as the basis for predicting shares among competing alternatives (i.e., how many respondents chose each particular alternative in the choice set). Descriptions of the alternatives included in the two holdout choice sets were input into this share simulator, and the simulated shares that were the output of this simulator were compared with the actual shares of the two holdout choice sets that were collected in the data collection phase using the pick-one scaling method. Residuals were calculated for each of the sets of shares, and a mean absolute residual was also calculated. In this way, an estimate was obtained regarding the predictive ability of the discrete choice conjoint methodology.

Preference scores for multiattribute health states can be calculated using the estimated model form as determined by modeling the DCCM data. However, in order to be comparable to the scores obtained using the conditional utility function-based procedure with the rating scale and standard gamble scaling methods (and therefore address Research Questions 7 through 14), it was necessary to convert these preference scores into a 0 to 1 scale, where (in order to be consistent with the scores obtained using the conditional utility function-based procedure with the rating scale and standard gamble scaling methods) 1 represented the preference score given to the all-best health state -- with the best level of functioning on each of the five attributes of health status -- and 0 represented the state of immediate death.
In order to make this conversion, the individual attribute-level parameter estimates for the main-effects plus self-care/usual activities model were summed for all the relevant effects for the health states in question. In this regard, for the all-best health state, the relevant effects were the parameter estimates for the highest level of functioning for each of the five attributes and the parameter estimate for the interaction between the highest level of self-care and the highest level of usual activities. The sum of these effects was set as the highest point (i.e., 1.0) in the universe of rescaled scores. Since each choice set contained the state of immediate death, the analysis provided a parameter estimate for this state, i.e., no summation procedures were necessary for the state of immediate death. This estimate was set as the lowest point (i.e., 0.0) in the universe of rescaled scores.

All the other health states needed for further analyses were mapped onto this rescaled continuum in the following manner: (a) the sum of relevant effects was calculated by adding the parameter estimates of each of the five main effects and the two-way interaction in question, (b) this sum was converted into the rescaled 0-1 continuum (with 0 as immediate death and 1 as the all-best health state) by dividing it by the sum of the all-best health state. In this way, the preference scores of each of the multiattribute health states of interest for addressing the different research questions in this study was converted into the 0-1 scale. The multiattribute health states which were thus converted were the five corner states and the three interior states that were used in the conditional utility function-based procedure.

As a result of this conversion procedure, preference scores for eight different multiattribute health states (i.e., five corner states and three interior
states) were obtained on a 0-1 scale, where 0 was the preference score for immediate death and 1 was the preference score for the all-best state. These were the preference scores which were used to address Research Questions 7 through 14. It should be noted that all analyses to address these questions were conducted at the aggregate level, using the arithmetic mean as the measure of central tendency. The specific analytical procedures used to address each of these research questions are described in turn in the next section. In all cases, the a priori level of significance was set at 0.05.

Research Questions Pertaining to Comparing Preference Scores

As noted in Chapter II, Research Questions 7 through 14 pertained to comparing the preference scores obtained using the three preference measurement methodologies. This section will discuss the analytical procedures followed for each of these eight questions. All the analytical procedures described in this section were conducted at three different levels: (a) at the overall level (i.e., all eight multiattribute health states), (b) at the level of the five corner states separately, and (c) at the level of the three interior states separately.

It was decided to conduct the analysis at these three levels separately in order to investigate whether there may be a difference in the results of the analyses due to the types of health states under evaluation. In this regard, the structural differences between the corner and interior states are important. Specifically, the corner states had only one health status dimension which was not at the best level of functioning; on the other hand, the interior states had several health status dimensions which were not on
the best level of functioning. It is possible that these structural differences may be associated with differences in preference scores of the alternative health states under evaluation.

For the purpose of consistency with the format adopted in Chapter II and to help organize the section, the data analysis procedures are described in five sub-sections, corresponding to the different bases of comparison of the preference scores obtained using the three preference measurement methodologies, i.e., preferential ordering, magnitude of preference scores, correlations, concordances, and functional relationships.

**Preferential Ordering**

The first aspect on which the different preference measurement methodologies were compared was in terms of the order in which they ranked the alternative health states under evaluation. Two specific research questions were investigated in this regard. The data analysis procedures for each will be discussed in turn in this sub-section.

**Research Question # 7:** Do the three health status preference measurement methodologies (i.e., the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) agree in terms of the order in which they rank alternative states of health?
The data used to address this question consisted of the preference scores of the eight multiattribute health states (i.e., five corner states and three interior states) determined via the three preference measurement methodologies, as explained in the previous section. The analysis was conducted at three different levels: (a) at the overall level (i.e., all eight multiattribute health states), (b) at the level of the five corner states separately, and (c) at the level of the three interior states separately.

This research question was addressed using Kendall's Coefficient of Concordance or Kendall's W. Kendall's W measures the extent of agreement in the rank ordering of a variable measured by three or more different measures. It is bounded between 0 and 1, with 0 indicating no agreement among the three measures and 1 indicating perfect agreement among the three measures (since there is no such thing as perfect disagreement with more than two measures, Kendall's W cannot take on a value of less than 0). However, if there are ties present in any of the rankings, the maximum value of Kendall's W is no longer 1 and a correction factor for ties needs to be incorporated in its computational formula (Gibbons, 1993). It should be noted that there were no ties in any of the rankings of multiattribute health states produced by any of the three preference measurement methodologies investigated in this study; therefore, the correction factor for ties will not be discussed further.

Given that Kendall's W is a descriptive measure of agreement among the rankings of three or more measures, it can also be used to test the null hypothesis of no agreement among the rankings produced by the three or
more measures. The only possible alternative hypothesis is that agreement exists among the rankings produced by the three or more measures.

As noted earlier, statistical analyses were conducted using SAS. However, SAS does not include a computational procedure for calculating Kendall’s W. Therefore, for the purpose of addressing this research question, Kendall’s W was calculated according to the following computational formula (which was manually programmed into SAS):

\[ W = \frac{12 S}{k^2 n (n^2 - 1)}, \]  \hspace{1cm} (5.7)

where

- \( S \) = sum of squares of deviations of rank sums
- \( k \) = number of measures among which agreement is measured (i.e., 3 for the purpose of all levels of analysis)
- \( n \) = number of observations (i.e., 3, 5, or 8, depending on the level of the analysis)
- \( W \) = Kendall’s Coefficient of Concordance

Further, \( S \) can be calculated by the following formula:

\[ S = \sum [R_i - k(n + 1)/2]^2, \]  \hspace{1cm} (5.8)

where

- \( i = 1 \ldots n \)
- \( R_i \) = rank sums of each of the measures

and all other terms are as previously defined.
As noted earlier in this section, the analytical procedures that were conducted in order to address this research question were performed at three different levels — for the three interior states, for the five corner states, and for all the eight multiattribute health states. Significance testing was not conducted for the analysis at the level of the three interior states on account of the small sample size. Regarding the analysis at the level of the five corner states, the appropriate p-value can be read off the right tail of the sampling distribution of either W or S (Gibbons, 1993). For the purpose of this study, the sampling distribution of S was used for hypothesis testing, which is appropriate for small sample sizes (Kendall and Gibbons, 1990). Regarding the analysis at the level of all eight multiattribute health states, Gibbons' (1993) recommendation of using the distribution of Q for sample sizes greater than five was followed; thus, the significance of Kendall's W at the level of all eight multiattribute health states was evaluated using the following computational formula for Q:

$$Q = \frac{12 S}{k n (n + 1)} = k (n - 1) W \quad (5.9)$$

where

- $Q$ = test statistic

and all other terms have been previously defined.

Gibbons (1993) has noted that for such sample sizes, the distribution of Q is approximately chi-square with $(n - 1)$ degrees of freedom. Thus, the approximate right-tail p-value was found from a chi-square table.
In multiple hypothesis testing, it is important to consider the cumulative probability of committing at least one Type I error. Specifically, when multiple hypotheses are tested -- and each test has a specified Type I error probability -- the probability that at least one Type I error is committed increases, often dramatically, with the number of hypotheses that are tested (Shaffer, 1995). This is known as the phenomenon of alpha slippage. Numerous methods have been proposed for dealing with this phenomenon. All begin with the definition of a family of multiple comparisons, defined as the set of hypotheses for which significance statements will be considered and errors controlled jointly (Shaffer, 1995), followed by implementation of a particular testing method which maintains the familywise error (i.e., the cumulative Type I error probability for the entire family of multiple comparisons) at this prespecified level.

In order to account for alpha slippage due to the multiple comparisons which were conducted to address this research question, a Bonferroni correction was applied to the a priori p-value of 0.05. Keppel (1982) has recommended a Bonferroni correction be used in case of planned comparisons on a data matrix. Further, unlike more stringent tests for multiple comparisons which can only be applied to comparisons between normally distributed means, the Bonferroni correction is general in its application, and can be applied to various types of comparisons (Shaffer, 1995).

Two families of multiple comparisons were defined -- the first on the data matrix of ordinal rankings for the five corner states and the second on the data matrix of ordinal rankings for all eight multiattribute health states.
Since a total of seven comparisons were planned for each of these data matrices (one for the present case, three with respect to the significance of Kendall's τ, and three with respect to the significance of Spearman's rho, both of which are discussed later), each hypothesis was evaluated at an alpha level of 0.05/7 or 0.007. In this way, the familywise error in both cases was maintained at 0.05.

In summary, for the purpose of addressing this research question, the following analytical procedures were followed: (a) Kendall's W was calculated using Equation (5.7), (b) the significance of Kendall's W was determined using the sampling distribution of either $S$ or $Q$ (as explained above) using a Bonferroni correction on the $a priori$ alpha level.

**Research Question # 8(a):** Do the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology agree in terms of the order in which they rank alternative states of health?

**Research Question # 8(b):** Do the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM agree in terms of the order in which they rank alternative states of health?
Research Question # 8(c): Do the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM agree in terms of the order in which they rank alternative states of health?

The data used to address this set of questions consisted of the preference scores of the eight multiattribute health states (i.e., five corner states and three interior states) determined via the three preference measurement methodologies, as explained in the previous section. The analysis was conducted at three different levels: (a) at the overall level (i.e., all eight multiattribute health states), (b) at the level of the five corner states separately, and (c) at the level of the three interior states separately.

This set of research questions was addressed using Kendall's $\tau$ to indicate agreement in preferential ordering of multiattribute health states between each of the preference measurement methodologies taken two at a time. Kendall's $\tau$ measures the extent of agreement in the rank ordering of a variable measured by two different measures. Specifically, it measures agreement between two different measures as the proportion of concordant pairs (i.e., pairs of observations for which both measures order the variable in the same way) minus the proportion of discordant pairs (i.e., pairs of observations for which the two measures order the variable in different ways) in the sample. In other words, Kendall’s $\tau$ is the proportion of pairs that agree minus the proportion of pairs that disagree. It is bounded between -1 and +1, with -1 indicating perfect disagreement between the two
measures, +1 indicating perfect agreement between the two measures, and 0 indicating no agreement between the two measures. However, if there are ties present in any of the rankings, the maximum absolute value of Kendall's τ is no longer 1 and a correction factor for ties needs to be incorporated in its computational formula (Gibbons, 1993). As noted earlier in this section, there were no ties in any of the rankings of multiattribute health states produced by any of the three preference measurement methodologies investigated in this study; therefore, the correction factor for ties will not be discussed further.

Since Kendall's τ is a descriptive measure of agreement between the rankings of two measures, it can also be used to test the null hypothesis of no agreement among the rankings produced by the two measures. It was decided to conduct a one-tailed significance test for hypothesis testing purposes, since the motivation behind this research question was to determine whether there was agreement -- not disagreement -- between the preference measurement methodologies. As explained earlier, since all three methodologies measure the same construct, it is important to measure the extent of agreement among them. Further, use of a one-tailed test provides one with more statistical power for hypothesis testing. Therefore, the alternative hypothesis was that agreement exists among the rankings produced by the two measures.

Unlike the case of Kendall's W described above, SAS readily calculates Kendall's τ. Accordingly, SAS was used to compute Kendall's τ for the three levels of the analysis. Significance testing was again not conducted for the analysis at the level of the three interior states on account of the small
sample size. Regarding the analysis at the levels of the five corner states and at the overall level of all eight multiattribute health states, the appropriate p-value was read off the right tail of the sampling distribution of τ (Gibbons, 1993).

In order to account for alpha slippage due to the multiple comparisons which were conducted to address this research question, a Bonferroni correction was applied to the a priori p-value of 0.05. The reasons for selecting a Bonferroni correction were the same as described earlier under the discussion for Kendall’s W; therefore, they will not be repeated here. For this purpose, two families of multiple comparisons were defined -- the first on the data matrix of ordinal rankings for the five corner states and the second on the data matrix of ordinal rankings for all eight multiattribute health states. Since a total of seven comparisons were planned for each of these data matrices (three with respect to the significance of Kendall’s τ, one with respect to the significance of Kendall’s W, and three with respect to the significance of Spearman’s rho), each hypothesis was evaluated at an alpha level of 0.05/7 or 0.007. In this way, the familywise error in both cases was maintained at 0.05.

In summary, for the purpose of addressing this set of research questions for this study using Kendall’s τ, the following analytical procedures were followed: (a) Kendall’s τ was calculated using the computational procedure in SAS, (b) the significance of Kendall’s τ was determined using the sampling distribution of τ using a Bonferroni correction on the a priori alpha level.
Magnitude of Preference Scores

The second aspect on which the different preference measurement methodologies were compared was in terms of the magnitude of the preference scores they ascribed to the alternative health states under evaluation. The specific research question investigated in this regard was stated as follows:

**Research Question # 9:** Do the three health status preference measurement methodologies (i.e., the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) agree in terms of the magnitude of the preference scores they ascribe to alternative states of health?

The data used to address this question consisted of the preference scores of the eight multiattribute health states (i.e., five corner states and three interior states) determined via the three preference measurement methodologies. This question was addressed using a repeated measures one-way analysis of variance with a Scheffe test for multiple comparisons. Since the rationale behind controlling for alpha slippage on account of multiple comparisons was described in the previous section of this chapter, it will not be repeated here. It was decided to use a Scheffe test as opposed to any other tests for multiple comparisons for this particular research question because
the Scheffe test is the most conservative and robust among the different tests for multiple comparisons under conditions of violation of sphericity, as explained below.

The null hypothesis for this test was that there were no significant differences between the preference scores obtained using the three health status preference measurement methodologies, while the alternative hypothesis was that at least one of the scores was significantly different than the others.

In repeated measures designs, it is important to consider the assumptions made concerning the correlations among multiple measures obtained from the same subject in the analysis of data collected using such designs. In this regard, the assumption of compound symmetry is important -- if the covariance matrix of the repeated measures has equal variances and equal covariances then the covariance matrix is said to have compound symmetry. Under conditions of compound symmetry, the usual F-tests are valid. More generally, Rouanet and Lepine (1970) as well as Huynh and Feldt (1970) have shown that the usual F-tests remain valid when the covariance matrix of the repeated measures displays the property of sphericity.

If, however, the covariance matrix of the repeated measures does not display the property of sphericity, the sampling distribution of the F-ratio is not distributed as F when the null hypothesis is true -- therefore, standard F-tables cannot be used to directly judge the significance of an observed F-value. This is because the actual sampling distribution shifts to the right of the central F-distribution when the covariance matrix of the repeated
measures does not display the property of sphericity -- thus, the critical values obtained from the tabled values of the F-distribution would be too small, leading to a much more lenient critical level than necessary, which in turn would lead to rejecting the null hypothesis falsely a greater percentage of the time than necessary.

Therefore, a necessary first step in hypothesis testing in the case of repeated measures designs is to determine whether the covariance matrix of the repeated measures displays the property of sphericity. SAS offers such a test for sphericity. Depending on the result of this test, one can decide whether to use the usual F-test or an adjusted F-test for the purpose of hypothesis testing.

If the covariance matrix of the repeated measures displays the property of sphericity, the usual F-test may be used for hypothesis testing. However, if the covariance matrix of the repeated measures does not display the property of sphericity, an adjustment to numerator and denominator degrees of freedom is needed for the purpose of hypothesis testing. SAS provides two such adjustments, both of which are based on a degrees of freedom adjustment factor known as epsilon or $\varepsilon$ (Box, 1954). Both adjustments estimate $\varepsilon$ and then multiply the numerator and denominator degrees of freedom by this estimate before determining significance levels for the F-tests.

The first estimate is the Greenhouse-Geisser (1958) estimate, which is the maximum likelihood estimate of $\varepsilon$ under normality, while the second is the Huynh and Feldt (1976) estimate, which is the ratio of unbiased estimates of the numerator and denominator of the population $\varepsilon$. Of the two
estimates, the Greenhouse-Geisser (1958) estimate is the more conservative (Huynh and Feldt, 1976).

If the omnibus F-test (as described above) is significant, multiple comparisons need to be conducted. If the covariance matrix of the repeated measures displays the property of sphericity, the properties of the regular multiple comparison tests regarding the control of familywise error remain unchanged. If, however, the covariance matrix of the repeated measures does not display the property of sphericity, the efficacy of the different multiple comparison tests in controlling familywise error is differentially affected. In a simulation study, Thomson (1990) showed that the Scheffe test provided the most stringent control of familywise error under conditions of nonsphericity. Therefore, rather than making an *a posteriori* assessment and decision regarding which particular multiple comparison test to use, it was decided to use a Scheffe test on account of its stringent control of familywise error under conditions of nonsphericity.

For the purpose of addressing this research question for this study, the following analytical procedures were followed: (a) a repeated measures one-way analysis of variance was run, (b) the results of the test for sphericity of the covariance matrix of repeated measures were used to determine whether the usual or adjusted F-test should be used for hypothesis testing of the null hypothesis, (c) if the omnibus F-test in (b) above was significant, multiple comparison tests were conducted using the Scheffe test for multiple comparisons in order to control the familywise error.
Correlations

The third aspect on which the different preference measurement methodologies were compared was in terms of the correlations between the preference scores of multiattribute health states obtained using the different methodologies. Two specific research questions were investigated in this regard. The data analysis procedures for each will be discussed in turn in this sub-section.

Research Question # 10(a): Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Research Question # 10(b): Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

Research Question # 10(c): Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?
The data used to address this set of questions consisted of the preference scores of the eight multiattribute health states (i.e., five corner states and three interior states) determined via the three preference measurement methodologies. As in the case of the previous questions, the analyses conducted in order to address this set of research questions were performed at three levels -- all eight multiattribute health states together, the five corner states, and the three interior states.

This set of research questions was addressed using Spearman’s rho or \( \rho \). Spearman’s \( \rho \) is a measure of the correlation between two measures in terms of the order in which they rank a particular variable. i.e., it is a rank-order correlation coefficient between two measures. It ranges in value between -1 and +1, where -1 indicates perfect negative correlation between the rank ordering of the two measures, +1 indicates perfect positive correlation between the rank ordering of the two measures, and 0 indicates no relationship or association between the rank ordering of the two measures. As with the case of Kendall’s \( \tau \) (which was explained earlier in this section), if there are ties present in any of the rankings, the maximum absolute value of is no Spearman’s \( \rho \) is longer 1 and a correction factor for ties needs to be incorporated in its computational formula (Gibbons, 1993). As noted earlier in this section, there were no ties in any of the rankings of multiattribute health states produced by any of the three preference measurement methodologies investigated in this study; therefore, the correction factor for ties will not be discussed further.

Since Spearman’s \( \rho \) is a descriptive measure of correlation or association between the rankings of two measures, it can also be used to test
the null hypothesis of no correlation or association among the rankings produced by the two measures. As noted in the case of Kendall's $\tau$ earlier in this section, it was decided to conduct a one-tailed significance test for hypothesis testing purposes. Since the reasons for choosing a one-tailed significance were the same as those discussed for Kendall's $\tau$, they will not be repeated here. The alternative hypothesis was that correlation or association exists among the rankings produced by the preference measurement methodologies, taken two at a time.

SAS readily calculates the value of Spearman's $\rho$. Accordingly, SAS was used to compute Spearman's $\rho$ for the three levels of the analysis. Significance testing was again not conducted for the analysis at the level of the three interior states on account of the small sample size. Regarding the analysis at the levels of the five corner states and at the overall level of all eight multiattribute health states, the appropriate $p$-value was read off the right tail of the sampling distribution of $\rho$ (Gibbons, 1993).

In order to account for alpha slippage due to the multiple comparisons which were conducted to address this research question, a Bonferroni correction was applied to the a priori $p$-value of 0.05. The reasons for selecting a Bonferroni correction were the same as described earlier under the discussion for Kendall's $W$; therefore, they will not be repeated here. For this purpose, two families of multiple comparisons were defined -- the first on the data matrix of ordinal rankings for the five corner states and the second on the data matrix of ordinal rankings for all eight multiattribute health states. Since a total of seven comparisons were planned for each of these data matrices (one with respect to the significance of Kendall's $W$,
three with respect to the significance of Kendall's $\tau$ as discussed earlier and three with respect to the significance of Spearman's $\rho$), each hypothesis was evaluated at an alpha level of 0.05/7 or 0.007. In this way, the familywise error in both cases was maintained at 0.05.

In summary, for the purpose of addressing this set of research questions, the following analytical procedures were followed: (a) Spearman's $\rho$ was calculated using the computational procedure in SAS, (b) the significance of Spearman's $\rho$ was determined using the sampling distribution of $\rho$ using a Bonferroni correction on the a priori alpha level.

**Research Question # 11(a):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

**Research Question # 11(b):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

**Research Question # 11(c):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling
method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?

This set of research questions was addressed using Pearson’s product moment correlation coefficient or Pearson’s r. Pearson’s r is a measure of the linear correlation between two measures in terms of the scores they give to a particular variable. It ranges in value between -1 and +1, where -1 indicates perfect negative correlation between the rank ordering of the two measures, +1 indicates perfect positive correlation between the rank ordering of the two measures, and 0 indicates no relationship or association between the rank ordering of the two measures.

Since Pearson’s r is a descriptive measure of correlation or association between the scores obtained by two measures, it can also be used to test the null hypothesis of no correlation or association among the scores obtained by the two measures. As noted in the case of Spearman’s ρ earlier in this section, it was decided to conduct a one-tailed significance test for hypothesis testing purposes. Since the reasons for choosing a one-tailed significance were the same as those discussed for Spearman’s ρ, they will not be repeated here. The alternative hypothesis was that correlation or association exists among the rankings produced by the preference measurement methodologies, taken two at a time.

Pearson’s r was computed using SAS. Significance testing was again not conducted for the analysis at the level of the three interior states on account of the small sample size. Regarding the analysis at the levels of the
five corner states and at the overall level of all eight multiattribute health states, the sample Pearson’s r and sample sizes were substituted into Equation (5.10):

\[ t = r \sqrt{(n - 2)} / \sqrt{(1 - r^2)} \quad (5.10) \]

where

- \( r \) = sample Pearson product moment correlation coefficient
- \( n \) = sample size
- \( t \) = observed test statistic for hypothesis testing

The appropriate \( p \)-value was read off the right tail of the sampling distribution of the \( t \) distribution with \((n - 2)\) degrees of freedom.

In order to account for alpha slippage due to the multiple comparisons which were conducted to address this research question, a Bonferroni correction was applied to the \emph{a priori} \( p \)-value of 0.05. The reasons for selecting a Bonferroni correction were the same as described earlier under the discussion for Kendall’s \( W \); therefore, they will not be repeated here. For this purpose, two families of multiple comparisons were defined -- the first on the data matrix of preference scores for the five corner states and the second on the data matrix of preference scores for all eight multiattribute health states. Since a total of seven comparisons were planned for each of these data matrices (one with respect to the significance of the overall intraclass correlation coefficient, three with respect to the significance of the pairwise intraclass correlation coefficients as discussed later, and three with respect to the significance of Pearson’s r), each hypothesis was evaluated at
an alpha level of 0.05/7 or 0.007. In this way, the familywise error in both cases was maintained at 0.05.

In summary, for the purpose of addressing this research question for this study using Pearson's $r$, the following analytical procedures were followed: (a) Pearson's $r$ was calculated using the computational procedure in SAS, (b) the test statistic was calculated using Equation (5.10), and (c) the significance of the observed test statistic was determined using the sampling distribution of $t$ using a Bonferroni correction on the \textit{a priori} alpha level.

**Concordances**

The fourth aspect on which the different preference measurement methodologies were compared was in terms of the concordance among the preference scores of multiattribute health states obtained using the different methodologies. The research questions related to this aspect were addressed using the intraclass correlation coefficient or ICC. Unlike the other statistics discussed in this sub-section, the ICC has not been commonly used in health status research. Therefore, a brief discussion on the ICC is appropriate.

The ICC has been defined by Deyo, Diehr, and Patrick (1991) as the proportion of total variability accounted for by the variability among persons (or more generally, the objects under evaluation in any study, e.g., multiattribute health states for the purpose of this study). Computationally, it can be viewed as the ratio of the variance of interest in the study to the sum of the variance of interest plus error (Shrout and Fleiss, 1979).

It is instructive to note the difference between the Pearson $r$ and the ICC. The Pearson $r$ is a statistic indicating the extent of the linear correlation
or association between two measures. Although the Pearson r is commonly used in the health status measurement literature as a measure of correlation between two measures, its disadvantage is that it does not account for systematic bias between measures. In other words, it is possible that two measures may be systematically different from each other but still be highly correlated with one another. This point is especially important to note for the purpose of this study, since it was expected that the preference scores of multiattribute health states would be significantly different from each other on account of differences in underlying theoretical principles (e.g., value versus utility measurement).

The ICC, on the other hand, assesses not only the strength of correlation, but also a test for the difference in means of the measures. As Kramer and Feinstein (1981) have noted, the ICC assesses both, the similarity in slopes (like the Pearson r) as well as the similarity in intercepts. Thus, unlike the Pearson r, if one measure is systematically higher or lower than the others, the ICC will correspondingly be lowered. Thus, the ICC provides an estimate of the correlation between two or more measures after adjusting for the fact that the measures might be significantly different from one another. In other words, if offers an estimate of the correlation corrected for systematic bias.

Another important difference between the ICC and the Pearson r is that the ICC is not restricted to the case of only two measures; it can be extended to several measures. Like the Pearson r, the values of the ICC range from -1 to +1, with -1 and +1 indicating perfect disagreement and agreement.
among the measures respectively and 0 indicating random agreement among the measures.

There are several different types of ICCs; the particular ICC chosen for use in any study depends on the assumptions of the study design and analysis. The first ICCs were developed at the end of the last century in the field of genetics (Muller and Buttner, 1994). First, a special formulation of Pearson’s r was defined for this purpose, which was reached by assuming equality of means and variances. Subsequently, Fischer (1925) introduced the ICC into the field of psychometrics, by showing that the special formulation referred to above was equivalent to a one-way analysis of variance (ANOVA), where the total variance was split into within-subject and between-subject variability (Muller and Buttner, 1994). This one-way ANOVA model has come to be known as the Model I formulation of the ICC.

Since then, more complicated ANOVA models of the ICC have been developed (see Bartko, 1966; Shrout and Fleiss, 1979). Of these models, a couple of models based on two-way ANOVAs have gained the most popularity -- the first, which has come to be known as the Model II formulation of the ICC, assumes that the raters (health status preference measurement methodologies for the purpose of this study) are a random sample, while the second, which has come to be known as the Model III formulation of the ICC, assumes that the raters are fixed. For both, Model II as well as Model III, the within-subject (within health state for the purpose of this study) variance component is split into a term which yields variability among the raters or preference measurement methodologies, an interaction
between the raters and the subjects, and an error term. Thus, both models permit the separation of bias or systematic error from random error (Muller and Buttner, 1994).

Given this multitude of formulations of the ICC, it is important to select the correct formulation for use in any particular study. This is because the different formulations can give quite different results when applied to the same data (see Muller and Buttner, 1994 for a demonstration). Bartko (1966), Muller and Buttner (1994), and Shrout and Fleiss (1979) have offered guidelines for deciding which particular formulation of the ICC to use in any study. In this regard, there are two major questions which guide the decision: (a) is a one-way or two-way ANOVA appropriate for the analysis? and (b) can the raters (preference measurement methodologies) be considered random effects or fixed effects? It can be seen that the first question differentiates between Model I on the one hand and Models II and III on the other, while the second question is specifically concerned with the choice between Models II and III.

For the purpose of this study: (a) a two-way ANOVA was considered appropriate, since such a formulation would permit the separation of between-raters variability (i.e., variability between the different health status preference measurement methodologies) and (b) the raters (i.e., preference measurement methodologies) were fixed, since they were specifically chosen from the different preference measurement methodologies which could have been used (as explained in Chapter II). Therefore, it was decided to adopt the Model III formulation of the ICC (also called the mixed model formulation) for the purpose of this study.
Specifically, two types of ICCs were calculated in order to address this research question. The first considered all three health status preference measurement methodologies together, while the second considered the methodologies two at a time. Both types of ICCs used the Model III formulation described above in their computation. As in the case of the other analyses described, both the analyses on the ICC were conducted at three different levels -- all eight multiattribute health states, the five corner states only, and the three interior states only.

Against this backdrop, the data analysis procedures for the specific research questions investigated regarding the concordance among the preference scores obtained using the three different preference measurement methodologies can be discussed. On account of the fact that both the research questions investigating concordance were addressed using the ICC and followed similar data analysis procedures, they will be discussed together. The differences in the data analysis procedures for the two questions will be pointed out in the discussion when appropriate.

**Research Question # 12**: Is there significant concordance among the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM?
Research Question # 13(a): Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Research Question # 13(b): Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

Research Question # 13(c): Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?

Research Question # 12 was addressed using the overall ICC among the three different preference measurement methodologies, while Research Questions # 13(a), 13(b), and 13(c) were addressed using the ICC between each of the two measurement methodologies in question. Since the ICC is a descriptive measure of concordance or agreement between the scores obtained by two or more measures, it can also be used to test the null hypothesis of no concordance or agreement among the scores obtained by these two or more measures. As explained for the other hypotheses tested in
this section, it was decided to conduct a one-tailed significance test for hypothesis testing purposes. The alternative hypothesis was that correlation or association exists among the rankings produced by the preference measurement methodologies.

Since SAS does not readily compute the ICC, the formula for the ICC had to be programmed into SAS. The specific formula used to compute the ICC is shown in Equation (5.11):

\[
\text{ICC} = \frac{\text{MSB} - \text{MSE}}{\text{MSB} + (k-1) \text{MSE}} \tag{5.11}
\]

where

\begin{align*}
\text{MSB} & = \text{between subjects (health states) mean square} \\
\text{MSE} & = \text{mean square error} \\
k & = \text{number of raters (preference measurement methodologies)} \\
\text{ICC} & = \text{intraclass correlation coefficient}
\end{align*}

The MSB and MSE values were obtained from a two-way ANOVA, which was run using SAS. These values were then substituted into Equation (5.11), in order to calculate the ICC. The only difference in the computational procedures for the overall and pairwise ICCs was in the value of \( k \) -- 3 for the computation of the overall ICC and 2 for the computation of the pairwise ICCs. The significance of the ICC was assessed on the basis of the F-ratio (provided by SAS) of MSB and MSE, with \((n - 1)\) and \((n - 1)(k - 1)\) degrees of freedom (Bartko, 1966, 1974; Shrout and Fleiss, 1979). As before, significance testing was again not conducted for the analysis at the level of the three interior states on account of the small sample size.
In order to account for alpha slippage due to the multiple comparisons which were conducted to address this research question, a Bonferroni correction was applied to the \textit{a priori} $p$-value of 0.05. The reasons for selecting a Bonferroni correction were the same as described earlier under the discussion for Kendall’s $W$; therefore, they will not be repeated here. For this purpose, two families of multiple comparisons were defined — the first on the data matrix of preference scores for the five corner states and the second on the data matrix of preference scores for all eight multiattribute health states. Since a total of seven comparisons were planned for each of these data matrices (one with respect to the significance of the overall intraclass correlation coefficient, three with respect to the significance of the ICC, and three with respect to the significance of Pearson’s $r$ as discussed earlier), each hypothesis was evaluated at an alpha level of $0.05/7$ or $0.007$. In this way, the familywise error in both cases was maintained at 0.05.

In summary, for the purpose of addressing this research question for this study using the ICC, the following analytical procedures were followed: (a) a two-way ANOVA was run on the data matrix of preference measurement scores using SAS, (b) the ICC was calculated using Equation (5.11), and (c) the significance of the observed $F$-ratio was estimated using a Bonferroni correction on the \textit{a priori} alpha level.

\textbf{Functional Relationships}

It was noted in Chapter II that estimating functional relationships between preference functions obtained using different preference measurement methodologies provide strong evidence for the construct
validity of the methodologies according to the principle of convergent validation. Three specific research questions were investigated in this regard, pertaining to the functional relationship between each set of two preference measurement methodologies. The first two of these research questions were concerned the functional relationship between value and utility functions, while the third was concerned with the functional relationship between two value functions. The data analysis procedures for the first two questions will be discussed together, followed by the third question.

The data used to address this set of questions consisted of the preference scores of the eight multiattribute health states (i.e., five corner states and three interior states) determined via the three preference measurement methodologies. This research question was addressed by using regression analysis. As in the case of the previous research questions, the analysis was conducted at three levels -- all eight multiattribute health states, the five corner states only, and the three interior states only.

**Research Question # 14(a): What is the nature of the functional relationship between the preference functions for alternative health states obtained using the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?**
Research Question #14(b): What is the nature of the functional relationship between the preference functions for alternative health states obtained using the discrete choice conjoint methodology or DCCM and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

As explained in Chapter IV, the relationship between value and utility functions — i.e., CUF-RS and CUF-SG and also DCCM and CUF-SG methodologies — was expected to be represented by a power function.

Notationally, the expected relationships between the value and utility functions investigated in this study were:

\[ \text{CUF-SG} = \text{CUF-RS}^x \]  
\[ \text{CUF-SG} = \text{DCCM}^y \]

where

- CUF-SG = conditional utility function-based procedure with standard gamble scaling method
- CUF-RS = conditional utility function-based procedure with rating scale scaling method
- DCCM = discrete choice conjoint methodology

\( x \) and \( y \) are constants

The methodology adopted by Torrance, Zhang, Feeny, et al, (1992) was followed in determining the functional relationship between the value...
and utility functions. Accordingly, the relationship was estimated on the basis of the observed disvalues and disutilities (not their mathematically equivalent value and utility counterparts). Therefore, notationally, the expected relationship in disvalue and disutility terms can be represented as follows:

\[(1 - CUF-SG) = (1 - CUF-RS)^x \]  
\[(1 - CUF-SG) = (1 - DCCM)^y \]

where

- CUF-SG = conditional utility function-based procedure with standard gamble scaling method
- CUF-RS = conditional utility function-based procedure with rating scale scaling method
- DCCM = discrete choice conjoint methodology
- \(x\) and \(y\) are constants

The task was to fit a curve through the appropriate number of points (three, five, or eight, depending on the level of the analysis) such that it also passed through the points \((0,0)\) and \((1,1)\). This is necessary to satisfy the health scale model, which is anchored at 0.0 and 1.0 (Torrance, 1976a). The first step in the analysis involved transforming the observed (in case of the CUF-RS and CUF-SG methodologies) and calculated (in case of the DCCM methodology) disvalues and disutilities into their natural logarithms. The curve fitting process then used straight line regression through the origin in
order to determine the parameters of the model. It was necessary to force the regression line through the origin in order to make it include the anchor points of the scale -- i.e., (0,0) or the all-best state on the disvalue and disutility scale and (1,1) or the state of immediate death on the disvalue and disutility scale. In both cases, the utilities (i.e., preference scores obtained using the CUF-SG methodology) were the dependent variable and the values (i.e., preference scores obtained using the CUF-RS or DCCM methodologies) were the independent variables. In order to determine the goodness of fit of the fitted regression equation, adjusted $R^2$ values were calculated. It was decided to use adjusted $R^2$ values because of the small sample sizes for all the levels of the analysis. The computational formula for the adjusted $R^2$ (also called the shrunken $R^2$) is given by:

$$\text{adj } R^2 = 1 - \frac{(1 - R^2) n}{(n - k - 1)}$$

(5.16)

where

- $n$ = sample size
- $k$ = number of independent variables in the regression equation
- $R^2$ = observed $R^2$ value
- adj $R^2$ = adjusted $R^2$ value

The problems of alpha slippage in multiple hypothesis testing was explained earlier in this chapter, and therefore will not be repeated here. In order to account for alpha slippage due to the multiple comparisons which were conducted to address this research question, a Bonferroni correction was applied to the $a$ priori $p$-value of 0.05. For this purpose, two families of
multiple comparisons were defined — the first on the data matrix of preference scores for the five corner states and the second on the data matrix of preference scores for all eight multiattribute health states (as explained before, significance testing was not conducted for the analysis of the three interior states on account of the small sample size). Since a total of three comparisons were planned for each of these data matrices (i.e., one for each estimated functional relationship), each hypothesis was evaluated at an alpha level of 0.05/3 or 0.017. In this way, the familywise error in both cases was maintained at 0.05.

In summary, for the purpose of estimating the functional relationship between the CUF-RS and CUF-SG as well as between the DCCM and CUF-SG methodologies, the following procedures were followed: (a) disvalues and disutilities were transformed into their natural logarithms, (b) straight line regression through the origin was used to estimate the parameters of the model, (c) adjusted $R^2$ values were calculated according to the formula in Equation (5.12), (d) the significance of the regression equation was determined using a Bonferroni correction on the a priori alpha level.

**Research Question # 14(c): What is the nature of the functional relationship between the preference functions for alternative health states obtained using the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?**
As explained in Chapter IV, the relationship between two value functions -- i.e., the CUF-RS and DCCM methodologies -- was expected to be represented by a simple linear function. Notationally, the expected relationship between the two value functions investigated in this dissertation was:

\[
\text{DCCM} = a + b \times \text{CUF-RS} \tag{5.17}
\]

where

- \( \text{CUF-RS} \) = conditional utility function-based procedure with rating scale scaling method
- \( \text{DCCM} \) = discrete choice conjoint methodology
- \( a \) and \( b \) are constants

This functional relationship was also estimated using straight line regression. In order to be consistent with procedures followed for the estimation of functional relationships between the other preference measurement methodologies investigated in this study (as described above) it was decided to use the disvalue formulations of the preference scores for the purpose of this analysis. Therefore, notationally, the expected relationship between the two disvalue functions investigated in this study was:

\[
(1 - \text{DCCM}) = a + b \times (1 - \text{CUF-RS}) \tag{5.18}
\]
where

\[ \text{CUF-RS} = \text{conditional utility function-based procedure with rating scale scaling method} \]

\[ \text{DCCM} = \text{discrete choice conjoint methodology} \]

\[ a \text{ and } b \text{ are constants} \]

The fitting process used straight line regression in order to determine the parameters of the model. However, in this case, the line was not forced to pass through the origin. The preference scores obtained using the DCCM methodology were treated as the dependent variable in the regression analysis, while the preference scores obtained using the CUF-RS methodology were treated as the independent variable. Again, adjusted $R^2$ values were calculated for the same reasons discussed above. The significance of the regression equation was evaluated using a Bonferroni correction on the *a priori* alpha level, as explained earlier in this section.

In summary, for the purpose of estimating the functional relationship between the CUF-RS and DCCM methodologies, the following procedures were followed: (a) the disvalue formulations of the preference scores were used as the data for the analysis, (b) straight line regression (not through the origin) was used to estimate the parameters of the model, (c) adjusted $R^2$ values were calculated according to the formula in Equation (5.12), (d) the significance of the regression equation was determined using a Bonferroni correction on the *a priori* alpha level.
Research Questions Pertaining to Comparing Respondent Evaluations

As noted in Chapter II, Research Questions 15 through 18 pertained to comparing the respondent evaluations of the three preference measurement methodologies. This section will discuss the analytical procedures followed for each of these four questions. Each question will be discussed in turn.

Reliability

**Research Question # 15(a):** What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?

**Research Question # 15(b):** What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

**Research Question # 15(c):** What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the discrete choice conjoint methodology or DCCM?

As noted in Chapter II, before directly addressing the issue of respondent evaluations of the different preference measurement
methodologies it was necessary to determine the psychometric properties -- i.e., reliability and validity -- of the data collected with the instrument developed to measure respondent evaluations of the individual preference measurement methodologies. This research question was the first of two for this purpose.

Reliability refers to the extent to which a score is free of measurement error (Nunnally, 1967, 1978). Since the instrument used to collect data on respondent evaluations of the three preference measurement methodologies was only administered once, it was not possible to estimate reliability by means of test-retest reliability coefficients. For the purpose of this study, it was decided to estimate the reliability on the basis of a measure of the average correlation among items of the measurement instrument, i.e., a measure of internal consistency. Reliability as measured by internal consistency has been defined as the extent to which all items of a multi-item scale measure the same underlying construct, or the convergence of the items on the construct being measured (Stewart, 1990). This is especially appropriate for the context of this study, wherein the estimation of reliability focused on the extent to which all items of the eight-item semantic differential measured the overall construct of respondent evaluation of each preference measurement methodology.

The specific measure of internal consistency which was used for the purpose of this study was coefficient alpha or Cronbach's alpha (Cronbach, 1951). Cronbach's alpha is the appropriate reliability coefficient for estimating the reliability of data collected using multi-item Likert-type or semantic differential scales which are summated, i.e., derived by adding
together several items that have similar response scales (Stewart, 1990). Indeed, it is the most widely used reliability coefficient for such purposes in social science research.

Peterson (1994) has aptly summarized the popularity of Cronbach's alpha as a reliability coefficient: "Whether by acclamation (see, e.g., Churchill, 1979; Gerbing and Anderson, 1988; Peter, 1979) or citation, coefficient alpha has effectively become the measure of choice for estimating the reliability of a multi-item scale. Indeed, coefficient alpha has become one of the foundations of measurement theory. Because it is a generalized intraclass correlation coefficient, coefficient alpha can be derived from the theory of true and error scores, as well as from the domain sampling model. According to the Social Science Citation Index, Cronbach's 1951 article has been referenced in more than 2,200 articles in the last 20 years" (p.382). SAS readily performs the necessary computation for Cronbach's alpha; therefore, for the purpose of estimating the reliability of the data collected using this instrument, Cronbach's coefficient alpha was calculated using SAS.

Exploratory Factor Analyses

Research Question # 16(a): What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?
Research Question # 16(b): What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Research Question # 16(c): What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the discrete choice conjoint methodology or DCCM?

This was the second of two questions that was intended to investigate the psychometric properties of the data collected with the instrument developed to measure respondent evaluations of the individual preference measurement methodologies. As noted earlier in this chapter, content and face validity of the instrument used to measure respondent evaluations of the three preference measurement methodologies was determined by the panel of experts as well as the test subjects for the pre- and pilot tests. Although a detailed investigation into the construct validity of the respondent evaluation data was considered to be beyond the scope of this study, an attempt was made to understand the structure of respondent evaluations, as explained below.

It was noted in an earlier chapter of this dissertation that one of the unaddressed issues in respondent evaluations of preference measurement methodologies was the lack of a structure to overall respondent evaluations. In an attempt toward addressing this particular issue (and, in the process,
providing an initial step toward the construct validation of the instrument used to measure respondent evaluations of the three preference measurement methodologies), an exploratory factor analysis was conducted on the data collected on the respondent evaluations of each of the three individual preference measurement methodologies.

Factor analysis is a generic name given to a class of multivariate statistical methods whose primary purpose is data reduction and summarization. These methods are concerned with analyzing the interrelationships among a large number of variables and then explaining these variables in terms of their common, underlying dimensions or factors (Hair, Anderson, and Tatham, 1987). In other words, these methods are directed toward determining the structure underlying different constructs. Therefore, factor analysis is well suited to the problem of interest in this particular research question, i.e., determining the underlying structure of respondent evaluations of preference measurement methodologies.

Having decided to use factor analysis to determine the underlying structure of respondent evaluations of preference measurement methodologies, four other decisions pertaining to the implementation of the analytical technique need to be made: (a) determination of the scope of the analysis, (b) selection of the type of the general factor model, (c) selection of the factor extraction method, and (d) selection of the factor rotation method. Each of these four decisions will be discussed in turn, with respect to how they were handled for the purpose of the factor analysis conducted in the course of addressing this particular research question.
Regarding the scope of the analysis, there are two basic types of factor analyses which may be conducted -- exploratory factor analysis and confirmatory factor analysis. As the names might suggest, exploratory factor analysis is best suited for situations in which the researcher has no prior knowledge of or expectation regarding the underlying structure of the construct being investigated; on the other hand, confirmatory factor analysis is best applied to situations in which the researcher is hypothesizing a particular structure underlying the construct being investigated and wishes to test this hypothesis. For the purpose of the situation germane to this study, it was noted earlier that one of the unaddressed issues in respondent evaluations of preference measurement methodologies was the lack of a structure to overall respondent evaluations. Therefore, since the purpose of conducting the factor analysis in this study was to provide some preliminary understanding of the structure underlying respondent evaluations of preference measurement methodologies and not to confirm any hypotheses regarding this structure, it was decided to conduct an exploratory factor analysis for the purpose of this study.

Regarding the type of the general factor model, although numerous variations of the general factor model are available, the two most frequently employed models are the common factor analysis model and the component analysis (also called principal components analysis) model. These two types of models differ in the way they handle the different types of variance. Hair, Anderson, and Tatham (1987) have noted that for the purpose of factor analysis, total variance consists of three types of variance: (a) common variance, which is the variance in a variable which is shared with all other
variables in the analysis, (b) unique or specific variance, which is the variance associated with only a particular variable, and (c) error variance, which is the variance due to the unreliability in the data gathering process or a random component in the measured construct.

In the case of component analysis, the factors that are derived as the result of the analysis are based on the total variance, and therefore contain proportions of unique and sometimes error variance (usually not enough to distort the overall factor structure) in addition to the common variance. Specifically, this is done by inserting unities into the diagonal of the correlation matrix which is used as the input for the analysis. On the other hand, in the case of common factor analysis, the factors that are derived as the result of the analysis are based on only the common variance. Specifically, this is done by inserting communalities into the diagonal of the correlation matrix which is used as the input for the analysis (Hair, Anderson, and Tatham, 1987).

Hair, Anderson, and Tatham (1987) have noted that the decision between the two types of general factor models should be based on two criteria -- the objective of the researcher and the amount of prior knowledge about the variance in the variables. These researchers have recommended that when the focus of the analysis is on prediction and determining the minimum number of factors needed to account for the maximum portion of the variance represented in the original set of variables and when the researcher has prior knowledge suggesting the error and unique variance represent a relatively small portion of the total variance, component analysis should be selected. On the other hand, they recommend that when the
objective of the analysis is on explanation and identifying the latent dimensions represented in the original variables and when the researcher has little knowledge about the amount of error and unique variance (and therefore wishes to eliminate this variance), the common factor analysis model should be selected (Hair, Anderson, and Tatham, 1987).

In the specific situation of interest in this study, i.e., understanding the structure of respondent evaluations of preference measurement methodologies, the objective of the factor analysis is on understanding as opposed to prediction -- specifically, the objective is to understand the latent dimensions in the construct of respondent evaluations of preference measurement methodologies. Moreover, as noted earlier, there is little knowledge about the amount of error and unique variance relative to the total variance. Therefore, it was decided to adopt the common factor analysis model for the purpose of this study.

Regarding the factor extraction methods (also referred to as factor solution methods and methods of estimation of factor loadings), the two most commonly used ones are principal axis factoring (PAF) and maximum likelihood (ML) estimation. The foundation for the method of PAF estimation was built by Pearson (1901), but it was Hotelling (see Harman, 1967, 1968) who developed it in the form of its current usage in factor analysis. This method involves extracting a set of factors in decreasing order of their contributions to the variance (total or common, depending on whether component analysis or common factor analysis is used) of a set of variables. The method begins with the extraction of the first factor, whose contribution to the variance of the variables has as great a total as possible.
Then the first factor residual correlations are obtained. Finally, a second factor with maximum contribution to the residual variance is extracted. The process is continued until all or a suitably large portion of the variance (total or common, depending on whether component analysis or common factor analysis is used) is analyzed (Harman, 1968).

In determining the number of factors to be extracted using PAF estimation, researchers use a variety of criteria, including the latent root criterion (i.e., extract all factors with latent roots or eigenvalues greater than one), a priori criteria regarding the number of factors to expect (in the case of confirmatory factor analysis), the percentage of variance criterion [e.g., Hair, Anderson, and Tatham (1987) note that some social science researchers consider a solution which accounts of 60% of the total variance as satisfactory], and the scree test criterion (i.e., plotting the latent roots against the number of factors in the order of their extraction -- the point at which the curve first begins to straighten out is considered to be the maximum number of factors to extract).

Although the method of PAF estimation has been widely used in psychometric as well as health status research, it is important to note that it is based on the tacit assumption that the sample correlation matrix (which is used as input for the analysis) yields results interpretable in terms of the population from which the sample was drawn. There is no explicit consideration of the statistical problems of the uncertainty of conclusions that might be derived from the empirical data (Harman, 1968). In contrast, the method of ML estimation is a statistically sound method of factor extraction. This method was developed for use in factor analysis by Lawley.
(1940, 1941), who worked on the premise of using Fisher's "method of maximum likelihood" to estimate the universe values of the factor loadings from given empirical data. The ML estimation method requires a hypothesis regarding the number of factors to be extracted, and then, according to this hypothesis, derives a factor solution. The adequacy of the hypothesis regarding the number of factors to be extracted can be tested statistically using the maximum likelihood function (Lawley, 1940, 1941).

As explained earlier (in the section where the principle behind maximum likelihood estimation of the parameters of discrete choice conjoint models was explained), the maximum likelihood function measures the relative likelihood of the occurrence of the data in the population on the basis of the sample data. Since the regular maximum likelihood function is quite complex to work with computationally, maximum likelihood factor analysis is usually implemented with the maximum likelihood discrepancy function, which is inversely related to the regular maximum likelihood function and measures the extent of the difference between the sample data and the population data estimated from the sample data (strictly speaking, it measures the discrepancy between the sample and population covariance matrices). Therefore, when the regular maximum likelihood function is maximized, the maximum likelihood discrepancy function is minimized. In other words, the smaller the value of the maximum likelihood discrepancy function, the better the fit. SAS readily computes the value of the maximum likelihood discrepancy function.

Specifically, two types of null hypotheses can be tested using ML estimation. The first, known as the hypothesis of exact fit, states that the
model with the hypothesized number of factors fits exactly in the population. This is the hypothesis which was initially introduced by Lawley (1940, 1941) in his pioneering work on ML estimation in factor analysis. The test statistic used to test the statistical significance of this hypothesis is maximum likelihood discrepancy function multiplied by the term \((n - 1)\), where \(n\) is the sample size. This test statistic asymptotically follows a chi-square distribution, the tables of which can be referred to in order to determine the significance of the observed test statistic. SAS includes the results of this test as part of its regular output of maximum likelihood factor analysis. Bartlett (1950, 1951) has proposed a correction for the observed value of the chi-square in order to account for small sample sizes (since, strictly speaking, the chi-square distribution is only appropriate in the asymptotic sense). The hypothesis testing procedure used by SAS considers Bartlett’s (1950, 1951) correction for the chi-square in its computation.

Tucker and Lewis (1973) have developed a descriptive measure of fit based on the chi-square value determined as above (referred to as the Tucker-Lewis Index or TLI) which can be used along with the results of significance testing of the hypothesis of exact fit to evaluate the model fit. The use of this index is analogous to variance components analysis for assessing whether the variance accounted for by a theory is large or small compared to the total variance in the data (Heeler, Whipple, and Hustad, 1977). Intuitively, the TLI provides a descriptive measure of where along the continuum of the worst to the best model is the model under evaluation in terms of goodness of fit. In theory, the TLI ranges in value from 0 to 1, with higher values indicating better fit of the model to the data. It should,
however, be noted that it is possible to obtain cases in which the value of the TLI is greater than 1.0; this simply indicates that the fit of the hypothesized model (i.e., number of factors) is greater than that what would be expected at the chosen level of significance, usually 5% -- i.e., the fit is excellent.

It has been noted in the published literature that use of the TLI requires some subjective judgment, since the acceptability of a given value of the TLI will vary with the study data and objectives (Heeler, Whipple, and Hustad, 1977). The acceptability of a given value of the TLI depends on balancing the objectives of explanation and parsimony. However, most published applications of the TLI have found that only very high values indicate good fits, i.e., 0.9 and above (see Bollen, 1989). SAS readily computes the TLI index and includes the value of the TLI index as part of the output of maximum likelihood factor analysis.

The second type of null hypothesis that can be tested using ML estimation is known as the hypothesis of close fit. This hypothesis and the statistical test of significance to test this hypothesis have recently been introduced by Browne and Cudeck (1992). Explaining their rationale for introducing the hypothesis of close fit, they note: "In applications of the analysis of covariance structures in the social sciences, it is implausible that any model we use is anything more than an approximation to reality. Since a null hypothesis that a model fits exactly in some population is known a priori to be false, it seems pointless to try to test whether it is true. ... Rather than trying to ask whether a model is correct, or fits the population covariance matrix exactly, it is sensible to assess the degree of lack of fit of the model" (Browne and Cudeck, 1992 p.231; emphasis added).
Further, Browne and Cudeck (1992) have noted since the hypothesis of exact fit is "... invariably false in practical situations, it will almost certainly be rejected if $n$ is sufficiently large. This discourages the use of large samples. It also encourages the inclusion of uninterpretable additional parameters in a model to reduce the power of the test and thereby avoid the rejection of $H_0$" (p.240). Accordingly, Browne and Cudeck (1992) have proposed the root mean square error of approximation or RMSEA developed by Steiger (Steiger, 1990; Steiger and Lind, 1980) as the test statistic to use for testing the hypothesis of close fit.

Mathematically, the RMSEA is defined as:

$$\text{RMSEA} = \sqrt{\frac{F_n}{df}}$$

where

$F_n$ = population discrepancy function

df = degrees of freedom of the model under evaluation

Further, the population discrepancy function is not known, it can be approximated from the sample value of the maximum likelihood discrepancy function by:

$$\text{est } F_n = \text{Max} \{ F_s - \frac{df}{n-1}; 0 \}$$

where

$F_s$ = sample value of the maximum likelihood discrepancy function

$n$ = sample size
\[ \text{est } F_1 = \text{estimated value of the population discrepancy function} \]

and all other terms have been previously defined.

The RMSEA has a lower bound of zero, and will be zero only if the model fits exactly in the population (note that this property allows the RMSEA to be used as the test statistic for the test of the hypothesis of exact fit also). Although there is no upper bound on the value of the RMSEA, Browne and Cudeck (1992), on the basis of their experience, recommend a value of 0.05 or less as indicating a close fit of the model, while a value of 0.1 or more as indicating poor fit. The interval between 0.05 and 0.1 is like a "gray area," in which the fit of the model needs to be assessed using other supplementary diagnostics, such as confidence intervals for the RMSEA (Browne and Cudeck, 1992). SAS does not provide any computations for the RMSEA. However, FITMOD (Browne, 1992), a computer package designed specifically for this purpose does so; therefore, the computation and significance testing of the RMSEA is easily done.

The advantages of hypothesis testing of ML estimation need to be considered in light of a few potential drawbacks of the method. First, it is computationally more demanding than the PAF estimation method, although the use of high-speed computers somewhat nullifies this potential drawback. Second, it is based on the assumption of multivariate normality. Although responses to single questions are often non-normal, Heeler, Whipple, and Hustad (1977) have noted that scores obtained from summated scales often satisfy the assumption of multivariate normality. Moreover, it should be noted that a simulation study by Fuller and
Hemmerler (1966) found that ML estimation was quite robust to violations of multivariate normality.

In making a decision about which particular factor extraction method to use for the factor analysis conducted as a part of this study, it was considered important to have a theoretical rationale for the selection of a given number of factors. As noted earlier in this section, the PAF estimation procedure has no such rationale, while the ML estimation procedure is firmly grounded in statistical theory. Secondly, a simulation study comparing the PAF and ML estimation methods (among others) found the ML estimation method to be superior in terms of the accuracy of the factor loadings obtained as a result of the estimation procedure (Browne, 1968). Therefore, it was decided to use ML estimation as the factor extraction method for the purpose of this study.

Specifically, a series of models were fit to the data, each with successively greater number of factors extracted than the previous one. The fit of each of these successive model was evaluated using both, tests of exact as well as tests of close fit. The statistics necessary to test the hypothesis of exact fit were obtained using SAS, while the statistics necessary to test the hypothesis of close fit were obtained using FITMOD (Browne, 1992). Apart from information about the statistical significance of the estimated model which were obtained as a result of the ML estimation procedure, Joreskog's (1971) advice that the ultimate criteria for the goodness of a model were its usefulness and the results it produces was adhered to in terms of being able to infer a meaningful interpretation of the fitted model.
The final decision that needs to be made in the conduct of any factor analysis study concerns the factor rotation method. The term rotation simply means that the reference axes of the factors are turned about the origin until some other position is reached (Child, 1970). The results of the factor extraction (as described above) yield an initial unrotated matrix of factor loadings. These unrotated factor solutions extract factors in order of their importance; therefore, the first factor tends to be a general factor with almost every variable loading significantly, and it accounts for the largest amount of variance. The second and subsequent factors will account for successively smaller portions of variance (Hair, Anderson, and Tatham, 1987). In other words, although the unrotated factor solution achieves the objective of data reduction, it does not provide useful information in order to permit meaningful interpretation of the factor solution.

The basic reason for rotation of the factors is to achieve simpler and theoretically more meaningful factor solutions. The ultimate effect of rotating the factor matrix is to redistribute the variance from the earlier factors to the later factors to achieve a simpler and theoretically more meaningful factor solution (Hair, Anderson, and Tatham, 1987). A set of criteria for the determination of a more meaningful factor solution has been formulated by Thurstone (1947), in what is known as his principle of "simple structure." Harman (1967) has restated Thurstone's criteria for simple structure: "(a) each row of the factor matrix should have at least one zero, (b) if there are \( m \) common factors, each column of the factor matrix should have at least \( m \) zeros, (c) for every pair of columns of the factor matrix, there should be several variables whose entries vanish in one column"
but not in the other, (d) for every pair of columns of the factor matrix, a large proportion of the variables should have vanishing entries in both columns when there are four or more factors, and (e) for every pair of columns of the factor matrix there should be only a small number of variables with nonvanishing entries in both columns” (p.98). For the purpose of clarification, it should be noted that what Harman (1967) refers to as “vanishing entries” are loadings of zero.

There are two general types of rotation techniques -- orthogonal and oblique. The difference between these two techniques lies in the degree of correlation permitted among the extracted factors. In this regard, in an orthogonal rotation, the factors are rotated in such a way that factor axes are maintained at 90 degrees -- i.e., each factor is independent from all other factors. In other words, the correlation between factors is arbitrarily determined to be zero. On the other hand, in an oblique rotation, correlation among extracted factors is permitted -- indeed encouraged -- since it is assumed that the original variables were environmentally correlated to some extent. In other words, the oblique rotation technique is probably more representative of reality.

Hair, Anderson, and Tatham (1987) have noted that the decision between an orthogonal and oblique rotation technique should be made on the basis of the specifics of the research problem. In this regard, they have noted that if the objective of the research is to reduce the number of original variables regardless of how meaningful the resulting factors may be, then the appropriate rotation technique would be the orthogonal one, especially if these factors are to be used subsequently in regression or other predictive
analyses. However, if the goal of the research is to obtain several theoretically meaningful factors, then an oblique rotation technique is appropriate since (as explained above) it is more realistic (Hair, Anderson, and Tatham, 1987).

As noted earlier, the objective of the factor analysis conducted for the purpose of addressing this research question was understanding the structure of respondent evaluations of preference measurement methodologies. Therefore, the focus of the factor analysis was on understanding as opposed to prediction. In such instances, it was considered necessary to allow the existence of intercorrelations among the extracted factors on account of the fact that this would be better reflective of reality.

Moreover, since oblique rotation allow for the existence of all possible degrees of correlation among the extracted factors, they also allow for the existence of no correlation among the extracted factors. Viewed in this way, an orthogonal rotation technique is a special case of the oblique rotation techniques; hence, any oblique rotation would also consider orthogonal rotations as one of the several possible solutions before deciding on the solution that best fit the data. In other words, the range of possible solutions is broader with oblique rotation techniques as opposed to orthogonal rotation techniques. Therefore, it was decided to use oblique rotation of the extracted factors for the purpose of this study.

The specific approach to oblique rotation (e.g., promax, oblimin, etc.) is usually a function of what is available in the statistical software. SAS performs oblique rotation using promax; hence, the specific approach to oblique rotation used in this study was promax. However, some analyses
needed to be performed on another factor analysis computer program, i.e., FÇAP (Cudeck, 1991), as explained in the next chapter. This program uses the direct quartimin approach to oblique rotation. Therefore, both promax as well as direct quartimin approaches to oblique rotation were used in this study. As explained in the next chapter, the final rotated factor matrix chosen for interpretation was based on the promax approach.

The result of performing the factor analysis on the basis of the above decisions was a rotated factor loading matrix. The following guidelines were followed in interpreting this matrix: (a) following Hair, Anderson, and Tatham (1987), a factor loading was considered significant at the 0.05 level if its absolute value was at least 0.19, (b) although it was desired to have measured variables load on only one factor, multiple loadings were permitted, and (c) the emphasis in interpretation was on the meaningfulness of the rotated factor loading matrix in terms of dimensions underlying respondent evaluations of preference measurement methodologies -- therefore, more emphasis was placed on the observed pattern of factor loadings as opposed to any mechanical rules of interpretation.

In summary, the specific steps followed in performing the exploratory factor analysis using the common factor analysis model in this study were: (a) the correlation matrix of the measured variables for each of three evaluations was computed using SAS; this matrix was used as the data for the factor analysis, (b) factors were extracted using the method of maximum likelihood estimation; multiple models were fit to the data, each extracting a different number of factors, (c) the fit of each of the models of (b) was evaluated using both, the test for exact fit as well as the test for close fit;
accordingly, a decision on the number of factors to extract was made; (d) the decided number of extracted factors was rotated obliquely using the promax as well as direct quartimin oblique rotation approach to an approximation of simple structure, (e) the final rotated factor loading matrix was interpreted in order to determine the underlying structure of respondent evaluations of preference measurement methodologies. It should be noted that the same procedure was followed for the respondent evaluations of each of the three preference measurement methodologies.

Comparing Respondent Evaluations

Having addressed the psychometric properties of the instruments used to measure respondents' evaluations of each individual preference measurement methodology, the question of significant differences among these evaluations on each of the attributes used to operationalize these evaluations was then addressed. The specific research question in this regard was:

**Research Question # 17:** Are there any significant differences among the respondent evaluations for the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM in terms of the following variables:
(a) interest
(b) sensibility
(c) difficulty in doing
(d) usefulness
(e) length
(f) success
(g) clarity
(h) difficulty in understanding
(i) confidence

This set of research questions was addressed using a series of repeated measures one-way analyses of variance, with the Scheffe test for multiple comparisons. The reasons for using the Scheffe test for multiple comparisons were the same as those mentioned under Research Question # 9; therefore, they will not be repeated here. Further, the issue of sphericity of the covariance matrix of the repeated measures was also discussed in Research Question # 9; it also will not be repeated here. The data used to address this question were the respondent evaluations of the three preference measurement methodologies on the seven-point semantic differential and the five-point Likert-type scale (as described earlier).

The null hypothesis for each of these tests was that there were no significant differences among the respondent evaluations of three health status preference measurement methodologies on the attribute in question, while the alternative hypothesis was that at least one of the preference
measurement methodologies was evaluated differently than the others by respondents.

For the purpose of addressing this research question for this study, the following analytical procedures were followed: (a) a repeated measures one-way analysis of variance was run, (b) the results of the test for sphericity of the covariance matrix of repeated measures were used to determine whether the usual or adjusted F-test should be used for hypothesis testing of the null hypothesis, (c) if the omnibus F-test in (b) above was significant, multiple comparison tests were conducted using the Scheffe test for multiple comparisons in order to control the familywise error.

Research Question #18: Is there a difference between the preferential rank ordering of the three health status preference measurement methodologies (the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) as determined by responses to questions using direct comparison and judgmental ratings as the response modes in terms of the following variables:

(a) interest
(b) sensibility
(c) difficulty in doing
(d) usefulness
(e) length
As noted in Chapter II, the purpose of this question was to investigate the existence of the phenomenon of preference reversal in the context of respondent evaluations of preference measurement methodologies. The data for the judgmental ratings were the same as were evaluated in Research Question # 17; the data for the direct comparisons were collected by the instrument presented in Appendix H.

In order to provide a direct comparative assessment of the three preference measurement methodologies in terms of each of the variables measuring respondent evaluations, simple frequency counts and percentages were computed for each of three places in the rankings of the different attributes. These were then compared to the preferential rank ordering of the three methodologies as inferred from the judgmental ratings of the methodologies on each of the variables. On account of the small sample size for this comparison (only three methodologies), inferential statistics were not computed; only descriptive statistics (i.e., frequencies, percentages, and Kendall’s $\tau$ are reported for this research question).
This chapter has described the methodology followed in order to address each of the research questions which was investigated in this dissertation. The description of the methodology has been presented in some detail, in order to enable replication by other researchers. In this regard, the guiding premise in writing this chapter has been on transparency of procedures. This same premise will dictate the flow of the next chapter of this dissertation, in which the results of each of the analyses conducted in this study are presented in a step-by-step manner to facilitate easy understanding.
CHAPTER VI

RESULTS

Chapter Overview

This chapter presents the results of the analyses conducted in order to address each of the research questions which were investigated in this dissertation. It begins with a presentation of the response analysis, which is followed by a section which describes the sample of respondents in terms of background characteristics such as demographics, severity of hypertension, and presence of comorbidities. This is followed by a presentation of the results of the analysis procedures conducted to address each of the research questions that was investigated in this dissertation. In presenting the results, the organization followed in Chapters II and V is followed -- i.e., three categories of research questions, pertaining to modeling the data, comparing preference scores, and comparing respondent evaluations. The emphasis in this chapter is on the presentation, and not the interpretation, of the results. The results are interpreted and discussed in the final chapter of this dissertation.
Response Analysis

It was noted in Chapter V that the estimated sample size for this study was 130 patients and that it was decided to approach 153 patients to participate in this study (assuming an 85% participation rate). Out of the 153 patients who were contacted over the telephone to participate in the study, only seven refused to do so. A further five did not show up for their interviews at the study booth. There were two major reasons for these no-shows: (a) they did not keep their appointments with their physicians at the VAOPC and therefore did not come into the VAOPC altogether (three such patients) and (b) they came into the VAOPC and kept their appointments with their physicians, but did not show up for their interviews at the study booth (two such patients).

It should be noted that one of the two patients who fell in the latter category of no-shows [i.e., (b)] stopped by the study booth to inform the study investigator that he was running late for an appointment and therefore needed to cancel his interview. The other patient apparently either forgot about the interview or chose not to be interviewed, since a check on the patient check-in schedule at the VAOPC revealed that he checked into the clinic a few minutes before his appointment with his physician and checked out very shortly after his appointment was over. Finally, one patient discontinued the interview, indicating that he was bothered by his allergies. It should be noted that this patient was sneezing frequently during the short period of time he was being interviewed. Therefore, a total of 13 patients who were contacted were unable to be interviewed. That left a total of 140 patients who were interviewed.
Out of these 140 interviews, one was not included for the analyses conducted to address the study questions because it contained markedly different data from the other interviews. Specifically, this particular respondent indicated that he preferred the state of immediate death to all other states of health. He further indicated that he understood the instructions perfectly (and the study investigator concurred with that opinion) and that his responses were not a function of misunderstanding the instructions. He explained his preference on account of his belief that he had lived his life to the fullest and “… was ready to go if need be.” Therefore, the effective sample size for all the analyses reported in this chapter is 139 patients.

It should be noted that this number (i.e., 139) is higher than the estimated sample size calculated in Chapter V. This was just a function of the fact that the vast majority of patients were very willing to take the time to be interviewed, and consequently the refusal rate was very low. In all, 90.85% (i.e., 139 out of 153) patients who were contacted over the telephone gave usable responses for the purpose of addressing the research questions which were investigated in this study. The response analysis is summarized in Table 7.

The data collection period of the study lasted approximately two months, with usually between three and four interviews being conducted every day. The VAOPC was open only between the hours of 8:00 a.m. and 5:00 p.m.; therefore, it was not possible to schedule more interviews per day. However, the time in the evenings was usually used to recruit patients over the telephone for the study.
<table>
<thead>
<tr>
<th>Category</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients contacted by telephone</td>
<td>153</td>
</tr>
<tr>
<td>less Number of refusals over the telephone</td>
<td>- 7</td>
</tr>
<tr>
<td>less Number of no-shows for appointment and interview</td>
<td>- 3</td>
</tr>
<tr>
<td>less Number of no-shows for interview only</td>
<td>- 2</td>
</tr>
<tr>
<td>less Number of patients who discontinued interview</td>
<td>- 1</td>
</tr>
<tr>
<td>less Number of patients whose interviews were not used</td>
<td>- 1</td>
</tr>
<tr>
<td>TOTAL (number of patients who gave useable data)</td>
<td>139</td>
</tr>
</tbody>
</table>
Descriptive Characteristics of Sample

The purpose of this section is to describe the characteristics of the sample of patients who provided usable interviews (n=139) in terms of demographic characteristics, severity of hypertension, and presence of comorbidities. Self-rated health status -- as measured by the EuroQol -- will also be provided. First, however, information will be provided regarding some interview-specific characteristics of the respondents -- the position of the interview (before or after their physician appointment), whether patients were accompanied by their spouses, and the order in which preference measurement methodologies were presented to patients.

Regarding the position of the interview, a greater number of respondents took the interview after their appointments at the VAOPC -- 77 (55.4%) versus 61 (43.9%). This was reflective of their preference to be interviewed while waiting for their prescriptions to be filled. One patient took the interview between his two appointments at the VAOPC, which were scheduled three hours apart from each other (sufficient time to complete the interview in one sitting). Nine of the patients were accompanied by their spouses. Finally, the breakdown of respondents according to the order of presentation of preference measurement methodologies is provided in Table 8.

Background information pertaining to the demographic characteristics, severity of hypertension, and presence of comorbidities was collected from all respondents at the beginning of the interview. Given the nature of the study setting (i.e., VAOPC), it is not surprising that the vast majority (137 out of 139 or 98.6%) of respondents were male.
Table 8: Order of Presentation of Preference Measurement Methodologies

<table>
<thead>
<tr>
<th>First Methodology</th>
<th>Second Methodology</th>
<th>Third Methodology</th>
<th>( n ) (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUF-RS(^a)</td>
<td>CUF-SG(^b)</td>
<td>DCCM(^c)</td>
<td>50 (36.0%)</td>
</tr>
<tr>
<td>CUF-RS</td>
<td>DCCM</td>
<td>CUF-SG</td>
<td>44 (31.6%)</td>
</tr>
<tr>
<td>DCCM</td>
<td>CUF-RS</td>
<td>CUF-SG</td>
<td>45 (32.4%)</td>
</tr>
</tbody>
</table>

Legend:

- \(^a\): CUF-RS = conditional utility function-based procedure with rating scale scaling method
- \(^b\): CUF-SG = conditional utility function-based procedure with standard gamble scaling method
- \(^c\): DCCM = discrete choice conjoint methodology
The demographic characteristics that were collected included respondents' age, highest education level, and main activity. The mean (± standard deviation) age of respondents was 65.79 (± 9.50) years. About two-thirds (65.5%) of the respondents said they had at least completed high school, while about a quarter (25.2%) said they had some college experience. The majority of respondents (114 out of 139 or 82.0%) reported they were retired. The breakdown of respondents by demographic sub-groups (except age, which was measured as a continuous variable) is provided in Table 9.

An indication of the severity of the respondents' hypertension was obtained by self-reports of the length of time since patients were diagnosed as having high blood pressure as well as the most recent systolic and diastolic blood pressure readings, which were obtained from respondents' medical records. Although the majority of patients (82 out of 139 or 59%) had their blood pressure checked on the same day as the interview, there were some patients whose most recent blood pressure reading was over six months before the interview. The mean (± standard deviation) "lag" between the most blood pressure reading and the day of the interview was 26.05 (± 46.35) days.

Most respondents had been diagnosed as having high blood pressure for a fair amount of time -- for instance, about three quarters (73.4%) of respondents had been diagnosed as having high blood pressure for over five years and the majority of respondents (56.8%) had been diagnosed as having high blood pressure for over ten years. A breakdown of respondents by length of time since diagnosis of high blood pressure is provided in Table 10.
Table 9: Demographic Characteristics of Respondents

<table>
<thead>
<tr>
<th>Demographic Variable</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex:</strong></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>137 (98.6%)</td>
</tr>
<tr>
<td>Female</td>
<td>2 (1.4%)</td>
</tr>
<tr>
<td><strong>Highest Education Level:</strong></td>
<td></td>
</tr>
<tr>
<td>Grade School or less</td>
<td>10 (7.2%)</td>
</tr>
<tr>
<td>Some High School</td>
<td>38 (27.3%)</td>
</tr>
<tr>
<td>Completed High School</td>
<td>33 (23.7%)</td>
</tr>
<tr>
<td>Technical School</td>
<td>9 (6.5%)</td>
</tr>
<tr>
<td>Some College</td>
<td>35 (25.2%)</td>
</tr>
<tr>
<td>College Degree</td>
<td>9 (6.5%)</td>
</tr>
<tr>
<td>Some Graduate Work</td>
<td>5 (3.6%)</td>
</tr>
<tr>
<td><strong>Main Activity:</strong></td>
<td></td>
</tr>
<tr>
<td>Retired</td>
<td>114 (82.0%)</td>
</tr>
<tr>
<td>In Employment or Self-employed</td>
<td>11 (7.9%)</td>
</tr>
<tr>
<td>Other</td>
<td>14 (10.1%)</td>
</tr>
<tr>
<td>Severity of Hypertension Variable</td>
<td>n (%)</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>Length of Time Since Diagnosis of High Blood Pressure:</strong></td>
<td></td>
</tr>
<tr>
<td>1 Year or Less</td>
<td>4 (2.9%)</td>
</tr>
<tr>
<td>More than 1 Year but No More than 5 Years</td>
<td>33 (23.7%)</td>
</tr>
<tr>
<td>More than 5 Years but No More than 10 Years</td>
<td>23 (16.5%)</td>
</tr>
<tr>
<td>More than 10 Years but No More than 15 Years</td>
<td>21 (15.1%)</td>
</tr>
<tr>
<td>More than 15 Years but No More than 20 Years</td>
<td>25 (18.0%)</td>
</tr>
<tr>
<td>More than 20 Years but No More than 25 Years</td>
<td>18 (12.9%)</td>
</tr>
<tr>
<td>More than 25 Years but No More than 30 Years</td>
<td>7 (5.0%)</td>
</tr>
<tr>
<td>More than 30 Years</td>
<td>8 (5.8%)</td>
</tr>
<tr>
<td><strong>Classification according to JNC V:</strong></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>51 (36.7%)</td>
</tr>
<tr>
<td>High Normal</td>
<td>20 (14.4%)</td>
</tr>
<tr>
<td>Mild Hypertension</td>
<td>44 (31.7%)</td>
</tr>
<tr>
<td>Moderate Hypertension</td>
<td>21 (15.5%)</td>
</tr>
<tr>
<td>Severe Hypertension</td>
<td>3 (2.2%)</td>
</tr>
</tbody>
</table>

Legend: a = description of classes is provided in Appendix 18
The mean (± standard deviation) systolic blood pressure of respondents was 139.16 (± 31.37) mm Hg, while the mean (± standard deviation) diastolic blood pressure of respondents was 78.66 (± 10.60) mm Hg. As noted in Chapter V, respondents were classified according to the categories of the Fifth Report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure or JNC V (1993) for severity of hypertension. According to this particular classification, the hypertension of most patients participating in this study appeared to be well-controlled, since the majority (51.1%) of patients did not even cross the cut-off figures to be diagnosed as hypertensive. A breakdown of patients according to severity level as determined by the JNC V classification is provided in Table 10.

Apart from demographic characteristics and severity of hypertension, background information was also collected on presence of comorbidities. For this purpose, data were only collected at a nominal level of measurement, i.e., yes/no. A listing of the self-reported comorbidities of patients is provided in Table 11. The pattern of comorbidities reported by patients is consistent with what might be expected for a sample of veterans, with arthritis or rheumatism, back pain, diabetes, heart trouble, and hyperlipidemia being the most frequently mentioned comorbidities, while vision, hearing, and prostate trouble were also quite frequently mentioned by respondents.

Respondents were also asked to describe their health status on the day of the interview, as operationalized by the EuroQol health status measurement instrument. Most respondents classified themselves on the
highest levels of functioning on most of the health status dimensions, with
the exception of the pain/discomfort dimension for which most respondents
classified themselves on the intermediate level of functioning. For each of
the five health status dimensions, very few respondents (i.e., less than 10%)
classified themselves on the lowest level of functioning -- once again, the
greatest number of respondents classified themselves on the lowest level of
functioning for the dimension of pain/discomfort among all the health
status dimensions of the EuroQol. Regarding the general health perceptions
of the respondents, the majority of respondents indicated that their level of
health was "much the same" over the past 12 months. These results are
presented in Table 12.

Finally, the total time for the entire interview was a little less than an
hour -- the mean (± standard deviation) time in minutes for the interview
was 59.02 (12.23) minutes. Although the time for the entire interview was
rather long, it should be noted that patients did not seem to mind this fact.
Indeed, several patients informally mentioned that they enjoyed spending
the time being interviewed for this study. It is also interesting to note that
the total time for the study interviews was less than that of the interviews
conducted for the pre-testing and pilot testing procedures (which were
reported in the previous chapter). This is probably reflective of learning
effects on the part of the interviewer (who conducted all three sets of
interviews).
Table 11: Respondent Comorbidities

<table>
<thead>
<tr>
<th>Comorbidity</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allergies</td>
<td>18 (12.9%)</td>
</tr>
<tr>
<td>Arthritis or Rheumatism</td>
<td>81 (58.3%)</td>
</tr>
<tr>
<td>Back Pain</td>
<td>44 (31.7%)</td>
</tr>
<tr>
<td>Blood Disorders</td>
<td>20 (14.4%)</td>
</tr>
<tr>
<td>Cancer or Leukemia</td>
<td>14 (10.1%)</td>
</tr>
<tr>
<td>Depression</td>
<td>20 (14.4%)</td>
</tr>
<tr>
<td>Diabetes</td>
<td>41 (29.5%)</td>
</tr>
<tr>
<td>Eye, Vision, or Sight Problems</td>
<td>38 (27.3%)</td>
</tr>
<tr>
<td>Hearing Problems</td>
<td>33 (23.7%)</td>
</tr>
<tr>
<td>Heart Trouble</td>
<td>54 (38.8%)</td>
</tr>
<tr>
<td>High Cholesterol/High Blood Fats</td>
<td>44 (31.7%)</td>
</tr>
<tr>
<td>Lung Disease (e.g., asthma, bronchitis, emphysema)</td>
<td>17 (12.2%)</td>
</tr>
<tr>
<td>Prostate Trouble</td>
<td>31 (22.3%)</td>
</tr>
<tr>
<td>Sleep Disorders</td>
<td>24 (17.3%)</td>
</tr>
<tr>
<td>Ulcers of Stomach and/or Intestine</td>
<td>26 (18.7%)</td>
</tr>
</tbody>
</table>
Table 12: Health Status of Respondents as Operationalized by the EuroQol

<table>
<thead>
<tr>
<th>Health Status Dimension</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mobility:</strong></td>
<td></td>
</tr>
<tr>
<td>no problems in walking about</td>
<td>70 (50.4%)</td>
</tr>
<tr>
<td>some problems in walking about</td>
<td>69 (49.6%)</td>
</tr>
<tr>
<td>confined to bed</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td><strong>Self-care:</strong></td>
<td></td>
</tr>
<tr>
<td>no problems with self-care</td>
<td>130 (93.5%)</td>
</tr>
<tr>
<td>some problems with washing or dressing</td>
<td>9 (6.5%)</td>
</tr>
<tr>
<td>unable to wash or dress</td>
<td>0 (0.0%)</td>
</tr>
<tr>
<td><strong>Usual Activities:</strong></td>
<td></td>
</tr>
<tr>
<td>no problems with performing usual activities</td>
<td>75 (54.0%)</td>
</tr>
<tr>
<td>some problems with performing usual activities</td>
<td>57 (41.0%)</td>
</tr>
<tr>
<td>unable to perform usual activities</td>
<td>7 (5.0%)</td>
</tr>
<tr>
<td><strong>Pain/Discomfort:</strong></td>
<td></td>
</tr>
<tr>
<td>no pain or discomfort</td>
<td>57 (41.0%)</td>
</tr>
<tr>
<td>moderate pain or discomfort</td>
<td>69 (49.6%)</td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>13 (9.4%)</td>
</tr>
<tr>
<td><strong>Anxiety/Depression:</strong></td>
<td></td>
</tr>
<tr>
<td>not anxious or depressed</td>
<td>91 (65.5%)</td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td>45 (32.4%)</td>
</tr>
<tr>
<td>extremely anxious or depressed</td>
<td>3 (2.2%)</td>
</tr>
<tr>
<td><strong>Compared with general level of health over past 12 months,</strong></td>
<td></td>
</tr>
<tr>
<td>health on the day of the interview:</td>
<td></td>
</tr>
<tr>
<td>better</td>
<td>35 (25.2%)</td>
</tr>
<tr>
<td>much the same</td>
<td>76 (54.7%)</td>
</tr>
<tr>
<td>worse</td>
<td>28 (20.1%)</td>
</tr>
</tbody>
</table>
Results of the Modeling Procedures

The specific modeling procedures for each of the three preference measurement methodologies implemented in this study were described in Chapter V. This section presents the results of these modeling procedures. These results are presented in turn, beginning with the conditional utility function-based procedure implemented with the rating scale scaling method (CUF-RS), followed by the conditional utility function-based procedure implemented with the standard gamble scaling method (CUF-SG), and finally by the discrete choice conjoint methodology (DCCM).

(a) Modeling the CUF-RS Data

The research questions which were investigated with respect to modeling the CUF-RS data were:

Research Question # 1: What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?

Research Question # 2: What is the predictive ability of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology as a measure of health status preferences, as indicated by the mean absolute residual?
The single attribute values determined using the CUF-RS methodology are presented in Table 13. As explained in the previous chapter, the first step in modeling the CUF-RS data was converting the single attribute values (shown in Table 13) into their respective and mathematically equivalent disvalue counterparts by taking their complements. These single attribute disvalues are presented in Table 14. The next step in the modeling procedure is to determine the appropriate model form of the preference function (in this case a value function). As noted earlier, this was done using the corner state approach, wherein the values of the single attribute corner states (using the disvalue scale) were used as the individual attribute scaling constants or $k_i$. The measured values of the multiattribute health states are presented in Table 15.

A few points need to be noted about the anchor states in Table 15. It was noted in Chapter V that respondents were given the three anchor states (i.e., the all-best, all-worst, and immediate death states) together and asked to place the all-best state at the top of the scale at the "best imaginable health state" or 100-mark. However, the option of selecting the bottom anchor of the scale was left to each respondent -- i.e., they could anchor the bottom of the scale with either the all-worst state or the immediate death state.

Although there were some respondents who opted to anchor the scale at the bottom with the all-worst state instead of the immediate death state (n=20), it was decided to use the immediate death state as the bottom anchor for this study. The reasons for this selection of the bottom anchor were discussed in Chapter V and will not be repeated here.
Table 13: Single Attribute Values Determined using CUF-RS

<table>
<thead>
<tr>
<th>Attribute Levels</th>
<th>Mean Value</th>
<th>SD'</th>
<th>SEM'</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mobility:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no problems in walking about</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>some problems in walking about</td>
<td>0.58</td>
<td>0.18</td>
<td>0.015</td>
</tr>
<tr>
<td>confined to bed</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Self-care:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no problems with self-care</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>some problems with washing or dressing</td>
<td>0.58</td>
<td>0.18</td>
<td>0.015</td>
</tr>
<tr>
<td>unable to wash or dress</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Usual Activities:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no problems with performing usual activities</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>some problems with performing usual activities</td>
<td>0.59</td>
<td>0.18</td>
<td>0.015</td>
</tr>
<tr>
<td>unable to perform usual activities</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Pain/Discomfort:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no pain or discomfort</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>moderate pain or discomfort</td>
<td>0.59</td>
<td>0.17</td>
<td>0.014</td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Anxiety/Depression:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not anxious or depressed</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td>0.58</td>
<td>0.20</td>
<td>0.017</td>
</tr>
<tr>
<td>extremely anxious or depressed</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: where 1.00 = best imaginable health state and 0.00 = worst imaginable health state
c: SD = standard deviation
d: SEM = standard error of the mean
Table 14: Single Attribute Disvalues Determined using CUF-RS<sup>a</sup>

<table>
<thead>
<tr>
<th>Attribute Levels</th>
<th>Mean Disvalue&lt;sup&gt;bc&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobility:</td>
<td></td>
</tr>
<tr>
<td>no problems in walking about</td>
<td>0.00</td>
</tr>
<tr>
<td>some problems in walking about</td>
<td>0.42</td>
</tr>
<tr>
<td>confined to bed</td>
<td>1.00</td>
</tr>
<tr>
<td>Self-care:</td>
<td></td>
</tr>
<tr>
<td>no problems with self-care</td>
<td>0.00</td>
</tr>
<tr>
<td>some problems with washing or dressing</td>
<td>0.42</td>
</tr>
<tr>
<td>unable to wash or dress</td>
<td>1.00</td>
</tr>
<tr>
<td>Usual Activities:</td>
<td></td>
</tr>
<tr>
<td>no problems with performing usual activities</td>
<td>0.00</td>
</tr>
<tr>
<td>some problems with performing usual activities</td>
<td>0.41</td>
</tr>
<tr>
<td>unable to perform usual activities</td>
<td>1.00</td>
</tr>
<tr>
<td>Pain/Discomfort:</td>
<td></td>
</tr>
<tr>
<td>no pain or discomfort</td>
<td>0.00</td>
</tr>
<tr>
<td>moderate pain or discomfort</td>
<td>0.41</td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>1.00</td>
</tr>
<tr>
<td>Anxiety/Depression:</td>
<td></td>
</tr>
<tr>
<td>not anxious or depressed</td>
<td>0.00</td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td>0.42</td>
</tr>
<tr>
<td>extremely anxious or depressed</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: mean disvalue = 1 - mean value (taken from Table 13)
c: where 1.00 = best imaginable health state and
    0.00 = worst imaginable health state
Table 15: Multiattribute Health State Values Determined using CUF-RS\textsuperscript{a}

<table>
<thead>
<tr>
<th>Multiattribute Health State\textsuperscript{b}</th>
<th>Mean Value\textsuperscript{c}</th>
<th>SD\textsuperscript{d}</th>
<th>SEM\textsuperscript{e}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anchor States:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-best (AB)</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>All-worst (AW)</td>
<td>0.06</td>
<td>0.07</td>
<td>0.006</td>
</tr>
<tr>
<td>Immediate Death (ID)</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Corner States:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility (CM)</td>
<td>0.40</td>
<td>0.17</td>
<td>0.014</td>
</tr>
<tr>
<td>Self-care (CS)</td>
<td>0.65</td>
<td>0.17</td>
<td>0.014</td>
</tr>
<tr>
<td>Usual Activities (CU)</td>
<td>0.68</td>
<td>0.19</td>
<td>0.016</td>
</tr>
<tr>
<td>Pain/Discomfort (CP)</td>
<td>0.56</td>
<td>0.22</td>
<td>0.019</td>
</tr>
<tr>
<td>Anxiety/Depression (CA)</td>
<td>0.63</td>
<td>0.24</td>
<td>0.020</td>
</tr>
<tr>
<td><strong>Interior States:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interior State #1 (BO)</td>
<td>0.32</td>
<td>0.16</td>
<td>0.014</td>
</tr>
<tr>
<td>Interior State #2 (E3)</td>
<td>0.37</td>
<td>0.18</td>
<td>0.015</td>
</tr>
<tr>
<td>Interior State #3 (E8)</td>
<td>0.16</td>
<td>0.11</td>
<td>0.009</td>
</tr>
</tbody>
</table>

Legend:
\textsuperscript{a}: n = 139
\textsuperscript{b}: the letters in parentheses correspond to the labels used to identify the multiattribute health states in the visual aids (see Appendix B)
\textsuperscript{c}: where 1.00 = best imaginable health state and 0.00 = worst imaginable health state
\textsuperscript{d}: SD = standard deviation
\textsuperscript{e}: SEM = standard error of the mean
In order to convert the values assigned to multiattribute health states by these 20 respondents to the immediate death to all-best scale, the following conversion formula was used:

\[ v(x_i) = \frac{(x_i - \text{ID})}{(1 - \text{ID})} \]  (6.1)

where

- \( x_{\text{in}} \) = value assigned to multiattribute health state on the all-worst to all-best scale
- \( \text{ID} \) = value assigned to immediate death on the all-worst to all-best scale
- \( v(x_i) \) = final transformed value of multiattribute health state on the immediate death to all-best scale

For each of these cases, the value of the all-worst state was recorded as zero. It should be noted that since the all-worst state was actually given a value of less than immediate death, its actual value was less than zero. However, in keeping with the decision to not measure preferences for health states which were considered worse than death, no attempt was made to infer this actual value. Along the same lines, values of any multiattribute health states which were initially (i.e., on the all-worst to all-best scale) considered worse than immediate death were recorded as zero.

This handling of health states considered worse than immediate death in no way denigrates the importance of recognizing patients' preferences for health states relative to the state of immediate death. Indeed, the existence of states worse than immediate death has been supported by empirical
evidence for over a decade (Bjork and Althin, 1992; Kind and Rosser, 1979; Torrance, 1984; Torrance, Boyle, and Horwood, 1982; Torrance, Zhang, Feeny, et al., 1992). The implications of the existence of such health status preferences will be discussed in Chapter VII.

The next step in the modeling procedure was to convert the corner state values (as they were measured) into disvalues -- this is because preferences for these corner states were measured on a 0-100 scale, where 0 = worst imaginable health state and 100 = best imaginable health state, i.e., on a value scale. The conversion to disvalues simply required taking the complement of the measured mean values, so that the resultant disvalue 0-100 scale would be anchored by 0 = best imaginable health state and 100 = worst imaginable health state. Table 16 presents the preference scores for the multiattribute health states on this disvalue scale.

As explained earlier, these disvalue scores for the corner states are nothing but the individual attribute scaling constants, which are needed to determine the particular model form which is appropriate for the preference function. The sum of these corner state disvalues or individual attribute scaling constants \( \sum k \) is 2.08. Since this sum is not equal to one, the additive model form is rejected for this preference function. Since the multilinear model form was not considered in this study (as explained earlier), the multiplicative model form was considered appropriate for this preference function. The interaction parameter \( k \) for the multiplicative model form was determined using the computer package SOLVE (Mundt, 1989), as explained in the previous chapter. The resulting value of the interaction parameter \( k \) was -0.914.
### Table 16: Multiattribute Health State Disvalues Determined using CUF-RS$^2$

<table>
<thead>
<tr>
<th>Multiattribute Health State$^b$</th>
<th>Mean Disvalue$^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anchor States:</strong></td>
<td></td>
</tr>
<tr>
<td>All-best (AB)</td>
<td>0.00</td>
</tr>
<tr>
<td>All-worst (AW)</td>
<td>0.94</td>
</tr>
<tr>
<td>Immediate Death (ID)</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Corner States:</strong></td>
<td></td>
</tr>
<tr>
<td>Mobility (CM)</td>
<td>0.60</td>
</tr>
<tr>
<td>Self-care (CS)</td>
<td>0.35</td>
</tr>
<tr>
<td>Usual Activities (CU)</td>
<td>0.32</td>
</tr>
<tr>
<td>Pain/Discomfort (CP)</td>
<td>0.44</td>
</tr>
<tr>
<td>Anxiety/Depression (CA)</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>Interior States:</strong></td>
<td></td>
</tr>
<tr>
<td>Interior State #1 (BO)</td>
<td>0.68</td>
</tr>
<tr>
<td>Interior State #2 (E3)</td>
<td>0.63</td>
</tr>
<tr>
<td>Interior State #3 (E8)</td>
<td>0.84</td>
</tr>
</tbody>
</table>

**Legend:**

- $a$: n = 139
- $b$: the letters in parentheses correspond to the labels used to identify the multiattribute health states in the visual aids (see Appendix B)
- $c$: where 0.00 = best imaginable health state and 1.00 = worst imaginable health state
Disvalues for multiattribute health states were then determined using the formula for the multiplicative model form:

\[ p(x) = \left(\frac{1}{k}\right) \left\{ \Pi [1 + kk_p(x)] - 1 \right\}; \quad (6.2) \]

where

\( i = 1 \ldots 5 \) (i.e., five health status dimensions in the EuroQol)

\( x \in X \) = multiattribute alternative

\( k = \) interaction parameter (-1 < k < 0 or k > 0)

\( 1 + k = \Pi [1 + kk_i] \)

\( k_i = \) scaling constant (0 ≤ k ≤ 1)

\( x_i = \) level of x on the i-th attribute

\( p_i(x) = \) i-th single attribute preference function

\( p(x) = \) overall preference function for multiattribute alternative x

Specifically, disvalues were calculated for the three multiattribute interior states (i.e., BO, E3, and E8). These calculated disvalues were then compared to the measured disvalues of these same three multiattribute health states in order to estimate the predictive ability of the CUF-RS methodology. Table 17 presents an illustration of the calculation for one of the interior states (i.e., BO), while Table 18 presents the calculated and measured disvalues of these three multiattribute health states, along with the each of the residuals.
In summary, the modeling of the data obtained using the CUF-RS methodology supported the existence of a multiplicative model form for the preference function for health status as operationalized by the EuroQol. The average error in prediction of health status preferences using the CUF-RS methodology is about 5%. Issues relating to these particular results of modeling of the data using the CUF-RS methodology will be discussed in Chapter VII.

(b) Modeling the CUF-SG Data

The research questions pertaining to modeling the CUF-SG data were:

**Research Question # 3**: What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

**Research Question # 4**: What is the predictive ability of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology as a measure of health status preferences, as indicated by the mean absolute residual?
Table 17: Illustration of Calculation of Multiatribute Health State Disvalues

Consider the multiatribute health state BO:*

- confined to bed (MOB-3)
- some problems with washing or dressing (SFC-2)
- no problems with performing usual activities (ACT-1)
- no pain or discomfort (PDF-1)
- not anxious or depressed (ANX-1)

The calculated disvalue is obtained using Equation (6.2):

\[ p(x) = \frac{1}{k} \prod (1 + k_i p_i(x_i) - 1) \]

where all terms have been previously defined.

From Tables 14 and 16,

\[ p(x_{MOB}) = 1; \quad p(x_{SFC}) = 0.42; \quad p(x_{ACT}) = 0; \quad p(x_{PDF}) = 0; \quad p(x_{ANX}) = 0 \]

\[ k_{MOB} = 0.60; \quad k_{SFC} = 0.35; \quad k_{ACT} = 0.32; \quad k_{PDF} = 0.44; \quad k_{ANX} = 0.37 \]

Further,

\[ k = -0.914 \]

Inputting these scores into Equation (6.2),

\[ p(BO) = \frac{1}{-0.914} \left\{ (1 - 0.914 \times 0.60 \times 1)(1 - 0.914 \times 0.35 \times 0.42)(1 - 0.914 \times 0.32 \times 0) \right\} \]

\[ = \frac{1}{-0.914} \left\{ 0.609095 \right\} \]

\[ p(BO) = 0.664. \]

Therefore, the calculated disvalue of the multiatribute state BO = 0.664.

Legend:

- the label BO was used for identification purposes in the visual aids (see Appendix B)
Table 18: Predictive Ability of CUF-RS Methodology*

<table>
<thead>
<tr>
<th>Multiattribute Health State(^b)</th>
<th>Calculated Disvalue (C)(^c)</th>
<th>Measured Disvalue (M)(^d)</th>
<th>Residual (C-M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BO</td>
<td>0.664</td>
<td>0.678</td>
<td>-0.014</td>
</tr>
<tr>
<td>E3</td>
<td>0.714</td>
<td>0.635</td>
<td>0.079</td>
</tr>
<tr>
<td>E8</td>
<td>0.896</td>
<td>0.837</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Mean Absolute Residual

Standard Deviation of Absolute Residual 0.051

Legend:

a: \(n = 139\)

b: the labels BO, E3, and E8 were used to identify the multiattribute health states in the visual aids (see Appendix B)

c: calculated using equation (6.2)

d: as reported in Table 16
The single attribute utilities determined using the CUF-SG methodology are presented in Table 19. The first step in modeling the CUF-SG data was converting the single attribute utilities (shown in Table 19) into their respective and mathematically equivalent disutility counterparts by taking their complements. These single attribute disutilities are presented in Table 20.

The next step in this modeling procedure is to determine the appropriate model form of the preference function (in this case a utility function). This was done using the corner state approach, wherein the utilities of the single attribute corner states (using the disutility scale) were used as the individual attribute scaling constants or \( k_r \). The measured utilities of the multiattribute health states are presented in Table 21.

A few points need to be noted about the anchor states in Table 21. It was noted in Chapter V that it was decided to scale the utilities on an immediate death to all-best health state continuum, and not estimate utilities for health states considered worse than immediate death. The reasons for this decision were explained earlier, and therefore will not be repeated here. However, the implications of the existence of such health status preferences will be discussed in Chapter VII.

The next step in the modeling procedure was to convert the corner state utilities (as they were measured) into disutilities, as explained earlier. The conversion to disutilities simply required taking the complement of the measured mean utilities. Table 22 presents the preference scores for the multiattribute health states on this disutility scale.
Table 19: Single Attribute Utilities Determined using CUF-SG
d

<table>
<thead>
<tr>
<th>Attribute Levels</th>
<th>Mean Utility</th>
<th>SD</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mobility:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no problems in walking about</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>some problems in walking about</td>
<td>0.81</td>
<td>0.16</td>
<td>0.014</td>
</tr>
<tr>
<td>confined to bed</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Self-care:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no problems with self-care</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>some problems with washing or</td>
<td>0.81</td>
<td>0.16</td>
<td>0.014</td>
</tr>
<tr>
<td>dressing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unable to wash or dress</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Usual Activities:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no problems with performing</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>usual activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>some problems with performing</td>
<td>0.82</td>
<td>0.15</td>
<td>0.013</td>
</tr>
<tr>
<td>usual activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>unable to perform usual</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>activities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Pain/Discomfort:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no pain or discomfort</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>moderate pain or discomfort</td>
<td>0.82</td>
<td>0.16</td>
<td>0.014</td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td><strong>Anxiety/Depression:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>not anxious or depressed</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td>0.81</td>
<td>0.17</td>
<td>0.014</td>
</tr>
<tr>
<td>extremely anxious or depressed</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: where 1.00 = best imaginable health state and 0.00 = worst imaginable health state
c: SD = standard deviation
d: SEM = standard error of the mean
Table 20: Single Attribute Disutilities Determined using CUF-SG

<table>
<thead>
<tr>
<th>Attribute Levels</th>
<th>Mean Disutility&lt;sup&gt;bc&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mobility:</strong></td>
<td></td>
</tr>
<tr>
<td>no problems in walking about</td>
<td>0.00</td>
</tr>
<tr>
<td>some problems in walking about</td>
<td>0.19</td>
</tr>
<tr>
<td>confined to bed</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Self-care:</strong></td>
<td></td>
</tr>
<tr>
<td>no problems with self-care</td>
<td>0.00</td>
</tr>
<tr>
<td>some problems with washing or dressing</td>
<td>0.19</td>
</tr>
<tr>
<td>unable to wash or dress</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Usual Activities:</strong></td>
<td></td>
</tr>
<tr>
<td>no problems with performing usual activities</td>
<td>0.00</td>
</tr>
<tr>
<td>some problems with performing usual activities</td>
<td>0.18</td>
</tr>
<tr>
<td>unable to perform usual activities</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Pain/Discomfort:</strong></td>
<td></td>
</tr>
<tr>
<td>no pain or discomfort</td>
<td>0.00</td>
</tr>
<tr>
<td>moderate pain or discomfort</td>
<td>0.18</td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Anxiety/Depression:</strong></td>
<td></td>
</tr>
<tr>
<td>not anxious or depressed</td>
<td>0.00</td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td>0.19</td>
</tr>
<tr>
<td>extremely anxious or depressed</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: mean disutility = 1 - mean utility (taken from Table 19)
c: where 1.00 = best imaginable health state and 0.00 = worst imaginable health state
Table 21: Multiattribute Health State Utilities Determined using CUF-SG^a

<table>
<thead>
<tr>
<th>Multiattribute Health State^b</th>
<th>Mean Utility^c</th>
<th>SD^d</th>
<th>SEM^e</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor States:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-best (AB)</td>
<td>1.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Immediate Death (ID)</td>
<td>0.00</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Corner States:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility (CM)</td>
<td>0.66</td>
<td>0.18</td>
<td>0.015</td>
</tr>
<tr>
<td>Self-care (CS)</td>
<td>0.84</td>
<td>0.14</td>
<td>0.012</td>
</tr>
<tr>
<td>Usual Activities (CU)</td>
<td>0.85</td>
<td>0.14</td>
<td>0.012</td>
</tr>
<tr>
<td>Pain/Discomfort (CP)</td>
<td>0.77</td>
<td>0.21</td>
<td>0.018</td>
</tr>
<tr>
<td>Anxiety/Depression (CA)</td>
<td>0.83</td>
<td>0.17</td>
<td>0.014</td>
</tr>
<tr>
<td>Interior States:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interior State #1 (BO)</td>
<td>0.63</td>
<td>0.19</td>
<td>0.016</td>
</tr>
<tr>
<td>Interior State #2 (E3)</td>
<td>0.64</td>
<td>0.23</td>
<td>0.020</td>
</tr>
<tr>
<td>Interior State #3 (E8)</td>
<td>0.45</td>
<td>0.25</td>
<td>0.021</td>
</tr>
<tr>
<td>All-worst (AW)</td>
<td>0.29</td>
<td>0.26</td>
<td>0.022</td>
</tr>
</tbody>
</table>

Legend:
- a: n = 139
- b: the letters in parentheses correspond to the labels used to identify the multiattribute health states in the visual aids (see Appendix B)
- c: where 1.00 = best imaginable health state and 0.00 = worst imaginable health state
- d: SD = standard deviation
- e: SEM = standard error of the mean
Table 22: Multiattribute Health State Disutilities Determined using CUF-SG

<table>
<thead>
<tr>
<th>Multiattribute Health State&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Mean Disutility&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anchor States:</td>
<td></td>
</tr>
<tr>
<td>All-best (AB)</td>
<td>0.00</td>
</tr>
<tr>
<td>Immediate Death (ID)</td>
<td>1.00</td>
</tr>
<tr>
<td>Corner States:</td>
<td></td>
</tr>
<tr>
<td>Mobility (CM)</td>
<td>0.34</td>
</tr>
<tr>
<td>Self-care (CS)</td>
<td>0.16</td>
</tr>
<tr>
<td>Usual Activities (CU)</td>
<td>0.15</td>
</tr>
<tr>
<td>Pain/Discomfort (CP)</td>
<td>0.23</td>
</tr>
<tr>
<td>Anxiety/Depression (CA)</td>
<td>0.17</td>
</tr>
<tr>
<td>Interior States:</td>
<td></td>
</tr>
<tr>
<td>Interior State #1 (BO)</td>
<td>0.37</td>
</tr>
<tr>
<td>Interior State #2 (E3)</td>
<td>0.36</td>
</tr>
<tr>
<td>Interior State #3 (E8)</td>
<td>0.55</td>
</tr>
<tr>
<td>All-worst (AW)</td>
<td>0.71</td>
</tr>
</tbody>
</table>

Legend:
a: n = 139
b: the letters in parentheses correspond to the labels used to identify the multiattribute health states in the visual aids (see Appendix B)
c: where 0.00 = best imaginable health state and 1.00 = worst imaginable health state
These disutility scores for the corner states are nothing but the individual attribute scaling constants, which are needed to determine the particular model form which is appropriate for the preference function. The sum of these corner state disvalues or individual attribute scaling constants (i.e., $\sum k$) is 1.05. Strictly speaking [i.e., according to the strict interpretation of the conditional utility function-based procedure as forwarded by Keeney and Raiffa (1976)], since this sum is not equal to one, the additive model form should be rejected for this preference function. However, one of the drawbacks of the conditional utility function-based procedure (which was discussed in Chapter III) is the fact that it lacks an error theory. In other words, one does not know how wide an interval about the sum of 1 should an additive model still be considered as tenable for the preference function.

It seems reasonable to allow for an interval about the sum of 1 of the scaling constants (i.e., $\sum k$) in order to account for random error. Although there are no accepted guidelines for the width of this interval, Krzysztofowicz and Duckstein (1980) have suggested an arbitrary cutoff of 0.1 units about the sum of 1 -- i.e., $1 \pm 0.1$. Using this cutoff for the purpose of determining the model form of the utility function in this study, the additive model form might be considered appropriate.

Still, as noted above, if one strictly follows the Keeney-Raiffa (1976) modeling procedure, one would not allow for any interval about the sum of 1; therefore, one would consider the multiplicative model form appropriate for the utility function in this study. For the purpose of modeling the data in this study, both model forms (i.e., additive as well as multiplicative) were fit to the CUF-SG data, and a decision about the appropriate model form was
deferred until after examination of the relative predictive ability of the model forms. For the purpose of the multiplicative model form, the interaction parameter $k$ was determined using the computer package SOLVE (Mundt, 1989), as explained in the previous chapter. The resulting value of the interaction parameter $k$ was -0.1196.

Disutilities for multiattribute health states were then determined using the formulae for the additive and multiplicative model forms. The additive model form is given by:

$$ p(x) = \sum k_i p_i(x), $$

(6.3)

while the multiplicative model form is given by:

$$ p(x) = \frac{1}{k} \left[ \prod \left[ 1 + k k_i p_i(x) \right] - 1 \right]; $$

(6.4)

where $i = 1 \ldots 5$ (i.e., five health status dimensions in the EuroQol)

$x \in X$ = multiattribute alternative

$k = $ interaction parameter ($-1 < k < 0$ or $k > 0$)

$1 + k = \prod \left[ 1 + k k_i \right]$

$k_i = $ scaling constant ($0 \leq k_i \leq 1$)

$x_i = $ level of $x$ on the $i$-th attribute

$p_i(x) = $ $i$-th single attribute preference function

$p(x) = $ overall preference function for multiattribute alternative $x$
Specifically, disutilities were calculated for the three multiattribute interior states (i.e., BO, E3, and E8) using both the additive as well as multiplicative model forms. These calculated disutilities were compared to the measured disutilities of these same three multiattribute health states to estimate the predictive ability of the CUF-RS methodology. This comparison also helped in determining which of the two alternative model forms was better supported by the empirical data. Table 23 presents an illustration of the calculation for one of the interior states (i.e., BO), while Table 24 presents the calculated and measured disutilities of these three multiattribute health states using both model forms, along with the each of the residuals.

On the basis of the results of the predictive ability of the alternative model forms, one might tentatively support the multiplicative model form as appropriate for the utility function of the respondents interviewed for this study. This is because of the smaller mean absolute residual and smaller standard deviation associated with the multiplicative model form as opposed to the additive model form. However, on account of the extremely small sample size of the holdout sample (n=3), caution needs to be exercised in interpreting this particular result.

In summary, the modeling of the data obtained using the CUF-SG methodology supported the existence of a multiplicative model form for the EuroQol health status preference function. The average error in prediction of health status preferences using this methodology is a little less than 5%. Issues pertinent to these results of modeling of the data using the CUF-SG methodology will be discussed in Chapter VII.
Consider the multiattribute health state \( BO \):
- confined to bed (\( MOB-3 \))
- some problems with washing or dressing (\( SFC-2 \))
- no problems with performing usual activities (\( ACT-1 \))
- no pain or discomfort (\( PDF-1 \))
- not anxious or depressed (\( ANX-1 \))

The calculated disutility is obtained using Equations (6.3) and (6.4):
\[
p(x) = \sum k_i p_i(x_i),
\]
and
\[
p(x) = \frac{1}{k} \prod \left[ 1 + k_i p_i(x_i) \right] - 1;
\]
where all terms have been previously defined.

From Tables 20 and 22,
\[
p(x_{MOB}) = 1; \ p(x_{SFC}) = 0.19; \ p(x_{ACT}) = 0; \ p(x_{PDF}) = 0; \ p(x_{ANX}) = 0
\]
\[
k_{MOB} = 0.34; \ k_{SFC} = 0.16; \ k_{ACT} = 0.15; \ k_{PDF} = 0.23; \ k_{ANX} = 0.17
\]

Further,
\[
k = -0.1196
\]

Inputting these scores into Equation (6.3),
\[
p(BO) = (0.34*1) + (0.16*0.19) + (0.15*0) + (0.23*0) + (0.17*0)
\]
\[
p(BO) = 0.3704.
\]

Inputting these scores into Equation (6.4),
\[
p(BO) = \frac{1}{1 - 0.1196} \left[ \prod \left[ 1 - k_i p_i(x_i) \right] \right] - 1
\]
\[
= \frac{1}{1 - 0.1196} \left[ 0.0442 \right] - 1
\]
\[
p(BO) = 0.3696.
\]

Therefore, the calculated disvalue of the multiattribute state \( BO \):
- with the additive model form = 0.3704.
- with the multiplicative model form = 0.3696.

Legend:
- the label \( BO \) was used for identification purposes in the visual aids
  (see Appendix B)
Table 24: Predictive Ability of CUF-SG Methodology

<table>
<thead>
<tr>
<th>Multiattribute Health State</th>
<th>Calculated Disvalue (C)</th>
<th>Measured Disvalue (M)</th>
<th>Residual (C-M)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Add</td>
<td>Mult</td>
<td>Add</td>
</tr>
<tr>
<td>BO</td>
<td>0.370</td>
<td>0.370</td>
<td>0.372</td>
</tr>
<tr>
<td>E3</td>
<td>0.354</td>
<td>0.350</td>
<td>0.357</td>
</tr>
<tr>
<td>E8</td>
<td>0.691</td>
<td>0.673</td>
<td>0.547</td>
</tr>
</tbody>
</table>

Mean Absolute Residual 0.050 0.045

Standard Deviation of Absolute Residual 0.082 0.071

Legend:

a: n = 139
b: the labels BO, E3, and E8 were used to identify the multiattribute health states in the visual aids (see Appendix B)
c: calculated using Equation (6.3)
d: calculated using Equation (6.4)
e: as reported in Table 22
(c) Modeling the DCCM Data

The research questions which were investigated with respect to modeling the CUF-RS data were:

Research Question # 5: What is the nature of the model form of the health status preference function as operationalized by the EuroQol and determined by the discrete choice conjoint methodology or DCCM?

Research Question # 6: What is the predictive ability of the discrete choice conjoint methodology or DCCM as a measure of health status preferences, as indicated by the mean absolute residual?

As noted in the previous chapter, the DCCM data were analyzed using maximum likelihood estimation in order to determine the parameters of the MNL choice model which was assumed to represent respondents' decision making processes. The model form of respondents' preference functions was empirically determined using a stepwise multinomial logit modeling procedure, the results of which are presented below.

The first task in this procedure was to fit a series of univariate models to the data, in order to determine the need for including each individual attribute in the model. Five such univariate models were fit, corresponding to the five health status dimensions of the EuroQol. Each of these five models was evaluated for significance against the null model, using the
likelihood ratio statistic \((G)\) as the test statistic (as explained in the previous chapter). For each of the models, there were two degrees of freedom, since two additional parameters were estimated for each of the models, compared to the null model. These two additional parameters corresponded to the two highest or most preferred levels of the attributes -- as explained later in this section, the coding of categorical variables in the CBC System omits the last level of the independent variable, e.g., MOB-1 and MOB-2 represent the highest or most preferred levels of mobility -- no problems in walking about and some problems in walking about respectively. The results of this first step in modeling the DCCM data are presented in Table 25.

Each of the individual health status attributes in Table 25 is significant. It is also important to consider the operational significance of each of the single attributes. As noted earlier, the EuroQol health status classification system was based on a review of existing health status measurement instruments, which in turn were developed after extensive interviews with respondents regarding the health status dimensions or attributes which were important to them. Therefore, it would stand to reason that each of the five dimensions of the EuroQol needs to be represented in any preference modeling exercise of the EuroQol. Therefore, on account of statistical as well as operational reasons, it was decided to fit a main-effects only model to the data, which included all the five health status dimensions of the EuroQol. The results of this step of the modeling procedure are shown in Table 26. As explained earlier, only two levels of each dimension are shown, since the third level was deleted for the purpose of the analysis.
### Table 25: Modeling the DCCM Data I: Univariate Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Coeff&lt;sup&gt;a&lt;/sup&gt;</th>
<th>SE Coeff&lt;sup&gt;b&lt;/sup&gt;</th>
<th>LL'</th>
<th>G&lt;sup&gt;d&lt;/sup&gt;</th>
<th>p&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Null Model&lt;sup&gt;i&lt;/sup&gt;</td>
<td></td>
<td></td>
<td>-1,926.949</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) MOB-1&lt;sup&gt;k&lt;/sup&gt;</td>
<td>1.109</td>
<td>0.068</td>
<td>-1,145.155</td>
<td>1,563.588</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>MOB-2</td>
<td>0.919</td>
<td>0.069</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) SFC-1&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.412</td>
<td>0.038</td>
<td>-1,441.372</td>
<td>971.154</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>SFC-2</td>
<td>0.157</td>
<td>0.041</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) ACT-1&lt;sup&gt;i&lt;/sup&gt;</td>
<td>0.298</td>
<td>0.037</td>
<td>-1,489.102</td>
<td>875.694</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>ACT-2</td>
<td>0.122</td>
<td>0.039</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) PDF-1&lt;sup&gt;i&lt;/sup&gt;</td>
<td>0.419</td>
<td>0.041</td>
<td>-1,360.871</td>
<td>1,132.156</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>PDF-2</td>
<td>0.442</td>
<td>0.038</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) ANX-1&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.472</td>
<td>0.038</td>
<td>-1,440.091</td>
<td>973.716</td>
<td>&lt;0.005</td>
</tr>
<tr>
<td>ANX-2</td>
<td>0.058</td>
<td>0.039</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- a: n = 139 for all six models
- b: maximum likelihood estimate of parameter
- c: LL = log likelihood
- d: G = likelihood ratio statistic
  - = - 2 (LL of test model - LL of null model)
- e: as determined from chi-square tables with df = 2 in all cases
- f: consisting of all coefficients equal to zero
- g: univariate model for mobility
- h: univariate model for self-care
- i: univariate model for usual activities
- j: univariate model for pain/discomfort
- k: univariate model for anxiety/depression
Table 26: Modeling the DCCM Data II: Main-effects Only Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coeff</th>
<th>SE Coeff</th>
<th>Wald statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOB-1</td>
<td>1.882</td>
<td>0.103</td>
<td>18.232*</td>
</tr>
<tr>
<td>MOB-2</td>
<td>1.362</td>
<td>0.093</td>
<td>14.603*</td>
</tr>
<tr>
<td>SFC-1</td>
<td>1.041</td>
<td>0.077</td>
<td>13.550*</td>
</tr>
<tr>
<td>SFC-2</td>
<td>0.422</td>
<td>0.070</td>
<td>6.073*</td>
</tr>
<tr>
<td>ACT-1</td>
<td>0.789</td>
<td>0.073</td>
<td>10.834*</td>
</tr>
<tr>
<td>ACT-2</td>
<td>0.344</td>
<td>0.066</td>
<td>5.181*</td>
</tr>
<tr>
<td>PDF-1</td>
<td>1.054</td>
<td>0.081</td>
<td>13.033*</td>
</tr>
<tr>
<td>PDF-2</td>
<td>0.760</td>
<td>0.068</td>
<td>11.178*</td>
</tr>
<tr>
<td>ANX-1</td>
<td>0.702</td>
<td>0.073</td>
<td>9.684*</td>
</tr>
<tr>
<td>ANX-2</td>
<td>0.496</td>
<td>0.072</td>
<td>6.920*</td>
</tr>
</tbody>
</table>

Null Model Log Likelihood = -1,926.949 (see Table 25)
Main-effects only Model Log Likelihood = -551.510

Therefore, Likelihood Ratio Statistic (G) = 2,750.878 (p<0.005)

Legend:
a: n = 139  
b: MOB = mobility; SFC = self-care; ACT = usual activities; PDF = pain/discomfort; ANX = anxiety/depression  
c: maximum likelihood estimate of parameters  
d: Wald statistic = coeff / SE coeff  
e: p<0.05
Table 26 shows that the main-effects only model was significant. Further, each of the individual coefficients was also significant. Thus, even with the multivariate model, each individual health status dimension was significant. The next step is to model interactions among health status dimensions. For this purpose, a series of models with main-effects plus single two-way interactions were fit to the data and compared to the main-effects only model using the likelihood ratio statistic (G) as the test statistic. Each model had four degrees of freedom (two parameters were estimated for each of the interacting variables; thus four additional parameters were estimated compared to the main-effects only model). Table 27 summarizes the results of this step of the modeling procedure.

On the basis of the results of adding single two-way interactions to the main-effects only model presented in Table 27, it can be seen that only one such interaction -- between self-care and usual activities or model 6 in Table 27-- is significant. This would hold true even at the rather liberal level of significance of 0.25 suggested by Hosmer and Lemeshow (1989) [which in turn was based on the work of Bendel and Afifi (1977) and Mickey and Greenland (1989)] for selection of variables in logistic regression modeling.

The next step in the modeling procedure is to fit multiple two-way interactions to the DCCM data, i.e., fit models containing main-effects plus two-way interactions between self-care and usual activities as well as other two-way interactions. The significance of each of these models was evaluated by comparing them to the main-effects plus self-care*usual activities interaction model using the likelihood ratio statistic (G) as the test statistic.
Table 27: Modeling the DCCM Data III: Main-effects and Single Two-way Interactions

<table>
<thead>
<tr>
<th>Model</th>
<th>LL</th>
<th>G</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Main-effects only</td>
<td>-551.510</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) MOB * SFC</td>
<td>-550.689</td>
<td>1.642</td>
<td>0.80</td>
</tr>
<tr>
<td>3) MOB * ACT</td>
<td>-550.595</td>
<td>1.830</td>
<td>0.77</td>
</tr>
<tr>
<td>4) MOB * PDF</td>
<td>-550.206</td>
<td>2.608</td>
<td>0.63</td>
</tr>
<tr>
<td>5) MOB * ANX</td>
<td>-550.345</td>
<td>2.330</td>
<td>0.68</td>
</tr>
<tr>
<td>6) SFC * ACT</td>
<td>-546.498</td>
<td>10.024</td>
<td>0.04</td>
</tr>
<tr>
<td>7) SFC * PDF</td>
<td>-549.836</td>
<td>3.348</td>
<td>0.50</td>
</tr>
<tr>
<td>8) SFC * ANX</td>
<td>-549.103</td>
<td>4.814</td>
<td>0.32</td>
</tr>
<tr>
<td>9) ACT * PDF</td>
<td>-550.567</td>
<td>1.886</td>
<td>0.76</td>
</tr>
<tr>
<td>10) ACT * ANX</td>
<td>-550.828</td>
<td>1.364</td>
<td>0.85</td>
</tr>
<tr>
<td>11) PDF * ANX</td>
<td>-550.583</td>
<td>1.854</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Legend:
a: n = 139
b: MOB = mobility; SFC = self-care; ACT = usual activities;
   PDF = pain/discomfort; ANX = anxiety/depression;
e.g., MOB * SFC = interaction between mobility and self-care
c: LL = log likelihood
c: G = likelihood ratio statistic
d: as determined from chi-square tables with df = 4 in all cases;
   linear extrapolation used for values in between entries of tables
Table 28: Modeling the DCCM Data IV: Main-effects and Multiple Two-way Interactions

<table>
<thead>
<tr>
<th>Model</th>
<th>LL</th>
<th>G</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Main-effects &amp; SFC * ACT</td>
<td>-546.498</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) MOB * SFC</td>
<td>-545.386</td>
<td>2.224</td>
<td>0.70</td>
</tr>
<tr>
<td>3) MOB * ACT</td>
<td>-544.934</td>
<td>3.128</td>
<td>0.54</td>
</tr>
<tr>
<td>4) MOB * PDF</td>
<td>-545.290</td>
<td>2.416</td>
<td>0.66</td>
</tr>
<tr>
<td>5) MOB * ANX</td>
<td>-545.465</td>
<td>2.066</td>
<td>0.72</td>
</tr>
<tr>
<td>6) SFC * PDF</td>
<td>-544.479</td>
<td>4.038</td>
<td>0.42</td>
</tr>
<tr>
<td>7) SFC * ANX</td>
<td>-544.107</td>
<td>4.782</td>
<td>0.33</td>
</tr>
<tr>
<td>8) ACT * PDF</td>
<td>-545.625</td>
<td>1.746</td>
<td>0.78</td>
</tr>
<tr>
<td>9) ACT * ANX</td>
<td>-545.616</td>
<td>1.764</td>
<td>0.78</td>
</tr>
<tr>
<td>10) PDF * ANX</td>
<td>-545.609</td>
<td>1.778</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Legend:
   a: n = 139
   b: MOB = mobility; SFC = self-care; ACT = usual activities;
       PDF = pain/discomfort; ANX = anxiety/depression;
       e.g., MOB * SFC = interaction between mobility and self-care
   c: LL = log likelihood
   c: G = likelihood ratio statistic
       = -2 (LL of test model - LL of main-effects & SFC*ACT model)
   d: as determined from chi-square tables with df = 4 in all cases;
       linear extrapolation used for values in between entries of tables
For each of the models, there were four degrees of freedom (since two parameters were estimated for each of the interacting variables, thus leading to four additional parameters over and above the main-effects plus self-care*usual activities interaction model). The results of this step of the modeling procedure are shown in Table 28.

On the basis of the results summarized in Table 28, it can be seen that none of the other two-way interactions are significant, even at the liberal 0.25 level of significance suggested for screening purposes by Hosmer and Lemeshow (1989). Therefore, on the basis of the results of the stepwise multinomial logit modeling procedure, the model form of the preference function (more specifically, the value function) obtained via discrete choice conjoint methodology is represented by:

\[
p(x) = MOB + SFC + ACT + PDF + ANX + SFC*ACT \quad (6.5)
\]

where

- **MOB** = maximum likelihood estimate for level of mobility
- **SFC** = maximum likelihood estimate for level of self-care
- **ACT** = maximum likelihood estimate for level of usual activities
- **PDF** = maximum likelihood estimate for level of pain/discomfort
- **ANX** = maximum likelihood estimate for level of anxiety/depression
- **SFC*ACT** = maximum likelihood estimate for levels of the interaction between self-care and usual activities
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
<th>SE Coef.</th>
<th>Wald Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 29: Modeling the OCM Data: Final Model Form
Table 29 (continued)

Null Model Log Likelihood = -1,926.949 (see Table 25)
This Model Log Likelihood = -546.498

Therefore, Likelihood Ratio Statistic (G) = 2,760.902 (p<0.005)

Legend:

a: n = 139
b: MOB = mobility; SFC = self-care; ACT = usual activities;
Pain/discomfort; ANX = anxiety/depression;
e.g., MOB * SFC = interaction between mobility and self-care
c: maximum likelihood estimate of parameters
d: Wald statistic = coeff / SE coeff
e: p<0.05
f: level deleted from analysis due to coding (see text for details)
The specific maximum likelihood parameter estimates for the variables in this model are presented in Table 29. A few points about Table 29 need to be noted. First, apart from the significance of the overall model, it is important to note that each of the individual variables in the model is also significant. Second, there are several entries in the table which are represented by dashed lines. This is because the particular levels of the variables corresponding to these entries were deleted from the analysis. Multinomial logit analysis (like least squares analysis) suffers from the dependencies among attribute levels when all the levels of an attribute are entered into the computational analysis. For this reason, the last level of each attribute is deleted from the analysis by the CBC System.

On account of the manner in which the levels of each attribute were programmed into the CBC System, the lowest or least preferred levels of each attribute were deleted from the analysis, i.e., MOB-3, SFC-3, ACT-3, PDF-3, and ANX-3. The remaining levels are represented using effects coding, which is a way of coding categorical attributes so that the average effect within each attribute will be zero. In other words, each maximum likelihood estimate of the coefficient of the level in question will represent a deviation from the mean within the attribute of which the level is descriptive.

After the logit computation, the CBC System provides estimates of the coefficients of the deleted levels of attributes by computing the negative of the sum of the included levels. For example, in Table 29, the maximum likelihood estimates of the coefficients of the two highest levels of mobility are 1.921 and 1.385 respectively. The inferred estimate of the coefficient of
the last (i.e., deleted) level of mobility is the negative of the sum of 1.921 and 1.385, or -3.306. Although this estimate of the coefficient of the deleted level is not required for modeling the data, it does prove useful in calculating overall preference scores for multiattribute alternatives, as explained later in this section. Since the standard errors of the deleted levels are not needed either for modeling purposes or for calculating overall preference scores for multiattribute alternatives, they are not provided as part of the routine output of the CBC System and are therefore not shown in Table 29.

The maximum likelihood estimates of the coefficients of the levels of the attributes included in the analysis and the inferred coefficients of the levels of the attributes not included in the analysis are used for the purposes of estimating the predictive accuracy of the modeling procedure as well as for calculating overall preference scores for multiattribute alternatives in order to address the research questions of the next section.

First, in terms of the predictive accuracy of the modeling procedure, the CBC System offers a share simulation estimation, which uses the estimated model form (i.e., main effects plus the self-care*usual activities interaction) as the basis for predicting shares among competing alternatives. As explained in Chapter V, descriptions of the alternatives included in the two holdout choice sets were input into this share simulator, and the estimated shares that were the output of this simulator were compared to the actual shares of the two holdout choice sets for which data were collected. Table 30 presents the calculated and observed shares of the three multiattribute health states, along with each of the residuals for both the holdout choice sets.
Table 30: Predictive Ability of DCCM

<table>
<thead>
<tr>
<th>Multiattribute Health State&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Calculated Share (C)&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Observed Share (M)</th>
<th>Residual (C-M)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Holdout Choice Set # 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BO</td>
<td>0.114</td>
<td>0.295</td>
<td>-0.181</td>
</tr>
<tr>
<td>E3</td>
<td>0.879</td>
<td>0.698</td>
<td>0.181</td>
</tr>
<tr>
<td>E8</td>
<td>0.003</td>
<td>0.000</td>
<td>0.003</td>
</tr>
<tr>
<td>ID</td>
<td>0.004</td>
<td>0.007</td>
<td>-0.003</td>
</tr>
<tr>
<td>Mean Absolute Residual</td>
<td></td>
<td></td>
<td>0.092</td>
</tr>
<tr>
<td>Standard Deviation of Absolute Residual</td>
<td></td>
<td></td>
<td>0.103</td>
</tr>
<tr>
<td><strong>Holdout Choice Set # 2&lt;sup&gt;d&lt;/sup&gt;</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS1</td>
<td>0.921</td>
<td>0.957</td>
<td>-0.036</td>
</tr>
<tr>
<td>HS2</td>
<td>0.004</td>
<td>0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>HS3</td>
<td>0.074</td>
<td>0.043</td>
<td>0.031</td>
</tr>
<tr>
<td>ID</td>
<td>0.001</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>Mean Absolute Residual</td>
<td></td>
<td></td>
<td>0.018</td>
</tr>
<tr>
<td>Standard Deviation of Absolute Residual</td>
<td></td>
<td></td>
<td>0.018</td>
</tr>
</tbody>
</table>

Legend:

<sup>a</sup>: n = 139
<sup>b</sup>: the labels BO, E3, and E8 were used to identify the multiattribute health states in the visual aids (see Appendix B)
<sup>c</sup>: calculated using equation (6.3)
<sup>d</sup>: HS = health state (these health states were not described by cards)
In summary, the modeling of the data obtained using the DCCM methodology supported the existence of a multiplicative model form for the preference function for health status as operationalized by the EuroQol, where the only two-way interaction which was significant was the one between self-care and usual activities. The average error in prediction of health status preferences using this methodology was variable across the two holdout choice sets -- for the first, it was about 9%, while for the second it was a little under 2%. Although this averages to about 5.5%, the variability in prediction error warrants caution in interpreting this "global average" of 5.5%. Issues pertinent to the results of modeling of the data using DCCM will be discussed in Chapter VII.

In order to permit comparison between the health status preference scores obtained using the different preference measurement methodologies, it was necessary to convert the DCCM preference scores of health states (obtained by summing up the relevant maximum likelihood coefficients) into a 0.0 to 1.0 scale, where -- in order to be consistent with the preference scores obtained using the conditional utility function-based procedure with the rating scale and standard gamble scaling methods -- a preference score of 1.0 is the score given to the all-best state (with the best level of functioning on all five attributes) and a preference score of 0.0 is the score given to the state of immediate death. Similarly, for each individual attribute, the highest or most preferred level would be given a score of 1.0 and the lowest or least preferred level would be given a score of 0.0.
In order to make this conversion for each of the levels, (a) the preference score of the lowest level was set to 0.0 by addition of the appropriate value (the negative of its own maximum likelihood coefficient) to its estimated coefficient, (b) this same appropriate value was added to the estimates of the other levels, (c) the preference score of the top level was set to 1.0 by dividing its converted value from (b) above by itself, and (d) the converted value of the middle level from (b) above was divided by the converted value of the top level, to yield a preference score for the middle level between 0.0 and 1.0 (thus making it comparable with the preference scores obtained using the conditional utility function-based procedures with the rating scale and standard gamble scaling methods). An illustration of this conversion procedure for individual levels of health status attributes is provided in Table 31.

In order to make this conversion for the multiattribute health states, the individual attribute-level maximum likelihood estimates for the main-effects plus self-care*usual activities interaction model were summed for all the relevant effects for the health states in question. In this regard, for the all-best health state, the relevant effects were the parameter estimates for the highest level of functioning for each of the five attributes and the parameter estimate for the interaction between the highest level of self-care and the highest level of usual activities. As each choice set contained the immediate death state, the analysis provided a parameter estimate for this state, i.e., no summation procedures were necessary for the state of immediate death.
Consider the health status attribute of mobility.

Maximum likelihood coefficients of levels of mobility:
\[
\begin{align*}
\text{MOB-1} &= 1.921 \\
\text{MOB-2} &= 1.385 \\
\text{MOB-3} &= -3.306
\end{align*}
\]

Step (a) -- Set the preference score of the lowest level to 0.0 by addition of appropriate value:
\[
\text{MOB-3} = -3.306 + 3.306 = 0.0
\]

Step (b) -- Add this same appropriate value to the maximum likelihood coefficients of the other levels:
\[
\begin{align*}
\text{MOB-1} &= 1.921 + 3.306 = 5.227 \\
\text{MOB-2} &= 1.385 + 3.306 = 4.691
\end{align*}
\]

Step (c) -- Set the preference score of the top level to 1.0 by self-division:
\[
\text{MOB-1} = 5.227 / 5.227 = 1.0
\]

Step (d) -- Divide the converted value of the middle level from Step (b) by the converted value of the top level, also from Step (b):
\[
\text{MOB-2} = 4.691 / 5.227 = 0.897
\]

Therefore, the preference scores of mobility on the 0.0-1.0 scale are:
\[
\begin{align*}
\text{MOB-1} &= 1.000 \\
\text{MOB-2} &= 0.897 \\
\text{MOB-3} &= 0.000
\end{align*}
\]
The specific steps in the conversion procedure of multiattribute health state preference scores from their maximum likelihood coefficients to the 0.0-1.0 scale were as follows: (a) the immediate death state was set as the bottom anchor of the 0.0-1.0 scale by adding the appropriate value (the negative of its own maximum likelihood coefficient) to its estimated coefficient, (b) this same appropriate value was added to the summated maximum likelihood coefficients of the relevant effects (i.e., all five main-effects and the two-way self-care*usual activities interaction) of the multiattribute health states, (c) the all-best state was set as the top anchor of the 0.0-1.0 scale, by dividing its converted score -- obtained from (b) above -- by itself, and (d) all the other health states needed for further analyses were mapped onto this rescaled 0.0-1.0 continuum by dividing their summated scores obtained in (b) above by the converted score of the all-best health state obtained from (b) above. An illustration of this conversion procedure for multiattribute health states is provided in Table 32.

Therefore, preference scores for eight multiattribute health states (i.e., five corner states and three interior states) were obtained on a 0.0 to 1.0 scale, where 0.0 was the preference score for immediate death and 1.0 was the preference score for the all-best state. These were the preference scores which were used to address each of the research questions investigated in this study. Preference scores for all levels of the five health status attributes on the 0.0 to 1.0 scale using all three methodologies are presented in Table 33; preference scores for the eight multiattribute health states on the 0.0 to 1.0 scale using all three methodologies are presented in Table 34.
Table 32: Illustration of Conversion of DCCM Scores of Multiattribute Health States to 0.0-1.0 Scale

Consider the multiattribute health state BO: i.e., MOB-3, SFC-2, ACT-1, PDF-1, ANX-1

Calculation of preference score for all-best health state using maximum likelihood coefficients:

\[
\begin{align*}
\text{MOB-1} &= 1.921 \\
\text{SFC-1} &= 1.039 \\
\text{ACT-1} &= 0.769 \\
\text{PDF-1} &= 1.064 \\
\text{ANX-1} &= 0.723 \\
\text{SFC-1} \times \text{ACT-1} &= 0.299 \\
\text{Total} &= 5.815
\end{align*}
\]

Preference score for immediate death state (maximum likelihood coefficient):

\[
\text{ID}^* = -3.856
\]

Step (a) -- Set the preference score of the immediate death state to 0.0 by addition of the appropriate value:

\[
\text{ID} = -3.856 + 3.856 = 0.0
\]
Table 32 (continued)

Step (b) -- Add this same appropriate value to the summated maximum likelihood coefficients of the relevant effects of BO:

(i) summated maximum likelihood coefficients for BO:
   - \( \text{MOB-3} = -3.306 \)
   - \( \text{SFC-2} = 0.395 \)
   - \( \text{ACT-1} = 0.769 \)
   - \( \text{PDF-1} = 1.064 \)
   - \( \text{ANX-1} = 0.723 \)
   - \( \text{SFC-2} \times \text{ACT-1} = -0.303 \)

\[ \text{Total} = -0.658 \]

(ii) adding appropriate value of ID to summated score of BO:
\[ \text{BO} = -0.658 + 3.856 = 3.198 \]

Step (c) -- Set the preference score of the all-best state to 1.0 by self-division:

(i) adding appropriate value of ID to summated score of AB:
\[ \text{AB} = 5.815 + 3.856 = 9.671 \]

(ii) setting this converted score to 1.0 by self-division:
\[ \text{AB} = 9.671 / 9.671 = 1.0 \]

Step (d) -- Mapping BO onto this 0.0-1.0 scale:
\[ \text{BO} = 3.198 / 9.671 = 0.331 \]

Therefore, the preference score of BO on the 0.0-1.0 scale is \( 0.331 \).

Legend:
a: the labels BO, AB, and ID were used for identification purposes in the visual aids (see Appendix B)
Table 33: Preference Scores for all Individual Attribute Levels on the 0.0 to 1.0 Scale*

<table>
<thead>
<tr>
<th>Level</th>
<th>CUF-RS</th>
<th>CUF-SG</th>
<th>DCCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOB-1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>MOB-2</td>
<td>0.58</td>
<td>0.81</td>
<td>0.90</td>
</tr>
<tr>
<td>MOB-3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SFC-1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>SFC-2</td>
<td>0.58</td>
<td>0.81</td>
<td>0.74</td>
</tr>
<tr>
<td>SFC-3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ACT-1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>ACT-2</td>
<td>0.59</td>
<td>0.82</td>
<td>0.76</td>
</tr>
<tr>
<td>ACT-3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>PDF-1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>PDF-2</td>
<td>0.59</td>
<td>0.82</td>
<td>0.89</td>
</tr>
<tr>
<td>PDF-3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ANX-1</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>ANX-2</td>
<td>0.58</td>
<td>0.81</td>
<td>0.88</td>
</tr>
<tr>
<td>ANX-3</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139  

b: MOB = mobility; SFC = self-care; ACT = usual activities; PDF = pain/discomfort; ANX = anxiety/depression;  
c: conditional utility function-based procedure with rating scale  
d: conditional utility function-based procedure with standard gamble  
e: discrete choice conjoint methodology
Table 34: Preference Scores for Multiattribute Health States on the 0.0 to 1.0 Scale\(^{ab}\)

<table>
<thead>
<tr>
<th>Multiattribute Health State(^c)</th>
<th>CUF-RS(^a)</th>
<th>CUF-SG(^e)</th>
<th>DCCM(^f)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Anchor States:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All-best (AB)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Immediate Death (ID)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td><strong>Corner States:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility (CM)</td>
<td>0.40</td>
<td>0.66</td>
<td>0.46</td>
</tr>
<tr>
<td>Self-care (CS)</td>
<td>0.65</td>
<td>0.84</td>
<td>0.71</td>
</tr>
<tr>
<td>Usual Activities (CU)</td>
<td>0.68</td>
<td>0.85</td>
<td>0.76</td>
</tr>
<tr>
<td>Pain/Discomfort (CP)</td>
<td>0.56</td>
<td>0.77</td>
<td>0.70</td>
</tr>
<tr>
<td>Anxiety/Depression (CA)</td>
<td>0.63</td>
<td>0.83</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>Interior States:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interior State #1 (BO)</td>
<td>0.32</td>
<td>0.63</td>
<td>0.33</td>
</tr>
<tr>
<td>Interior State #2 (E3)</td>
<td>0.37</td>
<td>0.64</td>
<td>0.53</td>
</tr>
<tr>
<td>Interior State #3 (E8)</td>
<td>0.16</td>
<td>0.45</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

Legend:
\(a\): \(n = 139\)
\(b\): where 1.00 = best imaginable health state and 0.00 = worst imaginable health state
\(c\): the letters in parentheses correspond to the labels used to identify the multiattribute health states in the visual aids (see Appendix B)
\(d\): conditional utility function-based procedure with rating scale
\(e\): conditional utility function-based procedure with standard gamble
\(f\): discrete choice conjoint methodology
It should be noted that all analyses were conducted at the aggregate level, using the arithmetic mean as the measure of central tendency. Further, only the preference scores for multiattribute health states (i.e., Table 34) were used for the analyses conducted to address Research Questions 7 through 14. This was on account of the fact that individual attribute levels are but building blocks for the multiattribute alternatives, and can be combined in different ways to form the overall preference score for the multiattribute alternatives. Therefore, the parameter estimates of individual attributes may be different if they are measured using different preference measurement methodologies, but the overall preference score for the multiattribute alternative obtained by the different methodologies may still be similar because, as posited by change of process theory (Mellers, Ordonez, and Birnbaum, 1992) -- see Chapter IV -- different measurement methodologies might prompt different combination functions or model forms for preference functions due to response mode effects.

Although the data presented in Table 33 were not be the subject of any analytical procedures conducted in order to address the research questions which were investigated in this study, it should be noted that there seems to be a systematic order to the preference scores for the individual attribute levels as determined by the three preference measurement methodologies. Specifically, the CUF-RS gives the lowest preference scores among all three methodologies for all the levels in question. The relative order of preference scores between the CUF-SG and DCCM methodologies is variable, but seems to be affected by the presence
of interactions between health status attributes -- the only two levels for which the DCCM methodology gave a lower preference score than the CUF-SG methodology were the levels of the health status attributes which were shown to interact in the DCCM data. Therefore, it is possible that the presence of the two-way interaction between self-care and usual activities suppressed the preference scores for the levels of these two attributes (since the "true" preference score of any multiattribute health state which included these two attributes would also take into account the preference score for the interaction between the attributes).

Moreover, a few points about Table 34 need to be noted. First, in terms of the relative ordering of multiattribute health states by preference measurement methodology, the CUF-RS methodology usually gives the lowest preference scores, while the CUF-SG methodology usually gives the highest preference scores to multiattribute health states. The only exception to this general pattern is for the case of the multiattribute health state E8, for which the DCCM methodology gave the lowest preference score. The issue of the statistical significance of these differences will be addressed in Research Question # 9, later in this chapter.

It is instructive to compare these relative orderings with those of the individual attribute levels, presented in Table 33. In both cases, the CUF-RS methodology gave the lowest preference scores. However, in the case of individual attribute levels the DCCM gave the highest scores, while in the case of the multiattribute health states the CUF-SG gave the highest scores. The implications of this difference will be discussed in Chapter VII.
The second point that needs to be noted about Table 34 is the preferential ordering of health states within each preference measurement methodology. In this regard, it can be seen that there is perfect agreement between the CUF-RS and CUF-SG methodologies in terms of the preferential ordering of multiattribute health states, while the DCCM methodology orders multiattribute health states differently than the other methodologies. The statistical significance of these differences will be addressed in Research Questions #7 and 8, which are considered in the next section of this chapter.

Results of the Comparison of Preference Scores

Research Questions 7 through 14 pertained to comparing the preference scores obtained using the three preference measurement methodologies. This section will discuss the results of each of these eight questions. As noted in the previous chapter, all analytical procedures were conducted at three different levels: (a) at the overall level (i.e., all eight multiattribute health states), (b) at the level of the five corner states separately, and (c) at the level of the three interior states separately. The results of all three levels of analysis will be presented in this section. For the purpose of consistency with the format adopted in Chapters II and V and to help organize the section, the results are presented in five sub-sections, corresponding to the different bases of comparison of the preference scores obtained using the three preference measurement methodologies, i.e., preferential ordering, magnitude of preference scores, correlations, concordances, and functional relationships.
Preferential Ordering

The first aspect on which the different preference measurement methodologies were compared was in terms of the order in which they ranked the alternative health states under evaluation. Two specific research questions were investigated in this regard. The results of the analytical procedures conducted to address each will be discussed in turn.

Research Question # 7: Do the three health status preference measurement methodologies (i.e., the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) agree in terms of the order in which they rank alternative states of health?

As noted in Chapter V, this question was addressed by computing Kendall's Coefficient of Concordance (W). The results of all three levels of the analysis are summarized in Table 35. Even after adjusting for multiple comparisons with the Bonferroni correction, there is significant agreement among all three preference measurement methodologies in terms of the order in which they rank alternative health states. This conclusion holds at all levels of the analysis (although no statistical significance tests were conducted for the analysis on the interior states since the sample size of three was too small, a value of Kendall’s W of 1.00 indicates perfect agreement).
Table 35: Overall Agreement in Rank Ordering of the Three Methodologies

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>Kendall’s W</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.96</td>
<td>0.007b</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.87</td>
<td>0.005</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>1.00</td>
<td>---c</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: estimated by linear interpolation between table entries
c: significance testing not carried out due to small sample size
Research Question # 8(a): Do the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology agree in terms of the order in which they rank alternative states of health?

Research Question # 8(b): Do the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM agree in terms of the order in which they rank alternative states of health?

Research Question # 8(c): Do the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM agree in terms of the order in which they rank alternative states of health?

This set of research questions was addressed using Kendall’s τ to indicate agreement in preferential ordering of multiattribute health states between each of the preference measurement methodologies taken two at a time. Regarding the comparison between the CUF-RS and CUF-SG methodologies, there was perfect agreement at all three levels of the analysis. Although the same result held at the level of the interior states in the comparison between the DCCM and other methodologies investigated in this study, there was less agreement between the preferential ordering of
multiattribute health states produced by the DCCM and the other methodologies investigated in this study in terms of the analysis of the five corner states as well as all the eight multiattribute health states. Indeed, the value of Kendall’s $\tau$ between the DCCM and the other methodologies for the analysis of the corner states did not even approach statistical significance -- even without applying the Bonferroni correction for multiple comparisons. These results are presented in Tables 36 through 38.

**Magnitude of Preference Scores**

The second aspect on which the different preference measurement methodologies were compared was in terms of the magnitude of the preference scores they ascribed to the alternative health states under evaluation. The specific research question investigated in this regard was stated as follows:

**Research Question # 9:** Do the three health status preference measurement methodologies (i.e., the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) agree in terms of the magnitude of the preference scores they ascribe to alternative states of health?
Table 36: Results of Kendall’s τ Between CUF-RS* and CUF-SG°

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>τ^d</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>1.000</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>1.000</td>
<td>0.008</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>1.000</td>
<td>---研</td>
</tr>
</tbody>
</table>

Legend:
- a: conditional utility function-based procedure with rating scale
- b: conditional utility function-based procedure with standard gamble
- c: n = 139
- d: Kendall’s τ
- e: significance testing not carried out due to small sample size
Table 37: Results of Kendall's $\tau$ Between CUF-RS$^a$ and DCCM$^{bc}$

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>$\tau^d$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.786</td>
<td>0.003</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.600</td>
<td>0.117</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>1.000</td>
<td>---&quot;</td>
</tr>
</tbody>
</table>

Legend:

a: conditional utility function-based procedure with rating scale  
b: discrete choice conjoint methodology  
c: n = 139  
d: Kendall's tau  
e: significance testing not carried out due to small sample size
Table 38: Results of Kendall's $\tau$ Between CUF-SG' and DCCM$^{bc}$

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>$\tau^d$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.786</td>
<td>0.003</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.600</td>
<td>0.117</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>1.00</td>
<td>---$^e$</td>
</tr>
</tbody>
</table>

Legend:

a: conditional utility function-based procedure with standard gamble
b: discrete choice conjoint methodology
c: $n = 139$
d: Kendall's tau
e: significance testing not carried out due to small sample size
This question was addressed by a series of repeated measures one-way analyses of variance with a Scheffe test for multiple comparisons. At the overall level (i.e., all eight multiattribute health states), the omnibus F-test was significant at the 0.05 level. The omnibus F-test was based on an adjusted F-test, since the covariance matrix of the repeated measures was found to be nonspherical. Pairwise comparisons using a Scheffe correction for multiple comparisons revealed there were significant differences (familywise p<0.05) between preference scores obtained using the CUF-RS and CUF-SG and between the CUF-SG and the DCCM methodologies. However, no significant differences were found between preference scores obtained using the CUF-RS and DCCM methodologies.

The analysis on the five corner states yielded slightly different results than the analysis on all eight multiattribute health states (i.e., five corner states and three interior states). Like the results of the previous analysis, the omnibus F-test for the analysis on the five corner states was significant at the 0.05 level. Unlike the omnibus F-test for the analysis on all eight multiattribute health states, the omnibus F-test for the analysis on the five corner states was based on the usual unadjusted F-test since the covariance matrix of the repeated measures was found to be spherical (on the basis of the test for sphericity). Pairwise comparisons using a Scheffe correction for multiple comparisons revealed that there were significant differences (familywise p<0.05) between the preference scores obtained using the CUF-RS and CUF-SG methodologies. However, in contrast to the results of the analysis on all eight multiattribute health states, no significant differences were found between the preference scores obtained using the CUF-SG and
the DCCM methodologies for the analysis on the five corner states, while significant differences were found between the preference scores obtained using the CUF-RS and the DCCM methodologies.

In terms of the analysis of the three interior states, the omnibus F-test was not significant at the 0.05 level. This maybe due the small sample size, and consequently less amount of statistical power to detect any significant differences. The results of the analyses pertinent to addressing this research question are presented in Table 39.

Correlations

The third aspect on which the different preference measurement methodologies were compared was in terms of the correlations between the preference scores of multiattribute health states obtained using the different methodologies. Two specific research questions were investigated in this regard. The data analysis procedures for each will be discussed in turn in this sub-section.

Research Question # 10(a): Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?
Table 39: Results of Repeated Measures One-way ANOVA on Preference Scores Obtained Using the Three Methodologies

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Omnibus F</th>
<th>Pairwise Comparisons&lt;sup&gt;b&lt;/sup&gt;</th>
<th>CUF-RS&lt;sup&gt;c&lt;/sup&gt;</th>
<th>CUF-RS</th>
<th>CUF-SG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>CUF-SG&lt;sup&gt;d&lt;/sup&gt;</td>
<td>DCCM&lt;sup&gt;e&lt;/sup&gt;</td>
<td>DCCM</td>
</tr>
<tr>
<td>Corner &amp; Interior States (n = 8)</td>
<td>19.12&lt;sup&gt;f&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;a&lt;/sup&gt;</td>
<td>No&lt;sup&gt;h&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;g&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Corner States Only (n = 5)</td>
<td>40.65&lt;sup&gt;f&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;g&lt;/sup&gt;</td>
<td>No&lt;sup&gt;h&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Interior States Only (n = 3)</td>
<td>7.89</td>
<td>---&lt;sup&gt;i&lt;/sup&gt;</td>
<td>---&lt;sup&gt;i&lt;/sup&gt;</td>
<td>---&lt;sup&gt;i&lt;/sup&gt;</td>
<td>---&lt;sup&gt;i&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Legend:
- a: n = 139
- b: evaluated using a Scheffe correction for multiple comparisons
- c: conditional utility function-based procedure with rating scale
- d: conditional utility function-based procedure with standard gamble
- e: discrete choice conjoint methodology
- f: p<0.05
- g: pairwise comparison was significant (familywise p<0.05)
- h: pairwise comparison was not significant
- i: pairwise comparisons not conducted due to nonsignificant omnibus F-test
Research Question # 10(b): Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

Research Question # 10(c): Is there a significant correlation between the preferential rank order of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?

This set of research questions was addressed using Spearman’s rho or $\rho$ to measure the rank-order correlation between the preference measurement methodologies taken two at a time. There was perfect positive rank-order correlation between the CUF-RS and CUF-SG methodologies at all three levels of the analysis. The same result held at the level of the interior states in the comparison between the DCCM and other methodologies investigated. However, the rank-order correlation coefficients between the preference scores of multiattribute health states obtained using the DCCM and the other methodologies investigated was less than perfect in the case of the analyses at the levels of the five corner states and eight multiattribute health states. Indeed, the value of Spearman’s $\rho$ between the DCCM and the other methodologies for the analysis of the corner states was not statistically significant, even without applying the Bonferroni correction. These results are presented in Tables 40 through 42.
Table 40: Results of Spearman's ρ Between CUF-RS* and CUF-SGbx

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>ρd</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>1.000</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>1.000</td>
<td>0.008</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>1.000</td>
<td>---e</td>
</tr>
</tbody>
</table>

Legend:
- a: conditional utility function-based procedure with rating scale
- b: conditional utility function-based procedure with standard gamble
- c: n = 139
- d: Spearman's rho
- e: significance testing not carried out due to small sample size
Table 41: Results of Spearman's ρ Between CUF-RS and DCCM

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>ρ ( ^d )</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.905</td>
<td>0.002</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.700</td>
<td>0.117</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>1.000</td>
<td>---</td>
</tr>
</tbody>
</table>

Legend:
- a: conditional utility function-based procedure with rating scale
- b: discrete choice conjoint methodology
- c: n = 139
- d: Spearman's rho
- e: significance testing not carried out due to small sample size
Table 42: Results of Spearman’s $\rho$ Between CUF-SG$^*$ and DCCM$^{bc}$

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>$\rho^d$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.905</td>
<td>0.002</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.700</td>
<td>0.117</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>1.000</td>
<td>---</td>
</tr>
</tbody>
</table>

Legend:
a: conditional utility function-based procedure with standard gamble  
b: discrete choice conjoint methodology  
c: $n = 139$  
d: Spearman’s rho  
e: significance testing not carried out due to small sample size
**Research Question # 11(a):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

**Research Question # 11(b):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

**Research Question # 11(c):** Is there a significant correlation between the magnitude of preference scores of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?

This set of research questions was addressed using Pearson's product moment correlation coefficient or Pearson's r. In all cases, the Pearson r was above 0.90. In the case of the comparison between the CUF-RS and CUF-SG methodologies, the value of Pearson’s r was significantly greater than zero (familywise p<0.05) for the analysis on corner states only and corner and interior states together. Although the value of Pearson’s r was significantly greater than zero (familywise p<0.05) for the analysis on corner and interior
states together for the comparison between the DCCM and other methodologies, it did not approach statistical significance (on the basis of the Bonferroni correction which was applied) for the analysis on corner states only. The results of the analyses at all three levels are presented in Tables 43 through 45.

Concordances

The fourth aspect on which the different preference measurement methodologies were compared was in terms of the concordance among the preference scores of multiattribute health states obtained using the different methodologies. The two research questions related to this aspect were addressed using the intraclass correlation coefficient or ICC. The results of the data analysis procedures conducted to address these questions will be presented in turn.

Research Question # 12: Is there significant concordance among the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM?
Table 43: Results of Pearson’s r Between CUF-RS’ and CUF-SG\textsuperscript{bc}

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>$r^d$</th>
<th>$t_{\text{obs}}$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.992</td>
<td>19.25</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.997</td>
<td>22.31</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>0.983</td>
<td>5.35</td>
<td>---</td>
</tr>
</tbody>
</table>

Legend:
\(a\): conditional utility function-based procedure with rating scale
\(b\): conditional utility function-based procedure with standard gamble
\(c\): \(n = 139\)
\(d\): Pearson’s product moment correlation coefficient
\(e\): significance testing not carried out due to small sample size
Table 44: Results of Pearson's r Between CUF-RS and DCCM

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>$r^d$</th>
<th>$t_{obs}$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.954</td>
<td>7.79</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.929</td>
<td>4.35</td>
<td>0.012*</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>0.992</td>
<td>7.86</td>
<td>---*</td>
</tr>
</tbody>
</table>

Legend:
- a: conditional utility function-based procedure with rating scale
- b: discrete choice conjoint methodology
- c: n = 139
- d: Pearson's product moment correlation coefficient
- e: determined by linear extrapolation between table entries
- f: significance testing not carried out due to small sample size
Table 45: Results of Pearson's $r$ Between CUF-SC$^\text{a}$ and DCCM$^\text{bc}$

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>$r^d$</th>
<th>$t_{obs}$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.967</td>
<td>9.30</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.937</td>
<td>4.65</td>
<td>0.009$^e$</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>0.951</td>
<td>3.08</td>
<td>---$^f$</td>
</tr>
</tbody>
</table>

Legend:

- $a$: conditional utility function-based procedure with rating scale
- $b$: conditional utility function-based procedure with standard gamble
- $c$: $n = 139$
- $d$: Pearson's product moment correlation coefficient
- $e$: determined by linear extrapolation between table entries
- $f$: significance testing not carried out due to small sample size
This particular research question was addressed using the overall ICC among the three different preference measurement methodologies taken together. The results of this overall ICC for all three levels of the analysis are presented in Table 46. It can be seen that even after adjusting for multiple comparisons with the Bonferroni correction, there is significant concordance among all three preference measurement methodologies investigated in this study. This conclusion holds at both the levels of the analysis for which the statistical significance of the overall ICC was estimated. As noted earlier, statistical significance testing was not carried out on the analysis of the three interior states only on account of the small sample size. The implications of these results will be discussed in Chapter VII.

**Research Question # 13(a):** Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

**Research Question # 13(b):** Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?
Table 46: Intraclass Correlation Coefficient Among All Three Methodologies

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>ICC$^b$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.856</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.892</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>0.685</td>
<td>---$^c$</td>
</tr>
</tbody>
</table>

Legend:
- a: n = 139
- b: intraclass correlation coefficient
- c: significance testing not carried out due to small sample size
Research Question # 13(c): Is there significant concordance between the preference scores of alternative states of health as determined by the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology and the discrete choice conjoint methodology or DCCM?

This set of research questions was addressed using the ICC between each of the two measurement methodologies in question. The results of these analyses are presented in Tables 47 through 49. It can be seen that the value of the ICC is significantly greater than zero (familywise p<0.05) for the comparison between the CUF-RS and CUF-SG and also between the CUF-RS and DCCM at all levels of the analysis. However, for the comparison between the CUF-SG and DCCM, the ICC approaches borderline statistical significance for the analysis on the corner and interior states together and is not significantly greater than zero for analysis on the corner states only.

Functional Relationships

Three specific research questions were investigated in this study, pertaining to the functional relationship between each set of two preference measurement methodologies. The first two of these research questions were concerned with the functional relationship between value and utility functions, while the third was concerned with the functional relationship between two value functions. The results of each question will be presented in turn.
### Table 47: Intraclass Correlation Coefficient Between CUF-RS* and CUF-SG**

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>ICC&lt;sup&gt;d&lt;/sup&gt;</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.951</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.939</td>
<td>0.003</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>0.983</td>
<td>---*</td>
</tr>
</tbody>
</table>

**Legend:**
- a: conditional utility function-based procedure with rating scale
- b: conditional utility function-based procedure with standard gamble
- c: n = 139
- d: intraclass correlation coefficient
- e: significance testing not carried out due to small sample size
Table 48: Intraclass Correlation Coefficient Between CUF-RS and DCCM

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>ICC$^d$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>0.881</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.916</td>
<td>0.005</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>0.667</td>
<td>---$^e$</td>
</tr>
</tbody>
</table>

Legend:
- a: conditional utility function-based procedure with rating scale
- b: discrete choice conjoint methodology
- c: n = 139
- d: intraclass correlation coefficient
- e: significance testing not carried out due to small sample size
Table 49: Intraclass Correlation Coefficient Between CUF-SG and DCCM

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>ICC&lt;sup&gt;d&lt;/sup&gt;</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior</td>
<td>8</td>
<td>0.773</td>
<td>0.007</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>0.825</td>
<td>0.022</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>0.628</td>
<td>---*</td>
</tr>
</tbody>
</table>

Legend:
- a: conditional utility function-based procedure with standard gamble
- b: discrete choice conjoint methodology
- c: n = 139
- d: intraclass correlation coefficient
- e: significance testing not carried out due to small sample size
Research Question # 14(a): What is the nature of the functional relationship between the preference functions for alternative health states obtained using the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Research Question # 14(b): What is the nature of the functional relationship between the preference functions for alternative health states obtained using the discrete choice conjoint methodology or DCCM and the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Research Question # 14(c): What is the nature of the functional relationship between the preference functions for alternative health states obtained using the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology and the discrete choice conjoint methodology or DCCM?

This set of questions was addressed by regression analysis. The results of the analyses conducted in order to address this particular research question are provided in Tables 50 through 52. It can be seen that all models are significant, with the exception of the relationship between the preference scores for corner states only obtained using the CUF-RS and DCCM methodologies. The implications of these results are discussed in the next chapter of this dissertation.
Table 50: Functional Relationship Between CUF-RS' and CUF-SG<sup>he</sup>

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>$x^d$</th>
<th>adj $R^2*$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>1.816</td>
<td>0.979</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>1.759</td>
<td>0.995</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>2.445</td>
<td>0.946</td>
<td>---&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Legend:

a: conditional utility function-based procedure with rating scale  
b: conditional utility function-based procedure with standard gamble  
c: n = 139  
d: parameter to be estimated, according to the model  
\[(1 - \text{CUF-SG}) = (1 - \text{CUF-RS})^*\]
e: adjusted $R^2 = 1 - \frac{(1 - R^2) n}{(n - k - 1)}$  
f: significance testing not carried out due to small sample size
Table 51: Functional Relationship Between CUF-RS and DCCM\textsuperscript{bc}

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>a\textsuperscript{a}</th>
<th>x\textsuperscript{c}</th>
<th>adj R\textsuperscript{2f}</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>-0.265</td>
<td>1.376</td>
<td>0.892</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>-0.144</td>
<td>1.101</td>
<td>0.773</td>
<td>0.022</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>-0.981</td>
<td>2.360</td>
<td>0.922</td>
<td>--\textsuperscript{a}</td>
</tr>
</tbody>
</table>

Legend:

a: conditional utility function-based procedure with rating scale
b: discrete choice conjoint methodology
c: n = 139
d: intercept, according to the model
DCCM = a + x (CUF-RS)
e: parameter to be estimated, according to the model
DCCM = a + x (CUF-RS)
f: adjusted R\textsuperscript{2} = 1 - [(1 - R\textsuperscript{2}) \cdot n / (n - k - 1)]
g: significance testing not carried out due to small sample size
Table 52: Functional Relationship Between CUF-SG and DCCM

<table>
<thead>
<tr>
<th>Stimuli</th>
<th>Health States</th>
<th>$x^d$</th>
<th>adj R$^{2e}$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corner &amp; Interior States</td>
<td>8</td>
<td>1.312</td>
<td>0.932</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Corner States Only</td>
<td>5</td>
<td>1.285</td>
<td>0.970</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Interior States Only</td>
<td>3</td>
<td>1.609</td>
<td>0.376</td>
<td>---</td>
</tr>
</tbody>
</table>

Legend:
a: conditional utility function-based procedure with standard gamble
b: discrete choice conjoint methodology
c: $n = 139$
d: parameter to be estimated, according to the model
\[(1 - \text{CUF-SG}) = (1 - \text{DCCM})^x\]
e: adjusted $R^2 = 1 - [(1 - R^2) n / (n - k - 1)]$
f: significance testing not carried out due to small sample size
Results of the Comparison of Respondent Evaluations

Research Questions 15 through 18 pertained to comparing the respondent evaluations of the three preference measurement methodologies. Research Questions 15 and 16 investigated the psychometric properties of the data collected on respondent evaluations of the three preference measurement methodologies, while Research Questions 17 and 18 involved the actual comparison of respondent evaluations of the three methodologies. This section will present the results of the analyses conducted in order to address each of these four questions. The results of each particular analysis will be presented in turn.

Reliability

Research Question # 15(a): What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?

Research Question # 15(b): What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?
Research Question # 15(c): What is the reliability coefficient of the data collected using the instrument to measure respondent evaluations of the discrete choice conjoint methodology or DCCM?

This set of research questions was addressed by calculating Cronbach's alpha for each of the sets of data on respondent evaluations (i.e., CUF-RS, CUF-SG, and DCCM). The estimates of reliability, as measured by Cronbach's coefficient alpha, for each of the three sets of data are presented in Table 53. This table also provides information regarding what the value of Cronbach's alpha would be if each of the items of the eight-item semantic differential scale were deleted. These results are discussed in Chapter VII.

Exploratory Factor Analyses

An exploratory factor analysis of the data collected on respondent evaluations of each of the three preference measurement methodologies was performed. The results of the factor analysis on the respondent evaluations of each of the three preference measurement methodologies will be presented in turn. Each set of results will begin with presenting the correlation matrix of measured variables, followed by the results of the factor extraction methods (for both, the test of exact fit as well as the test of close fit), and finally present the rotated factor matrix which was interpreted in order to understand the structure of respondent evaluations of the preference measurement methodology in question.
Table 53: Cronbach’s Coefficient Alpha for the Three Sets of Data*

<table>
<thead>
<tr>
<th>Item</th>
<th>Cronbach’s Alpha CUF-RSb</th>
<th>CUF-SCc</th>
<th>DCCMd</th>
</tr>
</thead>
<tbody>
<tr>
<td>All eight items</td>
<td>0.838</td>
<td>0.840</td>
<td>0.887</td>
</tr>
<tr>
<td>Deleting:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interest</td>
<td>0.832</td>
<td>0.839</td>
<td>0.882</td>
</tr>
<tr>
<td>Sensibility</td>
<td>0.823</td>
<td>0.837</td>
<td>0.878</td>
</tr>
<tr>
<td>Difficulty in Doing</td>
<td>0.842</td>
<td>0.855</td>
<td>0.886</td>
</tr>
<tr>
<td>Usefulness</td>
<td>0.823</td>
<td>0.836</td>
<td>0.882</td>
</tr>
<tr>
<td>Length</td>
<td>0.856</td>
<td>0.868</td>
<td>0.884</td>
</tr>
<tr>
<td>Success</td>
<td>0.816</td>
<td>0.845</td>
<td>0.882</td>
</tr>
<tr>
<td>Clarity</td>
<td>0.813</td>
<td>0.835</td>
<td>0.873</td>
</tr>
<tr>
<td>Difficulty in Understanding</td>
<td>0.823</td>
<td>0.843</td>
<td>0.878</td>
</tr>
</tbody>
</table>

Legend:
- a: n = 139
- b: conditional utility function-based procedure with rating scale
- c: conditional utility function-based procedure with standard gamble
- d: discrete choice conjoint methodology
Research Question # 16(a): What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology?

Table 54 presents the correlation matrix of measured variables which was computed using SAS. As noted in the previous chapter, it was this matrix which was used as the input data for the factor analysis. Table 55 presents the results of the factor extraction procedure using maximum likelihood estimation in terms of the test of exact fit. Each row of the table summarizes the results of the model with a given number of extracted factors (as specified in the first column of the table). The results of the test of exact fit indicate that the two-factor solution is the preferred one.

Table 56 presents the results of the factor extraction procedure using maximum likelihood estimation in terms of the test of close fit. As before, each row of the table summarizes the results of the model with a given number of extracted factors. The results of the test of close fit also indicate that the two-factor solution is the preferred one. Therefore, both, the test of exact fit as well as the test of close fit indicate that two factors represent the structure underlying respondent evaluations of the CUF-RS methodology.
Table 54: Correlation Matrix of Variables Measuring Respondent Evaluations of CUF-RS<sup>abc</sup>

<table>
<thead>
<tr>
<th></th>
<th>INT</th>
<th>SEN</th>
<th>DDO</th>
<th>USE</th>
<th>LEN</th>
<th>SUC</th>
<th>CLR</th>
<th>DUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEN</td>
<td>0.595</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDO</td>
<td>0.214</td>
<td>0.178</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USE</td>
<td>0.502</td>
<td>0.594</td>
<td>0.204</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEN</td>
<td>0.196</td>
<td>0.171</td>
<td>0.414</td>
<td>0.239</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUC</td>
<td>0.449</td>
<td>0.587</td>
<td>0.384</td>
<td>0.556</td>
<td>0.258</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>0.448</td>
<td>0.577</td>
<td>0.425</td>
<td>0.535</td>
<td>0.256</td>
<td>0.599</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>DUN</td>
<td>0.339</td>
<td>0.392</td>
<td>0.556</td>
<td>0.453</td>
<td>0.289</td>
<td>0.485</td>
<td>0.567</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139  
b: only portion below diagonal is shown due to symmetry of matrix  
c: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
Table 55: Results of Test of Exact Fit for Respondent Evaluations of CUF-RS

<table>
<thead>
<tr>
<th>Factors</th>
<th>$F_{c}$</th>
<th>$\chi^2_{d}$</th>
<th>df</th>
<th>parm</th>
<th>TLI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.145</td>
<td>433.995</td>
<td>28</td>
<td>8</td>
<td>---</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1</td>
<td>0.524</td>
<td>70.183</td>
<td>20</td>
<td>16</td>
<td>0.83</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.076</td>
<td>10.152</td>
<td>13</td>
<td>23</td>
<td>1.01</td>
<td>0.682</td>
</tr>
</tbody>
</table>

Legend:
a: n = 139  
b: number of factors extracted in the model  
c: sample value of the maximum likelihood discrepancy function  
d: Bartlett adjusted chi-square  
e: degrees of freedom of the model  
f: number of parameters in the model  
g: Tucker-Lewis Index
Table 56: Results of Test of Close Fit for Respondent Evaluations of CUF-R5

<table>
<thead>
<tr>
<th>Factors</th>
<th>$F_\nu$</th>
<th>df</th>
<th>n</th>
<th>RMSEA</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.145</td>
<td>28</td>
<td>8</td>
<td>0.324</td>
<td>0.292, 0.357</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1</td>
<td>0.524</td>
<td>20</td>
<td>16</td>
<td>0.138</td>
<td>0.098, 0.179</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.076</td>
<td>13</td>
<td>23</td>
<td>0.000</td>
<td>0.000, 0.080</td>
<td>0.865</td>
</tr>
</tbody>
</table>

Legend:
- a: n = 139
- b: number of factors extracted in the model
- c: sample value of the maximum likelihood discrepancy function
- d: degrees of freedom of the model
- e: number of parameters in the model
- f: root mean square error of approximation
- g: 95% confidence intervals for RMSEA
As noted in the previous chapter, this two-factor solution was subject to oblique rotation in order to obtain a more meaningful interpretation of the factor loadings. Table 57 presents the final rotated factor loading matrix (using the promax approach to oblique rotation), which was used to interpret the two-factor solution. Although this rotated factor loading matrix does not truly exhibit simple structure, it is still amenable to meaningful interpretation. However, interpretation of this matrix is deferred until Chapter VII.

**Research Question # 16(b):** What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology?

Table 58 presents the correlation matrix of measured variables which was computed using SAS. As noted in the previous chapter, it was this matrix which was used as the input data for the factor analysis. Table 59 presents the results of the factor extraction procedure using maximum likelihood estimation in terms of the test of exact fit. Each row of the table summarizes the results of the model with a given number of extracted factors (as specified in the first column of the table).
### Table 57: Factor Loading Matrix for Respondent Evaluations of CUF-RS^{ab}

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>0.674</td>
<td>-0.015</td>
<td>0.445</td>
</tr>
<tr>
<td>SEN</td>
<td>0.882</td>
<td>-0.116</td>
<td>0.696</td>
</tr>
<tr>
<td>DDO</td>
<td>-0.130</td>
<td>0.925</td>
<td>0.760</td>
</tr>
<tr>
<td>USE</td>
<td>0.744</td>
<td>-0.003</td>
<td>0.552</td>
</tr>
<tr>
<td>LEN</td>
<td>0.054</td>
<td>0.438</td>
<td>0.217</td>
</tr>
<tr>
<td>SUC</td>
<td>0.627</td>
<td>0.220</td>
<td>0.571</td>
</tr>
<tr>
<td>CLR</td>
<td>0.591</td>
<td>0.293</td>
<td>0.598</td>
</tr>
<tr>
<td>DUN</td>
<td>0.305</td>
<td>0.537</td>
<td>0.537</td>
</tr>
</tbody>
</table>

Legend:

- a: n = 139
- b: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
Table 58: Correlation Matrix of Variables Measuring Respondent Evaluations of CUF-SG

<table>
<thead>
<tr>
<th></th>
<th>INT</th>
<th>SEN</th>
<th>DDO</th>
<th>USE</th>
<th>LEN</th>
<th>SUC</th>
<th>CLR</th>
<th>DUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEN</td>
<td>0.663</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDO</td>
<td>0.322</td>
<td>0.334</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USE</td>
<td>0.588</td>
<td>0.633</td>
<td>0.348</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEN</td>
<td>0.284</td>
<td>0.249</td>
<td>0.382</td>
<td>0.249</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUC</td>
<td>0.473</td>
<td>0.600</td>
<td>0.223</td>
<td>0.652</td>
<td>0.271</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>0.563</td>
<td>0.527</td>
<td>0.434</td>
<td>0.515</td>
<td>0.337</td>
<td>0.477</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>DUN</td>
<td>0.397</td>
<td>0.360</td>
<td>0.643</td>
<td>0.430</td>
<td>0.345</td>
<td>0.378</td>
<td>0.591</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: only portion below diagonal is shown due to symmetry of matrix
c: INT = interest; SEN = sensibility; DDO = difficulty in doing;
   USE = usefulness; LEN = length; SUC = success; CLR = clarity;
   DUN = difficulty in understanding
Table 59: Results of Test of Exact Fit for Respondent Evaluations of CUF-SGa

<table>
<thead>
<tr>
<th>Factors&lt;sup&gt;b&lt;/sup&gt;</th>
<th>$F_{x}^{c}$</th>
<th>$\chi^{2d}$</th>
<th>df&lt;sup&gt;e&lt;/sup&gt;</th>
<th>parm&lt;sup&gt;f&lt;/sup&gt;</th>
<th>TLI&lt;sup&gt;g&lt;/sup&gt;</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.528</td>
<td>486.804</td>
<td>28</td>
<td>8</td>
<td>---</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1</td>
<td>0.689</td>
<td>92.258</td>
<td>20</td>
<td>16</td>
<td>0.78</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.177</td>
<td>23.511</td>
<td>13</td>
<td>23</td>
<td>0.95</td>
<td>0.036</td>
</tr>
<tr>
<td>3&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.058</td>
<td>8.054</td>
<td>7</td>
<td>29</td>
<td>0.99</td>
<td>0.328</td>
</tr>
</tbody>
</table>

Legend:
- a: $n = 139$
- b: number of factors extracted in the model
- c: sample value of the maximum likelihood discrepancy function
- d: Bartlett adjusted chi-square
- e: degrees of freedom of the model
- f: number of parameters in the model
- g: Tucker-Lewis Index
- h: obtained using PCAP (see text for explanation)
The results of the test of exact fit indicate that the three-factor solution is the preferred one. It should, however, be noted that the three-factor solution involved the occurrence of one Heywood case, i.e., a solution in which a communality was greater than unity. SAS stopped processing the program on account of this Heywood case. Therefore, the three-factor solution shown above was not obtained using SAS. Instead, the factor analysis program FCAP (Cudeck, 1991) was used for this solution. FCAP is based on the work of Swain (1975), and permits the estimation of parameters in the face of Heywood cases (albeit with an error message). The reason why it was decided to run a three-factor solution in the face of a Heywood case was to view the factor loading matrix of the three-factor solution, so that a better understanding of the potential structure of respondent evaluations of the CUF-SG methodology could be gained. Further, Browne (1968) has noted that although a population communality greater than or equal to unity is not possible, the occurrence of a Heywood case in a sample does not necessarily imply that the factor analysis model does not fit the population from which the sample is drawn, but may be due to sampling fluctuation in the estimate. Still, it should be noted that in his simulation study, Browne (1968) selected the factor matrix obtained on the iteration preceding the occurrence of the Heywood case as the final solution.

Table 60 presents the results of the factor extraction procedure using maximum likelihood estimation in terms of the test of close fit. As before, each row of the table summarizes the results of the model with a given number of extracted factors.
Table 60: Results of Test of Close Fit for Respondent Evaluations of CUF-SG*  

<table>
<thead>
<tr>
<th>Factors</th>
<th>$F_{sc}$</th>
<th>df</th>
<th>parm</th>
<th>RMSEA</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3.528</td>
<td>28</td>
<td>8</td>
<td>0.345</td>
<td>0.313, 0.377</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1</td>
<td>0.689</td>
<td>20</td>
<td>16</td>
<td>0.165</td>
<td>0.126, 0.205</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.177</td>
<td>13</td>
<td>23</td>
<td>0.080</td>
<td>0.000, 0.136</td>
<td>0.146</td>
</tr>
<tr>
<td>3</td>
<td>0.058</td>
<td>7</td>
<td>29</td>
<td>0.032</td>
<td>0.000, 0.125</td>
<td>0.554</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139  
b: number of factors extracted in the model  
c: sample value of the maximum likelihood discrepancy function  
d: degrees of freedom of the model  
e: number of parameters in the model  
f: root mean square error of approximation  
g: 95% confidence intervals for RMSEA  
h: obtained using FCAP (see text for explanation)
In contrast to the results of the test of exact fit, the results of the test of close fit indicate that the two-factor solution is the preferred one (although the RMSEA values of the three-factor solution are lower than those of the two-factor solution, both solutions fail to reject the null hypothesis of close fit in the population; therefore, the principle of parsimony in modeling would suggest that the two-factor solution be preferred). The final decision on how many factors to extract was based on the interpretability of the final rotated factor loading matrices of the two- and three-factor solutions, which are presented in Tables 61 and 62. Although these rotated factor loading matrices do not truly exhibit simple structure, they are still amenable to meaningful interpretation. More importantly, they serve the purpose of deciding how many factors to extract for the final solution.

Both these solutions are based on an oblique rotation of the extracted factors using the direct quartimin approach to oblique rotation. The reason why this particular approach was used for rotation is that it is the only one available in FCAP. Still, it is reassuring to note that the factor loadings of the two-factor solutions using the promax approach in SAS and the direct quartimin approach in FCAP were very similar, i.e., there was an absolute difference of only about 0.02 units in the magnitude of the factor loadings obtained using both approaches, and there were no differences in the signs of any of the factor loadings obtained using the two approaches.

Although interpretation of these factor loading matrices and the rationale for the decision of how many factors to extract for the final solution is deferred until the next chapter of this dissertation, at this point it needs to be noted that the two-factor solution was chosen over the three-factor one.
Table 61: Factor Loading Matrix for Respondent Evaluations of CUF-SG for the Two-factor Solution using Direct Quartimin$^{ab}$

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>0.729</td>
<td>0.047</td>
</tr>
<tr>
<td>SEN</td>
<td>0.873</td>
<td>-0.073</td>
</tr>
<tr>
<td>DDO</td>
<td>-0.052</td>
<td>0.752</td>
</tr>
<tr>
<td>USE</td>
<td>0.763</td>
<td>0.049</td>
</tr>
<tr>
<td>LEN</td>
<td>0.123</td>
<td>0.356</td>
</tr>
<tr>
<td>SUC</td>
<td>0.737</td>
<td>-0.012</td>
</tr>
<tr>
<td>CLR</td>
<td>0.414</td>
<td>0.421</td>
</tr>
<tr>
<td>DUN</td>
<td>-0.050</td>
<td>0.917</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
Table 62: Factor Loading Matrix for Respondent Evaluations of CUF-SG for the Three-factor Solution$^{ab}$

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>0.028</td>
<td>0.859</td>
<td>-0.085</td>
</tr>
<tr>
<td>SEN</td>
<td>-0.036</td>
<td>0.760</td>
<td>0.131</td>
</tr>
<tr>
<td>DDO</td>
<td>0.740</td>
<td>0.036</td>
<td>-0.104</td>
</tr>
<tr>
<td>USE</td>
<td>0.106</td>
<td>0.495</td>
<td>0.294</td>
</tr>
<tr>
<td>LEN</td>
<td>0.347</td>
<td>0.077</td>
<td>0.080</td>
</tr>
<tr>
<td>SUC</td>
<td>0.020</td>
<td>0.035</td>
<td>0.971</td>
</tr>
<tr>
<td>CLR</td>
<td>0.430</td>
<td>0.371</td>
<td>0.064</td>
</tr>
<tr>
<td>DUN</td>
<td>0.937</td>
<td>-0.117</td>
<td>0.068</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139  
b: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
For the purpose of consistency, the final rotated two-factor solution using promax rotation is presented in Table 63. There are two reasons why this particular solution is presented: (a) SAS is more commonly used as a software package for factor analysis compared to FCAP; therefore, it is useful to provide the actual factor loading matrix obtained using SAS for the purpose of comparisons with future studies in this area, and (b) the factor analysis solution for the respondent evaluations of the CUF-RS methodology was obtained using a promax rotation; therefore, in order to be consistent with the previous analysis and to permit comparisons across respondent evaluations of the different preference measurement methodologies investigated in this study, it was decided to present the results of the two-factor solution using promax rotation. It is noteworthy that the two-factor solutions using the promax and direct quartimin approaches to oblique rotation give very similar results in terms of the magnitudes and signs of the observed factor loadings.

Research Question # 16(c): What is the underlying factor structure of the data collected using the instrument designed to measure respondent evaluations of the discrete choice conjoint methodology or DCCM?

Table 64 presents the correlation matrix of measured variables which was computed using SAS. As noted in the previous chapter, it was this matrix which was used as the input data for the factor analysis.
Table 63: Factor Loading Matrix for Respondent Evaluations of CUF-SG for the Two-factor Solution using Promax\textsuperscript{ab}

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>0.716</td>
<td>0.072</td>
<td>0.575</td>
</tr>
<tr>
<td>SEN</td>
<td>0.852</td>
<td>-0.040</td>
<td>0.691</td>
</tr>
<tr>
<td>DDO</td>
<td>-0.018</td>
<td>0.732</td>
<td>0.522</td>
</tr>
<tr>
<td>USE</td>
<td>0.750</td>
<td>0.075</td>
<td>0.630</td>
</tr>
<tr>
<td>LEN</td>
<td>0.135</td>
<td>0.352</td>
<td>0.194</td>
</tr>
<tr>
<td>SUC</td>
<td>0.722</td>
<td>0.015</td>
<td>0.533</td>
</tr>
<tr>
<td>CLR</td>
<td>0.424</td>
<td>0.426</td>
<td>0.558</td>
</tr>
<tr>
<td>DUN</td>
<td>-0.009</td>
<td>0.893</td>
<td>0.788</td>
</tr>
</tbody>
</table>

Legend:
\textsuperscript{a}: n = 139
\textsuperscript{b}: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
Table 64: Correlation Matrix of Variables Measuring Respondent Evaluations of DCCM

<table>
<thead>
<tr>
<th></th>
<th>INT</th>
<th>SEN</th>
<th>DDO</th>
<th>USE</th>
<th>LEN</th>
<th>SUC</th>
<th>CLR</th>
<th>DUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEN</td>
<td>0.709</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDO</td>
<td>0.357</td>
<td>0.394</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USE</td>
<td>0.437</td>
<td>0.546</td>
<td>0.368</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LEN</td>
<td>0.507</td>
<td>0.516</td>
<td>0.501</td>
<td>0.477</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUC</td>
<td>0.476</td>
<td>0.621</td>
<td>0.377</td>
<td>0.682</td>
<td>0.450</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLR</td>
<td>0.585</td>
<td>0.559</td>
<td>0.591</td>
<td>0.485</td>
<td>0.529</td>
<td>0.486</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>DUN</td>
<td>0.476</td>
<td>0.396</td>
<td>0.733</td>
<td>0.534</td>
<td>0.474</td>
<td>0.469</td>
<td>0.647</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Legend:
- a: n = 139
- b: only portion below diagonal is shown due to symmetry of matrix
- c: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
Table 65 presents the results of the factor extraction procedure using maximum likelihood estimation in terms of the test of exact fit. Each row of the table summarizes the results of the model with a given number of extracted factors (as specified in the first column of the table). These results indicate that the four-factor solution is the preferred one. It should, however, be noted that both, the three- as well as the four-factor solutions involved the occurrence of one Heywood case, i.e., a solution in which a communality was greater than unity. As noted earlier, the factor analysis solutions with such cases were obtained using FCAP instead of SAS.

Table 66 presents the results of the factor extraction procedure using maximum likelihood estimation in terms of the test of close fit. As before, each row of the table summarizes the results of the model with a given number of extracted factors. In contrast to the results of the test of exact fit, these results indicate that the three-factor solution is the preferred one (although the RMSEA values of the four-factor solution are lower than those of the three-factor solution, both solutions fail to reject the null hypothesis of close fit in the population; therefore, the principle of parsimony in modeling would suggest that the three-factor solution be preferred). The final decision on how many factors to extract was based on the interpretability of the final rotated factor loading matrices of the two-, three-, and four-factor solutions, which are presented in Tables 67 through 69.
Table 65: Results of Test of Exact Fit for Respondent Evaluations of DCCM

<table>
<thead>
<tr>
<th>Factors&lt;sup&gt;b&lt;/sup&gt;</th>
<th>$F_{c}$</th>
<th>$\chi^2$</th>
<th>df&lt;sup&gt;e&lt;/sup&gt;</th>
<th>parm&lt;sup&gt;f&lt;/sup&gt;</th>
<th>TLI&lt;sup&gt;g&lt;/sup&gt;</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4.447</td>
<td>613.677</td>
<td>28</td>
<td>8</td>
<td>---</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1</td>
<td>0.993</td>
<td>132.880</td>
<td>20</td>
<td>16</td>
<td>0.73</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.382</td>
<td>50.826</td>
<td>13</td>
<td>23</td>
<td>0.86</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>3&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.121</td>
<td>16.721</td>
<td>7</td>
<td>29</td>
<td>0.93</td>
<td>0.019</td>
</tr>
<tr>
<td>4&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.016</td>
<td>2.268</td>
<td>2</td>
<td>34</td>
<td>0.99</td>
<td>0.459</td>
</tr>
</tbody>
</table>

Legend:
- a: $n = 139$
- b: number of factors extracted in the model
- c: sample value of the maximum likelihood discrepancy function
- d: Bartlett adjusted chi-square
- e: degrees of freedom of the model
- f: number of parameters in the model
- g: Tucker-Lewis Index
- h: obtained using FCAP (see text for explanation)
Table 66: Results of Test of Close Fit for Respondent Evaluations of DCCM

<table>
<thead>
<tr>
<th>Factors</th>
<th>$F_s$</th>
<th>df</th>
<th>parm</th>
<th>RMSEA</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4.447</td>
<td>28</td>
<td>8</td>
<td>0.389</td>
<td>0.358, 0.422</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>1</td>
<td>0.993</td>
<td>20</td>
<td>16</td>
<td>0.206</td>
<td>0.168, 0.245</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>2</td>
<td>0.382</td>
<td>13</td>
<td>23</td>
<td>0.149</td>
<td>0.100, 0.200</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>3$^h$</td>
<td>0.121</td>
<td>7</td>
<td>29</td>
<td>0.100</td>
<td>0.018, 0.174</td>
<td>0.082</td>
</tr>
<tr>
<td>4$^h$</td>
<td>0.016</td>
<td>2</td>
<td>34</td>
<td>0.031</td>
<td>0.000, 0.076</td>
<td>0.439</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: number of factors extracted in the model
c: sample value of the maximum likelihood discrepancy function
d: degrees of freedom of the model
e: number of parameters in the model
f: root mean square error of approximation
g: 95% confidence intervals for RMSEA
h: obtained using FCAP (see text for explanation)
Table 67: Factor Loading Matrix for Respondent Evaluations of DCCM for the Two-factor Solution using Direct Quartimin$^{ab}$

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>0.754</td>
<td>0.029</td>
</tr>
<tr>
<td>SEN</td>
<td>1.025</td>
<td>-0.202</td>
</tr>
<tr>
<td>DDO</td>
<td>0.043</td>
<td>0.757</td>
</tr>
<tr>
<td>USE</td>
<td>0.508</td>
<td>0.241</td>
</tr>
<tr>
<td>LEN</td>
<td>0.470</td>
<td>0.248</td>
</tr>
<tr>
<td>SUC</td>
<td>0.642</td>
<td>0.105</td>
</tr>
<tr>
<td>CLR</td>
<td>0.409</td>
<td>0.451</td>
</tr>
<tr>
<td>DUN</td>
<td>-0.014</td>
<td>0.941</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139  
b: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
Table 68: Factor Loading Matrix for Respondent Evaluations of DCCM for the Three-factor Solution$^{ab}$

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>0.130</td>
<td>0.725</td>
<td>-0.031</td>
</tr>
<tr>
<td>SEN</td>
<td>-0.096</td>
<td>0.924</td>
<td>0.082</td>
</tr>
<tr>
<td>DDO</td>
<td>0.840</td>
<td>0.025</td>
<td>-0.071</td>
</tr>
<tr>
<td>USE</td>
<td>0.021</td>
<td>-0.037</td>
<td>1.010</td>
</tr>
<tr>
<td>LEN</td>
<td>0.305</td>
<td>0.337</td>
<td>0.136</td>
</tr>
<tr>
<td>SUC</td>
<td>0.058</td>
<td>0.351</td>
<td>0.458</td>
</tr>
<tr>
<td>CLR</td>
<td>0.525</td>
<td>0.333</td>
<td>0.034</td>
</tr>
<tr>
<td>DUN</td>
<td>0.884</td>
<td>-0.101</td>
<td>0.143</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
Table 69: Factor Loading Matrix for Respondent Evaluations of DCCM for the Four-factor Solution*\textsuperscript{ab}

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>-0.060</td>
<td>0.964</td>
<td>-0.055</td>
<td>0.085</td>
</tr>
<tr>
<td>SEN</td>
<td>0.341</td>
<td>0.552</td>
<td>0.134</td>
<td>-0.184</td>
</tr>
<tr>
<td>DDO</td>
<td>-0.052</td>
<td>-0.062</td>
<td>0.955</td>
<td>0.089</td>
</tr>
<tr>
<td>USE</td>
<td>0.834</td>
<td>-0.033</td>
<td>-0.051</td>
<td>0.116</td>
</tr>
<tr>
<td>LEN</td>
<td>0.214</td>
<td>0.268</td>
<td>0.338</td>
<td>-0.035</td>
</tr>
<tr>
<td>SUC</td>
<td>0.817</td>
<td>0.021</td>
<td>0.034</td>
<td>-0.040</td>
</tr>
<tr>
<td>CLR</td>
<td>0.135</td>
<td>0.367</td>
<td>0.305</td>
<td>0.175</td>
</tr>
<tr>
<td>DUN</td>
<td>0.177</td>
<td>0.117</td>
<td>0.197</td>
<td>0.747</td>
</tr>
</tbody>
</table>

Legend:
\textsuperscript{a}: n = 139  
\textsuperscript{b}: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding  
\textsuperscript{c}: may not equal sum of squared factor loadings due to rounding error
It was decided to include a two-factor solution in order to compare the results of such a solution for respondent evaluations of the DCCM methodology to the results of similar solutions for respondent evaluations of the other preference measurement methodologies investigated in this study. Specifically, the interest was in determining whether extracting any more factors than those shown to be quite meaningful for the other preference measurement methodologies investigated in this study provided a better solution in terms of interpretability. As explained earlier, these solutions are based on an oblique rotation of the extracted factors using the direct quartimin approach to oblique rotation.

The interpretation of these factor loading matrices is deferred until Chapter VII. At this point, however, it needs to be noted that the two-factor solution was selected over the other solutions (the rationale behind this decision is explained in the next chapter). For the purpose of consistency (as explained earlier), the final rotated two-factor solution using promax rotation is presented in Table 70.

Comparing Respondent Evaluations

Having established the psychometric properties of the data collected on respondent evaluations of the three preference measurement methodologies investigated in this study, the issue of comparing the three methodologies on the basis of respondent evaluations can be addressed.
Table 70: Factor Loading Matrix for Respondent Evaluations of DCCM for the Two-factor Solution using Promax$^{ab}$

<table>
<thead>
<tr>
<th>Variables</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Communality</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>0.715</td>
<td>0.095</td>
<td>0.597</td>
</tr>
<tr>
<td>SEN</td>
<td>0.970</td>
<td>-0.110</td>
<td>0.833</td>
</tr>
<tr>
<td>DDO</td>
<td>0.047</td>
<td>0.757</td>
<td>0.616</td>
</tr>
<tr>
<td>USE</td>
<td>0.484</td>
<td>0.284</td>
<td>0.468</td>
</tr>
<tr>
<td>LEN</td>
<td>0.448</td>
<td>0.288</td>
<td>0.428</td>
</tr>
<tr>
<td>SUC</td>
<td>0.610</td>
<td>0.161</td>
<td>0.507</td>
</tr>
<tr>
<td>CLR</td>
<td>0.392</td>
<td>0.486</td>
<td>0.600</td>
</tr>
<tr>
<td>DUN</td>
<td>-0.004</td>
<td>0.935</td>
<td>0.869</td>
</tr>
</tbody>
</table>

Legend:

a: $n = 139$

b: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding
Research Question # 17: Are there any significant differences among the respondent evaluations for the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM in terms of the following variables:

(a) interest
(b) sensibility
(c) difficulty in doing
(d) usefulness
(e) length
(f) success
(g) clarity
(h) difficulty in understanding
(i) confidence

This set of research questions was addressed using a series of repeated measures one-way analyses of variance, with the Scheffe test for multiple comparisons. Table 71 presents the descriptive statistics -- i.e., mean (standard deviation) -- for the respondent evaluations of the three preference measurement methodologies investigated in this study. For all entries in the table (except for Total Time) higher values indicate more positive evaluations. In this sense, the items which were negative in the original
instrument have been recoded to facilitate easy interpretation and comparison with other items.

The results of the series of repeated measures ANOVAs on each of the variables used to measure respondent evaluations of the three preference measurement methodologies are summarized in Table 72. The variables for which there were significant differences in respondent judgmental ratings were difficulty in doing, usefulness, length, success, difficulty in understanding, and confidence. Further, there were also significant differences in the total time taken by respondents to complete the tasks required of each preference measurement methodology.

On the other hand, there were no significant differences in respondent judgmental ratings in terms of the interest, sensibility, and clarity of the preference measurement methodologies. The interpretation of the specific pairwise comparisons and the discussion of the results of these one-way repeated measures analyses of variance is deferred until the next chapter.

**Research Question #18**: Is there a difference between the preferential rank ordering of the three health status preference measurement methodologies (the conditional utility function-based procedure with the rating scale scaling method or CUF-RS methodology, the conditional utility function-based procedure with the standard gamble scaling method or CUF-SG methodology, and the discrete choice conjoint methodology or DCCM) as determined by responses to questions using direct comparison and judgmental ratings as the response modes in terms of the following variables:
Table 71: Respondent Evaluations of the Three Preference Measurement Methodologies Using Judgmental Ratings

<table>
<thead>
<tr>
<th>Variable</th>
<th>CUF-RS</th>
<th>CUF-SG</th>
<th>DCCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Time</td>
<td>15.02 (4.65)</td>
<td>16.18 (5.44)</td>
<td>8.17 (3.46)</td>
</tr>
<tr>
<td>INT</td>
<td>6.03 (1.17)</td>
<td>6.03 (0.93)</td>
<td>6.22 (1.18)</td>
</tr>
<tr>
<td>SEN</td>
<td>5.99 (1.20)</td>
<td>5.99 (1.06)</td>
<td>6.15 (1.06)</td>
</tr>
<tr>
<td>DDO</td>
<td>5.70 (1.60)</td>
<td>4.67 (1.80)</td>
<td>5.85 (1.62)</td>
</tr>
<tr>
<td>USE</td>
<td>5.94 (1.06)</td>
<td>5.99 (1.02)</td>
<td>6.21 (0.98)</td>
</tr>
<tr>
<td>LEN</td>
<td>4.63 (1.37)</td>
<td>4.11 (1.52)</td>
<td>5.65 (1.37)</td>
</tr>
<tr>
<td>SUC</td>
<td>5.86 (1.05)</td>
<td>5.91 (0.99)</td>
<td>6.10 (0.94)</td>
</tr>
<tr>
<td>CLR</td>
<td>6.00 (1.38)</td>
<td>5.99 (1.13)</td>
<td>6.18 (1.32)</td>
</tr>
<tr>
<td>DUN</td>
<td>5.78 (1.51)</td>
<td>5.35 (1.60)</td>
<td>6.16 (1.37)</td>
</tr>
<tr>
<td>Confidence</td>
<td>4.22 (0.80)</td>
<td>4.12 (0.72)</td>
<td>4.36 (0.77)</td>
</tr>
</tbody>
</table>
Table 71 (continued)

Legend:

a: n = 139
b: entries in the table are mean (standard deviation)
c: conditional utility function-based procedure with rating scale
d: conditional utility function-based procedure with standard gamble
e: discrete choice conjoint methodology
f: in minutes
g: interest, measured on a 7-point scale, where
   1 = extremely boring and 7 = extremely interesting
h: sensibility, measured on a 7-point scale, where
   1 = extremely senseless and 7 = extremely sensible
i: difficulty in doing, measured on a 7-point scale, where
   1 = extremely difficult to do and 7 = extremely easy to do
j: usefulness, measured on a 7-point scale, where
   1 = extremely useless and 7 = extremely useful
k: length, measured on a 7-point scale, where
   1 = extremely long and 7 = extremely short
l: success, measured on a 7-point scale, where
   1 = extremely unsuccessful and 7 = extremely successful
m: clarity, measured on a 7-point scale, where
   1 = extremely vague and 7 = extremely clear
n: difficulty in understanding, measured on a 7-point scale, where
   1 = extremely difficult to understand and
   7 = extremely easy to understand
o: confidence, measured on a 5-point scale,
   where 1 = very unconfident and 5 = very confident
Table 72: Results of Repeated Measures ANOVAs on Respondent Evaluations of the Three Preference Measurement Methodologies

<table>
<thead>
<tr>
<th>Variable</th>
<th>Omnibus F</th>
<th>CUF-RS vs CUF-SG</th>
<th>CUF-RS vs DCCM</th>
<th>CUF-SG vs DCCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Time</td>
<td>174.08&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>INT</td>
<td>2.19</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>SEN</td>
<td>1.84</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>DDO</td>
<td>36.26&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>No&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>USE</td>
<td>6.07&lt;sup&gt;a&lt;/sup&gt;</td>
<td>No&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>No&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>LEN</td>
<td>86.12&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>SUC</td>
<td>5.16&lt;sup&gt;a&lt;/sup&gt;</td>
<td>No&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>No&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>CLR</td>
<td>1.44</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
<td>---&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>DUN</td>
<td>19.98&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Confidence</td>
<td>6.33&lt;sup&gt;a&lt;/sup&gt;</td>
<td>No&lt;sup&gt;1&lt;/sup&gt;</td>
<td>No&lt;sup&gt;1&lt;/sup&gt;</td>
<td>Yes&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
Table 72 (continued)

Legend:
- **a**: n = 139
- **b**: INT = interest; SEN = sensibility; DDO = difficulty in doing;
  USE = usefulness; LEN = length; SUC = success; CLR = clarity;
  DUN = difficulty in understanding
- **c**: evaluated using a Scheffe correction for multiple comparisons
- **d**: conditional utility function-based procedure with rating scale
- **e**: conditional utility function-based procedure with standard gamble
- **f**: discrete choice conjoint methodology
- **g**: p<0.05
- **h**: pairwise comparison was significant (familywise p<0.05)
- **i**: pairwise comparison was not significant
- **j**: pairwise comparisons not conducted due to nonsignificant omnibus F-test
Simple frequency counts were calculated for each of the variables in this comparative evaluation as well as for the overall evaluation. These frequencies -- numerical as well as in the form of a percentage -- are presented in Table 73. As a necessary prerequisite to addressing Research Question # 18, the preferential rank ordering of the methodologies using the direct comparison and judgmental rating response modes was prepared. On the basis of these preferential rank orderings, Kendall's $\tau$ was computed to indicate the agreement in preferential rank ordering between the direct comparison and judgmental rating response modes for each of the variables of interest. As explained in the previous chapter, on account of the small sample size for this comparison, inferential statistics were not computed. The preferential rank ordering by each response mode and the computed value of Kendall's $\tau$ for each of the variables of interest are presented in Table 74. The interpretation of these results is deferred until Chapter VII.
Table 73: Frequency Counts for Comparative Evaluations of the Three Preference Measurement Methodologies

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Most</th>
<th>Middle</th>
<th>Least</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>(a) for interest</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>73</td>
<td>52.52</td>
<td>41</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>38</td>
<td>27.34</td>
<td>47</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>28</td>
<td>20.14</td>
<td>51</td>
</tr>
<tr>
<td>(b) for sensibility</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>73</td>
<td>52.52</td>
<td>44</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>36</td>
<td>25.90</td>
<td>42</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>30</td>
<td>21.58</td>
<td>53</td>
</tr>
<tr>
<td>(c) for difficulty in doing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(variable is reverse coded, so that most = easiest to do)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>85</td>
<td>61.15</td>
<td>38</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>23</td>
<td>16.55</td>
<td>29</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>31</td>
<td>22.30</td>
<td>72</td>
</tr>
<tr>
<td>(d) for usefulness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>74</td>
<td>53.24</td>
<td>47</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>33</td>
<td>23.74</td>
<td>41</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>32</td>
<td>23.02</td>
<td>51</td>
</tr>
<tr>
<td>(e) for length</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(variable is reverse coded, so that most = shortest)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>117</td>
<td>84.17</td>
<td>14</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>10</td>
<td>7.20</td>
<td>40</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>12</td>
<td>8.63</td>
<td>85</td>
</tr>
</tbody>
</table>
Table 73 (continued)

<table>
<thead>
<tr>
<th></th>
<th>DCCM&lt;sup&gt;a&lt;/sup&gt;</th>
<th>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</th>
<th>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f) for success</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;a&lt;/sup&gt;</td>
<td>72</td>
<td>51.80</td>
<td>45</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>40</td>
<td>28.78</td>
<td>35</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>27</td>
<td>19.42</td>
<td>59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DCCM&lt;sup&gt;b&lt;/sup&gt;</th>
<th>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</th>
<th>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>(g) for clarity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;b&lt;/sup&gt;</td>
<td>70</td>
<td>50.36</td>
<td>47</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>38</td>
<td>27.34</td>
<td>34</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>31</td>
<td>22.30</td>
<td>58</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DCCM&lt;sup&gt;d&lt;/sup&gt;</th>
<th>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</th>
<th>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>(h) for difficulty in understanding</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(variable is reverse coded, so that most = easiest to understand)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;d&lt;/sup&gt;</td>
<td>88</td>
<td>63.31</td>
<td>33</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>28</td>
<td>20.14</td>
<td>32</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>23</td>
<td>16.55</td>
<td>74</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DCCM&lt;sup&gt;d&lt;/sup&gt;</th>
<th>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</th>
<th>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) for confidence</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;d&lt;/sup&gt;</td>
<td>80</td>
<td>57.56</td>
<td>41</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>33</td>
<td>23.74</td>
<td>42</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>26</td>
<td>18.70</td>
<td>56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DCCM&lt;sup&gt;d&lt;/sup&gt;</th>
<th>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</th>
<th>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>(j) for the overall evaluation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(in this regard, most = best overall)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCCM&lt;sup&gt;d&lt;/sup&gt;</td>
<td>78</td>
<td>56.12</td>
<td>41</td>
</tr>
<tr>
<td>CUF-SG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>36</td>
<td>25.90</td>
<td>39</td>
</tr>
<tr>
<td>CUF-RS&lt;sup&gt;d&lt;/sup&gt;</td>
<td>25</td>
<td>17.98</td>
<td>59</td>
</tr>
</tbody>
</table>

Legend:

a: n = 139
b: discrete choice conjoint methodology
c: conditional utility function-based procedure with rating scale
d: conditional utility function-based procedure with standard gamble
### Table 74: Rank Ordering and Kendall’s $\tau$ Between Direct Comparison and Judgmental Rating Response Modes for Respondent Evaluations$^{ab}$

<table>
<thead>
<tr>
<th>Variable</th>
<th>Direct Comparison Order$^{cd}$</th>
<th>Judgmental Rating Order$^{def}$</th>
<th>Kendall’s $\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>INT</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>DCCM/CUF-SG=CUF-RS$^b$</td>
<td>0.817</td>
</tr>
<tr>
<td>SEN</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>DCCM/CUF-SG=CUF-RS$^b$</td>
<td>0.817</td>
</tr>
<tr>
<td>DDO</td>
<td>DCCM/CUF-RS/CUF-SG</td>
<td>DCCM/CUF-RS/CUF-SG</td>
<td>1.000</td>
</tr>
<tr>
<td>USE</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>1.000</td>
</tr>
<tr>
<td>LEN</td>
<td>DCCM/CUF-RS/CUF-SG</td>
<td>DCCM/CUF-RS/CUF-SG</td>
<td>1.000</td>
</tr>
<tr>
<td>SUC</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>1.000</td>
</tr>
<tr>
<td>CLR</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>DCCM/CUF-RS/CUF-SG</td>
<td>0.333</td>
</tr>
<tr>
<td>DUN</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>DCCM/CUF-RS/CUF-SG</td>
<td>0.333</td>
</tr>
<tr>
<td>CON</td>
<td>DCCM/CUF-SG/CUF-RS</td>
<td>DCCM/CUF-RS/CUF-SG</td>
<td>0.333</td>
</tr>
</tbody>
</table>

Legend:

- **a**: $n = 139$
- **b**: INT = interest; SEN = sensibility; DDO = difficulty in doing; USE = usefulness; LEN = length; SUC = success; CLR = clarity; DUN = difficulty in understanding; CON = confidence
- **c**: inferred from Table 73; entered in table as best/middle/worst
- **d**: inferred from Table 71; entered in table as best/middle/worst
- **e**: conditional utility function-based procedure with rating scale
- **f**: conditional utility function-based procedure with standard gamble
- **g**: discrete choice conjoint methodology
- **h**: tie between CUF-RS and CUF-SG
This chapter has presented the results of the analyses which were conducted in order to address each of the research questions investigated in this dissertation. The emphasis in this chapter has been more on presenting the results, as opposed to discussing them. This is because the issue of the discussion of results serves as the point of departure for the next chapter, which not only summarizes and discusses the study findings, but also places the study results in the broad context of health status measurement and the outcomes movement in health care.
HEALTH STATUS PREFERENCE MEASUREMENT:
A COMPARISON OF DISCRETE CHOICE CONJOINT AND CONDITIONAL UTILITY MODELING

Volume III

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Zafar Mohomed Anwar Hakim, B.Pharm., M.S.

The Ohio State University

1995

Dissertation Committee:  Approved by

Dev S. Pathak
Suzan Kucukarslan
Jon C. Schommer

Adviser
College of Pharmacy
CHAPTER VII

DISCUSSION AND CONCLUSIONS

Chapter Overview

This final chapter of this dissertation discusses the results of this study in the broader context of health status measurement and the outcomes movement in health care. The chapter begins with a summary of the results of this study, with respect to each of the research questions which was investigated. The results of each set of research questions (i.e., modeling the data, comparing preference scores, and comparing respondent evaluations) are then discussed in terms of how they relate to the nature of the problem investigated in this dissertation. This section is followed by a discussion of the implications of this study, for not only health status preference measurement but also health status measurement and the outcomes movement in health care in general. The next section of this chapter discusses some of the limitations of this study. Finally, some potential avenues for future research in this area are outlined.
Summary of Results

There were 18 specific research questions which were investigated in this dissertation. This section summarizes the results of the analytical procedures conducted in order to address each of these questions. The summary is organized according to the three categories initially defined in Chapter II of this dissertation -- modeling the data, comparing preference scores, and comparing respondent evaluations. Each category is addressed in turn.

Modeling the Data

The first six research questions investigated in this study pertained to modeling the data obtained using the preference measurement methodologies. Health status preferences were modeled using three different measurement methodologies: (a) the conditional utility function-based procedure with a rating scale scaling method or CUF-RS, (b) the conditional utility function-based procedure with a standard gamble scaling method or CUF-SG, and (c) the discrete choice conjoint methodology or DCCM. For each of the methodologies, two aspects of the modeling procedure were of interest in this study -- the nature of the model form of the health status preference function as determined by each methodology (Research Questions # 1, 3, and 5) and the predictive ability of each methodology (Research Questions # 2, 4, and 6).

Regarding the nature of the model form, all three methodologies supported the existence of a multiplicative model form for the health status (as operationalized by the EuroQol) preference function of the respondents
of this study. As explained in the previous chapter, this conclusion was more tentative in the case of health status preference function as determined by the CUF-SG methodology as opposed to the other methodologies.

Regarding the predictive ability of the three methodologies, both the CUF-RS and the CUF-SG methodologies had mean absolute residuals of about 5%, with the CUF-SG methodology being slightly more accurate in terms of predictive ability compared to the CUF-RS methodology. The predictive ability of the DCCM was more variable across choice sets, varying between a little under 2% to about 9% in the two holdout choice sets used in this study.

Comparing Preference Scores

Research Questions 7 through 14 pertained to comparing the preference scores obtained as a result of the different measurement methodologies. Specifically, the preference scores were subject to five different bases of comparison -- preferential ordering (Research Questions # 7 and 8), magnitude of preference scores (Research Question # 9), correlations (Research Questions # 10 and 11), concordances (Research Questions # 12 and 13), and functional relationships (Research Question # 14). The results of each of these bases of comparison will be summarized in this section. All analytical procedures were conducted at three levels -- for all eight multiattribute health states, for the five corner states only, and for the three interior states only. As noted in Chapter V, it was decided to conduct all analytical procedures at these three levels in order to investigate
the effects of structural differences between the corner and interior states on the preferences for these health states.

With respect to Research Questions #7 and 8, two types of statistical measures were used to indicate the extent of agreement in preferential ordering by methodologies -- Kendall's $W$ and Kendall's $\tau$. The analysis at the level of all eight multiattribute health states revealed that there was significant agreement in rank ordering of the health states among all three methodologies taken together (as indicated by Kendall’s $W$). This same conclusion held with respect to the comparison between each pair of methodologies (as indicated by Kendall’s $\tau$). The analysis at the level of the five corner states revealed that there was significant agreement in rank ordering of the health states among all methodologies taken together (as indicated by Kendall’s $W$). However, with respect to the comparison between each pair of methodologies, it was found that there was significant agreement in rank ordering of health states only between the CUF-RS and CUF-SG methodologies (as indicated by Kendall’s $\tau$). Although no significance testing was performed on the analysis at the level of the three interior states, it should be noted that there was perfect agreement among all three methodologies in rank ordering of health states (as indicated by both Kendall’s $W$ and Kendall’s $\tau$).

With respect to Research Question #9, a one-way repeated measures analysis of variance was used to determine whether there were significant differences among the magnitude of preference scores obtained using the three methodologies. The omnibus F-test was significant at the 0.05 level for the analysis at the level of all eight multiattribute health states; pairwise
comparisons using the Scheffe test revealed that there were significant differences (familywise p<0.05) between the preference scores obtained using the CUF-RS and CUF-SG and also between the CUF-SG and the DCCM methodologies, but not between the CUF-RS and DCCM methodologies. The analysis on the five corner states yielded slightly different results. Pairwise comparisons using a Scheffe test (following a significant omnibus F-test) revealed that there were significant differences (familywise p<0.05) between the preference scores obtained using the CUF-RS and CUF-SG as well as between the CUF-RS and DCCM methodologies, but not between the CUF-SG and DCCM methodologies. The omnibus F-test for the analysis at the level of the three interior states was not significant at the 0.05 level.

With respect to Research Questions # 10 and 11, two types of statistical measures were used to indicate the extent of correlation between the preference scores obtained using the three preference measurement methodologies -- Spearman’s ρ and Pearson’s r. In both cases, the results of all three levels of the analysis were very similar to those using Kendall’s τ -- i.e., there was significant correlation between each set of measurement methodologies taken two at a time for the analysis at the level of all eight multiattribute health states, but there was significant correlation between only the CUF-RS and CUF-SG methodologies for the analysis at the level of the five corner states only; further, like the results using Kendall’s τ, there was perfect rank-order correlation between each set of measurement methodologies taken two at a time for the analysis at the level of the three interior states only.
With respect to Research Questions # 12 and 13, the intraclass correlation coefficient or ICC was used to determine the extent of concordance among the preference scores obtained using the three preference measurement methodologies. The analysis at the level of all eight multiattribute health states revealed that there was significant concordance among all three methodologies taken together (as indicated by the overall ICC). This same conclusion held with respect to the comparison between each pair of methodologies (as indicated by pairwise ICC). The analysis at the level of the five corner states revealed that there was significant concordance among all methodologies taken together (as indicated by the overall ICC). However, with respect to the comparison between each pair of methodologies, it was found that there was significant concordance only between the CUF-RS and CUF-SG as well as the CUF-RS and DCCM methodologies (as indicated by the pairwise ICC). Although no significance testing was performed on the analysis at the level of the three interior states, it should be noted that the magnitude of the pairwise ICC between the CUF-RS and CUF-SG methodologies was nearly perfect, at 0.983.

With respect to Research Question # 14, the hypothesized functional relationships were tested using regression analysis. The analysis at the level of all eight multiattribute health states supported the hypothesized functional relationships between respondents' health status preference functions as determined by the different measurement methodologies -- i.e., a power curve relationship between the CUF-RS and CUF-SG as well as between the DCCM and CUF-SG methodologies, while a linear relationship between the CUF-RS and DCCM methodologies. The analysis at the level of
the five corner states supported the hypothesized functional relationships between the CUF-RS and CUF-SG as well as between the DCCM and CUF-SG methodologies, but not between the CUF-RS and DCCM methodologies. Although significance testing was not performed for the analysis at the level of the three interior states, it should be noted that the adjusted R² values for the fit of the functional relationship between the CUF-RS and CUF-SG as well as between the CUF-RS and DCCM methodologies were both above 0.90; however, the adjusted R² value for the fit of the functional relationship between the CUF-SG and DCCM methodologies was quite low -- 0.376.

**Comparing Respondent Evaluations**

Research Questions 15 through 18 pertained to comparing the respondent evaluations of the three preference measurement methodologies. Before addressing the comparison of respondent evaluations, it was necessary to determine the psychometric properties of the data collected on respondent evaluations.

Regarding the reliability of each set of respondent evaluations (Research Question # 15), Cronbach's alpha for each set was in the mid-to-high 0.80's. The discussion on the interpretation and acceptability of this value is deferred until a subsequent section of this chapter. An exploratory factor analysis was conducted on each dataset of respondent evaluations in order to determine the nature of the underlying factor structure of respondent evaluations (Research Question # 16). In each case, a two-factor solution was found to be appropriate for describing the underlying factorial structure of respondent evaluations.
Regarding the respondent evaluations of the three preference measurement methodologies using judgmental ratings (Research Question # 17), a series of repeated measures one-way analyses of variance with Scheffe tests for multiple comparisons were conducted in order to determine whether there were significant differences among the ratings given by respondents to the methodologies on each of the following variables -- interest, sensibility, difficulty in doing, usefulness, length, success, clarity, difficulty in understanding, and confidence. It was found that there were significant differences in respondent evaluations of difficulty in doing (specifically between the CUF-RS v CUF-SG and CUF-SG v DCCM methodologies, as indicated by the Scheffe test), usefulness (specifically between the CUF-RS v DCCM methodologies, as indicated by the Scheffe test), length (between each pairwise comparison, as indicated by the Scheffe test), success (specifically between the CUF-RS v DCCM methodologies, as indicated by the Scheffe test), difficulty in understanding (between each pairwise comparison, as indicated by the Scheffe test), and confidence (specifically between the CUF-SG and DCCM methodologies, as indicated by the Scheffe test). Further, there were also significant differences in total time taken by respondents to complete the tasks required of each methodology.

Regarding the preferential ordering of the three methodologies on each of the variables of interest for the respondent evaluations (Research Question # 18), it was found that the DCCM was the most preferred of all three methodologies on all the variables using both, direct comparisons as well as judgmental ratings as the response mode. For the variables of
interest, sensibility, usefulness, and success, the CUF-SG methodology was the second most preferred of all three methodologies using both response modes. However, differences in preferential rank ordering as a function of response mode were observed with the variables of clarity, difficulty in understanding, and confidence -- specifically, using direct comparisons as the response mode, the CUF-SG methodology was the second most preferred of all three methodologies, while using judgmental ratings as the response mode, the CUF-RS methodology was the second most preferred of all three methodologies.

**Discussion of Results**

The previous section summarized the results of the analyses conducted in order to address each of the research questions investigated in this study. This section discusses these results in terms of not only how they address each individual research question but also how they compare to results of similar studies in the published literature. As in the case of the previous section, the discussion is organized into three categories -- modeling the data, comparing preference scores, and comparing respondent evaluations. First, however, a few comments about the response analysis, respondent characteristics, and total time taken for the entire interview are provided.
Response Analysis

The usable response rate of patients was quite high, at 90.85%. Actually, the participation rate was even higher, at 92%, since one subject's responses were not used for analytical purposes (as explained in Chapter VI). In order to make a judgment regarding the acceptability of this participation rate, it is necessary to compare it to the participation rates obtained in similar studies in the literature. Torrance (1987) has noted that participation rates in health status preference measurement studies are lowest for the general public (70-84%) and highest for those with a special interest in the research, like patients or clinicians (83-100%). Since patients were used as the subjects of this study, the appropriate comparators for the participation rates obtained in this study would be published studies which have also used patients as subjects.

Table 75 presents the response analysis of selected health status preference measurement studies which have used patients as respondents. These studies are by no means an exhaustive listing of all studies which have used patients as respondents for the purpose of health status preference measurement; rather, they represent the most frequently cited studies in the literature which provide sufficient information to infer the entries in the table. The notable exception is the Wolfson, Sinclair, Bombardier, et al. (1982) study, which does not provide information about refusals and terminations. It was decided to include this study to provide an example of a study which had lower participation rates; further, this study is one of the more frequently cited studies in the literature.
Further, it should be noted that none of the studies included in Table 75 were conducted by any of the members of the EuroQol group. This is simply because the published literature on the EuroQol has mainly used mail surveys with the general population as respondents, as opposed to personal interviews with patients as respondents. Therefore, although it would be instructive to compare the response rates obtained in the present study with those obtained in studies using the EuroQol, the current status of the published literature does not permit such a comparison.

Comparing the response analysis of the present study to that of the other studies in Table 75, it can be seen that the participation rate of patients obtained in this study was the fourth highest among all studies. The refusal rate in this study was also among the lower ones of the studies compared in Table 75, i.e., the fifth lowest. Further, the fact that only one respondent broke-off the interview in midstream -- and that also because of problems with his allergies -- is quite encouraging. Overall, it may be concluded that the response rates obtained in this study were quite satisfactory, compared to those obtained in similar studies in the published literature.

This conclusion is especially noteworthy since the present study was conducted with outpatients, who were present at the clinic for a fixed period of time; several studies in Table 75 were conducted on inpatients, who might have been more inclined toward participating in any ongoing study in the hospital simply because they were admitted into the hospital and had nowhere else to go. The high participation rate obtained in this study could have been due to the fact that it was conducted on veterans, who, for the most part (as shown in Table 9) were retired.
Table 75: Response Analysis of Selected Health Status Preference Measurement Studies

<table>
<thead>
<tr>
<th>Condition</th>
<th>Patients</th>
<th>Approaches</th>
<th>Participants</th>
<th>Refusals</th>
<th>Terminations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>109 (100%)</td>
<td>109 (100%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>Cancer (unspecified)</td>
<td>80 (100%)</td>
<td>66 (83%)</td>
<td>14 (17%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>Chronic Stable Angina</td>
<td>311 (100%)</td>
<td>220 (71%)</td>
<td>91 (29%)</td>
<td>0 (0%)</td>
<td></td>
</tr>
<tr>
<td>End-stage Renal Disease</td>
<td>236 (100%)</td>
<td>194 (82%)</td>
<td>34 (14%)</td>
<td>8 (4%)</td>
<td></td>
</tr>
<tr>
<td>End-stage Renal Disease</td>
<td>50 (100%)</td>
<td>42 (84%)</td>
<td>6 (12%)</td>
<td>2 (4%)</td>
<td></td>
</tr>
<tr>
<td>End-stage Renal Disease</td>
<td>32 (100%)</td>
<td>29 (91%)</td>
<td>2 (6%)</td>
<td>1 (3%)</td>
<td></td>
</tr>
<tr>
<td>Laryngeal Cancer</td>
<td>97 (100%)</td>
<td>61 (63%)</td>
<td>30 (31%)</td>
<td>6 (6%)</td>
<td></td>
</tr>
<tr>
<td>Rheumatoid Arthritis</td>
<td>288 (100%)</td>
<td>279 (97%)</td>
<td>0 (0%)</td>
<td>9 (3%)</td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>15 (100%)</td>
<td>10 (67%)</td>
<td>--*</td>
<td>--*</td>
<td></td>
</tr>
<tr>
<td>Seriously Ill</td>
<td>1,897 (100%)</td>
<td>1,438 (76%)</td>
<td>247 (13%)</td>
<td>212 (11%)</td>
<td></td>
</tr>
<tr>
<td>Testicular Cancer</td>
<td>31 (100%)</td>
<td>30 (97%)</td>
<td>0 (0%)</td>
<td>1 (3%)</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>153 (100%)</td>
<td>140 (92%)</td>
<td>12 (8%)</td>
<td>1 (&lt;1%)</td>
<td></td>
</tr>
</tbody>
</table>
Table 75 (continued)

Legend:
- **a**: entries in table indicate number (percentage) of patients
- **b**: number (percentage) of patients approached to participate in study after meeting selection criteria
- **c**: number (percentage) of patients who were actually interviewed
- **d**: number (percentage) of patients who refused to participate when approached as well as patients who were unable to schedule appointments and no-shows
- **e**: number (percentage) of patients who broke-off interviews
- **f**: Boyd, Selby, Sutherland, et al. (1988)
- **g**: Llewellyn-Thomas, Sutherland, Tibshirani, et al. (1982)
- **h**: Sutherland, Dunn, and Boyd (1983)
- **i**: Llewellyn-Thomas, Sutherland, Tibshirani, et al. (1984)
- **j**: figures shown in table are for the first of two interviews per patient; there were two dropouts for the second interview due to poor health
- **k**: Nease, Kneeland, O'Connor, et al. (1995)
- **m**: figures shown in table are for the first of two interviews per patient; there were 23 dropouts for the second interview due to poor health (n=10), refusal (n=6) and interview scheduling conflicts (n=7)
- **o**: Torrance (1976a)
- **p**: Llewellyn-Thomas, Thiel, and McGreal (1992)
- **q**: Balaban, Sagi, Goldfarb, et al. (1986)
- **r**: Wolfson, Sinclair, Bombardier, et al. (1982)
- **s**: information not provided in article
- **t**: Tsevat, Cook, Green, et al. (1995); “seriously ill” defined as patients having clinical characteristics predictive of an aggregate 6-month mortality rate of about 50%
- **u**: Stiggelbout, Kiebert, Kievit, et al. (1994)
- **v**: present study
Respondent Characteristics

Regarding the descriptive characteristics of the subjects who participated in this study, it can be concluded that they are consistent with what might be expected in a sample of veterans. In this regard, the vast majority were male and retired. Although most had hypertension for over ten years, their hypertension seemed to be well-controlled, as evidenced by the breakdown of patients according to severity of hypertension provided in Table 10. Further, the pattern of comorbidities presented in Table 11 is also consistent with what might be expected for a sample of veterans, with arthritis or rheumatism, back pain, diabetes, heart trouble, and hyperlipidemia being the most frequently mentioned comorbidities; moreover, vision, hearing, and prostate trouble were also frequently mentioned comorbidities.

It is interesting to note that in spite of these comorbidities, the respondents of this study classified themselves as being at the highest levels of functioning on most of the health status dimensions of the EuroQol. The only exception was for the dimension of pain/discomfort, for which most respondents classified themselves as being on the intermediate level of functioning. This observation is consistent with the fact that arthritis or rheumatism and back pain were frequently occurring comorbidities in this patient population.

Taken together, the pattern of comorbidities and self-reported health status provide some support for the clinical validity of the EuroQol, in that the pain/discomfort dimension of the EuroQol was able to reflect the fact that the most frequently mentioned comorbidities were arthritis or
rheumatism and back pain (this was also supported by an individual level analysis, wherein each individual questionnaire was checked for correspondence between these items in the EuroQol and patient comorbidities).

Total Time Taken for the Entire Interview

The last point that needs to be noted before discussing the results of the analyses conducted to address each of the research questions investigated in this study is that of the total time taken for the entire interview. At little under an hour, the mean total time for the interview is quite long; however, as noted earlier, patients did not seem to mind the length of time they spent being interviewed. This maybe because of two reasons: (a) patients were waiting for their prescriptions to be filled by the pharmacy (as noted earlier, most patients were interviewed after their physician appointments at the clinic) and (b) patients truly enjoyed being interviewed -- some support for this notion lies in the fact that several patients voluntarily mentioned that they enjoyed the interview; moreover, it was the interviewer's observation that several patients appeared to empathize with some of the scenarios they were asked to evaluate (as evidenced by the fact that several patients noted this fact themselves -- again voluntarily). Therefore, it is concluded that although an hour is a long time to spend per patient interview, it was acceptable to the patients of this study.
Modeling the Data

There were two aspects of interest with regard to modeling the data obtained using each preference measurement methodology -- the nature of the model form of the health status preference function and the predictive ability of each preference measurement methodology. Each of these aspects will be discussed in turn in this section.

(a) Nature of the Model Form of the Health Status Preference Function

The fact that all three preference measurement methodologies supported the existence of a multiplicative model form for respondents' health status preference function is consistent with previous research on the EuroQol. As reviewed in Chapter II, there have been two previously published attempts to model the health status preference function as operationalized by the EuroQol (Dolan and Kind, 1994; van Hout and McDonnell, 1992). It is instructive to compare the results of these previous two modeling attempts with the results of the modeling procedures of this study.

Table 76 provides a listing of the different two-way interactions which were found to be statistically significant (at an alpha level of 0.05) in the two published modeling studies on the EuroQol and the present study. It can be seen that although there is agreement in terms of the general nature of the model form of the EuroQol health status preference function among the three studies (i.e., a multiplicative model form), there are some differences in terms of which specific health status dimensions interact with each other. First, the present study found only one significant two-way
interaction, between self-care and usual activities. This particular two-way interaction was also found to be significant in the Dolan and Kind (1994) study, but not in the van Hout and McDonnell (1992) study. Second, both the van Hout and McDonnell (1992) as well as Dolan and Kind (1994) found several (six, to be exact) two-way interactions to be significant, whereas the present study found only one two-way interaction to be significant.

In attempting to provide an explanation for these divergent results, it is necessary to revisit the specifics of each study. The van Hout and McDonnell (1992) study was based on data collected from a mail survey of the general population of Rotterdam, The Netherlands; the Dolan and Kind (1994) study was based on data collected from a mail survey of the general population of Finland; the present study collected data using personal interviews on veterans in Columbus, Ohio (USA). It can be seen that the three studies are not strictly comparable in terms of sample characteristics. Therefore, it is possible that the differences in interacting health status dimensions found in this study could be explained by the fact that the sample of respondents in this study was quite different than that of the other published studies which have modeled the EuroQol health status preference function.

In order to explore this possible explanation further, it is necessary to understand why veterans might have different health status preference functions as opposed to the general population and why the previous studies could not have adequately represented the health status preferences of this particular demographic sub-group.
Table 76: Comparison of Modeling Studies on the EuroQol*

<table>
<thead>
<tr>
<th>Interaction&lt;sup&gt;b&lt;/sup&gt;</th>
<th>I&lt;sup&gt;c&lt;/sup&gt;</th>
<th>II&lt;sup&gt;d&lt;/sup&gt;</th>
<th>III&lt;sup&gt;e&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOB &amp; SFC</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>MOB &amp; ACT</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>MOB &amp; PDF</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>MOB &amp; ANX</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>SFC &amp; ACT</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SFC &amp; PDF</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>SFC &amp; ANX</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>ACT &amp; PDF</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>ACT &amp; ANX</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>PDF &amp; ANX</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Legend:

a: Yes = interaction was significant; No = interaction was not significant
b: MOB = mobility; SFC = self-care; ACT = usual activities;
    PDF = pain/discomfort; ANX = anxiety/depression
c: van Hout and McDonnell (1992)
d: Dolan and Kind (1994)
e: present study
Regarding the issue of why veterans might have different health status preferences as opposed to the general population, it can be argued that this might stem from the plurality of medical conditions experienced by veterans compared to the general population. In this regard, veterans -- for the most part being elderly -- may have a greater incidence of serious health problems and comorbidities than the general population. Therefore, it is possible that they may not take some basic activities of daily living for granted, such as mobility and self-care. Specifically, with regard to the results of this study, it is possible that the mobility level of having no problems in walking about may have been so important to patients that they might have chosen an alternative which had this particular level regardless of the levels of the other health status attributes. This is, of course, an example of noncompensatory decision making, which -- as reviewed in Chapter IV -- is likely to occur when choice is used as the response mode. In support of this, it should be noted that several respondents informally thought aloud as they were making their choices, and their pattern of voiced-aloud thoughts suggested noncompensatory decision making with respect to the attribute of mobility.

Regarding the issue of why veterans (or respondents similar to veterans in terms of demographic characteristics) might not have been adequately represented in the previous EuroQol modeling studies, as noted in Table 9, the majority of the patients in this study were retired. Further, it was noted earlier that the mean age of respondents in this study was over 65 years. On the other hand, the respondents of the other modeling studies were largely employed and fairly young. For instance, in the case of the
Rotterdam study (Bonsel, Essink-Bot, van Hout, et al., 1992; Essink-Bot and Stouthard, 1992; Essink-Bot, Stouthard, and Bonsel, 1993; and Stouthard and Essink-Bot, 1992) -- the data of which was used for the modeling study by van Hout and McDonnell (1992) -- the majority of respondents who provided usable data were employed (54%) and under 45 years of age (53%). In the case of the Finnish study (Ohinmaa, Sintonen, and Pekurinen, 1994) -- the data of which was used for the modeling study by Dolan and Kind (1994) -- the vast majority of respondents who provided usable data were under 50 years of age (73%); no information about the main activity of respondents in the Finnish study was reported.

Further, a response analysis of the Rotterdam study showed that nonresponse was selectively higher among elderly and retired respondents (Essink-Bot and Stouthard, 1992; Essink-Bot, Stouthard, and Bonsel, 1993). Although a detailed nonresponse analysis for the Finnish study has not been reported, it was noted in the original study that older respondents were more likely to be nonrespondents, and further provide non-usable data even if they do respond (Ohinmaa, Sintonen, Pekurinen, 1994). These results are consistent with those reported in the survey sampling literature -- for instance, Cobb, King, and Chen (1957) have reported that elderly subjects are less likely to respond to mail surveys than younger subjects. Indeed, as noted in Chapter V, these results were instrumental in the decision to collect data using personal interviews for the purpose of this study.

Therefore, since the veterans of this study might have different health status preferences than the samples of the previous EuroQol modeling studies and the fact that respondents similar in demographic characteristics
were not adequately represented in these previous studies, the discrepancies in the specific interacting health status dimensions in the studies might be due to inherent differences in characteristics of the respondents of the studies.

It should be noted that there are other possible explanations for these discrepancies. One such explanation is due to change of process theory (Mellers, Ordonez, and Birnbaum, 1992), which was described in Chapter IV. In this regard, it may be argued that using choice as the response mode resulted in a different decision making process by respondents as compared to the process they might have used when asked to evaluate health states using judgmental response modes. Since the previous EuroQol modeling studies used judgmental ratings as the response mode, the differences observed between the present and previous EuroQol modeling studies might be explained by change of process theory. It should be noted that this possible explanation is not mutually exclusive from the previous one (on different sample characteristics), since the possibility of noncompensatory decision making could serve as the specific mechanism for operationalization of both possible explanations.

Finally, the possibility of respondent misunderstanding and confusion cannot be denied as a potential explanation for the discrepancy in results of the present and previous EuroQol modeling studies. In this regard, it should be noted that there were some instances of logically inconsistent valuations observed in both the previous EuroQol modeling studies. Logically inconsistent valuations for a pair of health states are those which exist when the “better” health state (which has at least the same or higher
level of functioning on all health status dimensions compared to the other
health state in the pair) is given a lower valuation. The existence of such
valuations is suggestive of misunderstanding of instructions or purpose of
the valuation task on the part of respondents.

Further, in a mail survey, it is not possible to clarify any doubts or
misunderstandings respondents might have. Specifically, Ohinmaa and
Sintonen (1994) reported that many elderly subjects did not understand the
valuation task in the Finnish study. The health status preferences measured
in the present study are logically consistent. This was revealed from an
individual-level eyeballing of the data collected by each preference
measurement methodology used in this study, and maybe reflective of the
consistency checks which were built-in into the interview schedule (see
Appendix C). Moreover, there were no counterintuitive preferences
observed in this study, unlike the Dolan and Kind (1994) study -- as
explained in Chapter II. Therefore, it is possible that the discrepancies in the
results of the present and previous EuroQol modeling studies may be
explained by respondent misunderstanding and confusion in the previous
studies as opposed to the present study.

It should be noted that the entire discussion thus far on the modeling
of the EuroQol health status preference function in this study has focused on
the results of the discrete choice conjoint methodology (DCCM). The
EuroQol health status preference function was also modeled using two other
methodologies in this study, i.e., the conditional utility function-based
procedure with a rating scale scaling method (CUF-RS) and the conditional
utility function-based procedure with a standard gamble scaling method
(CUF-SG). The reason for emphasizing only the results of the DCCM in this section so far is that it provides the most directly comparable results with the previous EuroQol modeling studies. This is because it allows estimation of which specific health status dimensions interact with each other, which is along the lines of the modeling procedure used by the EuroQol group of researchers (who mainly used stepwise linear regression).

The multiplicative preference function as determined by the conditional utility function-based procedure, on the other hand, does not permit the estimation of which specific attributes or dimensions interact with one another. This is best illustrated by expanding the compact version of the equation of the multiplicative model form presented in previous chapters of this dissertation. As presented in earlier chapters, the compact version of the multiplicative model form is given by:

\[ p(x) = \frac{1}{k} \left( \prod \left[ 1 + k_i \cdot p_i(x) \right] - 1 \right) ; \quad (7.1) \]

where

- \( i = 1 \ldots 5 \) (i.e., five health status dimensions in the EuroQol)
- \( x \in X \) = multiattribute alternative
- \( k \) = interaction parameter \((-1 < k < 0 \text{ or } k > 0)\)
  \[ 1 + k = \prod \left[ 1 + k_i \right] \]
- \( k_i \) = scaling constant \((0 \leq k_i \leq 1)\)
  \[ x_i = \text{level of } x \text{ on the } i\text{-th attribute} \]
- \( p_i(x) = \text{i-th single attribute preference function} \)
- \( p(x) = \text{overall preference function for multiattribute alternative } x \)
Keeney and Raiffa (1976) have shown that equation (7.1) is mathematically equivalent to its expanded form (for \( i = 5 \) attributes):

\[
p(x) = k_p(x_1) + k_p(x_2) + k_p(x_3) + k_p(x_4) + k_p(x_5) +
\]

\[
kk_kk_p(x_1)p_2(x_2) + kk_kk_p(x_1)p_3(x_3) + kk_kk_p(x_1)p_4(x_4) +
\]

\[
nk_kk_kk_kk_p(x_1)p_3(x_3)p_4(x_4) + nk_kk_kk_kk_p(x_1)p_5(x_5) +
\]

\[
nk^2k_kk_kk_kk_kk_p(x_1)p_3(x_3)p_4(x_4)p_5(x_5)
\]

Equation (7.2) clarifies the role of the interaction parameter \( k \) in the multiplicative model form. It can be seen that \( k \) serves as a global (i.e., non-specific) interaction parameter in the multiplicative model form. This is because there is no separate parameter for each of the possible interactions (if there was, the model form would be multilinear, not multiplicative). Therefore, when modeling a preference function using the conditional utility function-based procedure, it is not possible to specify which particular dimensions interact with one another in the case of the multiplicative model form. On the other hand, the multiplicative model form shown in equation (7.2) is not limited to the case of two-way interactions, as evidenced by the last ten terms in equation (7.2). However, these higher-order interactions are again represented by way of a global interaction constant -- i.e., there is no way of determining the parameters of each specific higher order interaction.
This brings up the issue of the distinction between the model forms of the EuroQol health status preference function as determined by the CUF-RS and CUF-SG methodologies on one hand and by the DCCM on the other hand. As determined by the CUF measurement strategy, the multiplicative model form does not permit specification of which particular health status dimensions interact with one another. Rather, both two-way as well as higher order interactions are represented by way of a global interaction constant, i.e., $k$. On the other hand, the DCCM as used in this study permits the specification of which particular two-way interactions are significant; however, higher-order interactions are not estimated.

Therefore, it can be seen that as used in this study, none of the three methodologies permits the complete specification of the model form of the EuroQol health status preference function. In this regard, the CUF-RS and CUF-SG methodologies allow for higher-order interactions among health status dimensions but do not permit the specification of which particular dimensions interact with one another. The DCCM, on the other hand, permits the specification of which particular dimensions interact with one another, but only for the case of two-way interactions -- not higher-order interactions.

It should be noted that these limitations are with respect to the implementation of the respective measurement strategies in this study, since the conditional utility function-based measurement strategy permits the specification of which particular dimensions interact with one another if the multilinear model form is estimated, and the discrete choice conjoint measurement strategy permits the estimation of higher-order interactions.
The decision to not utilize these capabilities of the respective measurement strategies for the purpose of this study was explained in Chapter V.

To briefly reiterate, Keeney and Raiffa (1976), who developed the conditional utility function-based procedure, have noted that the multilinear model form is rarely seen in practical situations involving more than three dimensions. Further, in order to estimate the multilinear model form using the conditional utility function-based procedure, it is necessary for each respondent to evaluate \((2^n - 2)\) alternatives, where \(n\) is the number of dimensions. Therefore, in the case of the present study, the each respondent would be required to evaluate 30 different health states for the estimation sample; this was deemed excessive, especially since this would be the load for only one of the three preference measurement methodologies which respondents would be required to implement. Regarding the DCCM, the decision not to estimate higher-order interactions for the purpose of this study was based on the advice of Batsell and Louviere (1991), who in their state-of-the-art review of DCCM observed that there was little to be gained from estimating higher-order interactions over and above all two-way interactions between dimensions. Further, the additional parameters which would need to be estimated in order to model higher-order interactions would dramatically increase either the number of respondents or the number of choice tasks presented to each respondent or both.

Thus, it can be seen that the model forms of the EuroQol health status preference function as determined by the CUF-RS and CUF-SG methodologies on one hand and the DCCM on the other hand are not strictly similar in terms of the specific dimensions which interact with one another.
and the estimation of higher-order interactions. This explains the observation relevant to Tables 33 and 34, which was noted in Chapter VI -- i.e., the difference in relative orderings of multiattribute health states and individual attribute levels, wherein the DCCM gave the highest preference scores for the individual attribute levels in Table 33 but the CUF-SG methodology gave the highest preference scores for multiattribute health states in Table 34. In this regard, although the CUF-SG preference scores for the individual attribute levels were lower than the DCCM preference scores, the greater degree of interactions (in terms of the general two-way and higher-order interactions) afforded by the model form determined by the CUF-SG methodology as opposed to the DCCM would allow the multiattribute health state preference scores determined by the CUF-SG methodology to be higher than those determined by the DCCM.

Still, in spite of the differences in the specifics of the model forms of the EuroQol health status preference function determined by the CUF-RS and CUF-SG on one hand and the DCCM on the other, it is noteworthy that both types of modeling procedures agreed in terms of the general nature of the EuroQol health status preference function -- i.e., a multiplicative model form, characterized by interactions among health status dimensions. This is especially reassuring since the previous research on modeling the EuroQol health status preference function also supports the existence of the multiplicative model form (Dolan and Kind, 1994; van Hout and McDonnell, 1992). Further, although a case may be made that the CUF-SG methodology supported an additive model form (on the basis of the sum of the scaling constants, with a margin for random error), it was shown in Chapter VI that
the multiplicative model form was superior to the additive model form in terms of predictive ability. Therefore, it may be concluded that health status as operationalized by the EuroQol has a multiplicative preference function. Further, it is noteworthy that the general model form of the EuroQol health status preference function is not sensitive to the particular response mode used for data collection purposes, i.e., judgments or choices.

(b) Predictive Ability of Each Preference Measurement Methodology

The second aspect of interest in modeling the EuroQol health status preference functions was the predictive ability of each preference measurement methodology. As explained in Chapter II, it was important to estimate the predictive ability of each methodology in order to gauge how accurate they were in measuring the construct they were intended to measure, i.e., health status preferences. In order to make this judgment, it is necessary to compare the estimates of predictive ability -- i.e., the mean absolute residuals -- calculated for each methodology with those obtained in similar studies in the published literature.

In this regard, the work of the McMaster group of researchers serves as a good comparator since it is this group which pioneered the use of the conditional utility function-based procedure in health status research. Further, the work of this group of researchers continues to be one of the more influential forces in health status preference measurement, as indicated by a review of citations of published articles in the field. Since the results of modeling health status as operationalized by the Mark III version of the
Health Utilities Index is not yet available, this discussion will focus on the previous versions (i.e., Mark I and Mark II) of the Health Utilities Index.

Regarding the Mark I version, there were two hold out validation health states -- the first had a residual of 0.17 and the second had a residual of 0.30 units; thus, the mean (standard deviation) absolute residual was about 0.24 (0.09) units (Torrance, Boyle, and Horwood, 1982). Although this estimate is quite high, it should be noted that it may not be strictly comparable to the methodologies used in this study, since it used a conditional utility function-based procedure with a time tradeoff scaling method.

The research on Mark II version provides a more direct comparison standard with that of the present study, since it used two of the preference measurement methodologies used in this study -- CUF-RS and CUF-SG. In each case, four hold out validation health states were used. In the case of the CUF-RS methodology, the mean (standard deviation) absolute residual was 0.08 (0.10) units, while in the case of the CUF-SG methodology, the mean (standard deviation) absolute residual was 0.04 (0.06) units (Torrance, Zhang, Feeny, et al., 1992). In comparison, in the case of the present study, the mean (standard deviation) absolute residuals of the CUF-RS and CUF-SG methodologies were 0.051 (0.03) and 0.045 (0.07) units respectively.

It can be seen that the predictive ability of the CUF-RS and CUF-SG methodologies in the present study was quite acceptable, compared to the predictive ability of the methodologies used to measure health status preferences for the Mark II version of the Health Utilities Index. Specifically, the CUF-RS methodology was more accurate in the present study, while the
CUF-SG methodology was slightly less accurate in the present study as compared to their counterparts in the Torrance, Zhang, Feeny, et al. (1992) study. Therefore, it may be concluded that the CUF-RS and CUF-SG methodologies were acceptable in terms of predictive ability of health status preferences.

It is somewhat more difficult to find a comparative standard for the case of the DCCM, since there have been no published applications of this methodology to the measurement of health status preferences. In the absence of such applications, it was decided to compare the estimate of predictive ability calculated in this study with similar estimates obtained in the marketing research literature, where this methodology is increasingly being used. The results of two studies -- both of which involved Louviere, who introduced DCCM into the field of marketing research (see Louviere and Woodworth, 1983) -- are useful in this regard. The first study, by Louviere and Hensher (1983b) found a mean (standard deviation) absolute residual of 0.02 (0.02) units, while the second study, by Elrod, Louviere, and Davey (1992) found a mean absolute residual of 0.04 units (standard deviations were not reported in this latter case). In comparison, the mean (standard deviation) absolute residual for the two holdout choice sets in the present study was 0.09 (0.10) and 0.02 (0.02) units respectively. Although the estimates in the present study are variable, they are still acceptable compared to the above-noted comparators (especially the second holdout choice set). Therefore, it may be concluded that the DCCM methodology was acceptable in terms of predictive ability of health status preferences.
Summarizing the discussion of the results of the modeling procedures, it can be stated that the present study supported the existence of a multiplicative model form for health status preferences as operationalized by the EuroQol and the each of the three preference measurement methodologies was acceptable in terms of predictive ability. Therefore, each methodology was acceptably accurate in measuring the construct it was intended to measure. The next set of research questions was intended to explore the extent to which the three preference measurement methodologies were measuring the same construct, i.e., health status preferences.

Comparing Preference Scores

There were five bases of comparison of the preference scores obtained using the three preference measurement methodologies -- preferential ordering, magnitude of preference scores, correlations, concordances, and functional relationships. Further, each set of analyses was performed at three different levels -- all eight multiattribute health states, the five corner states only, and the three interior states only. The results of each set of analyses for each of the bases of comparison will be discussed in this section.

Before discussing each set of results, it is necessary to revisit the purpose of using each of the five bases of comparison in the context of the overall purpose of this dissertation. As explained in Chapter II, the overall objective of this dissertation was to validate discrete choice conjoint methodology (DCCM) as a measure of health status preferences. The approach to validation adopted was by way of construct validation,
specifically convergent validation against two commonly used measurement methodologies for measuring health status preferences -- the conditional utility function-based procedure with a rating scale scaling method or CUF-RS methodology and the conditional utility function-based procedure with a standard gamble scaling method or CUF-SG methodology. The premise behind this approach, as explained in Chapter II, is that different measures of the same construct should covary in predictable ways.

Hence, it was decided to compare the measures or preference scores obtained using the different methodologies on the basis of five indicators. Further, specific hypotheses were forwarded for each of these five bases of comparison; if the results of the analyses provide empirical support for these hypotheses, the construct validity of DCCM as a measure of health status preferences would be supported. The rationale for the selection of these five bases of comparison was discussed in Chapter II, and will not be repeated here. This discussion will focus on linking the hypotheses forwarded in Chapter II to the results presented in Chapter VI, with the overall intent of determining whether the construct validity of DCCM as a measure of health status preferences was supported by the results of this study.

The discussion is organized into three sections. The first section discusses the results of each of the five bases of comparison of the preference scores obtained using the three methodologies. The second section addresses the issue of whether the present study was able to replicate the relationship between the CUF-RS and CUF-SG methodologies. As noted in Chapter II, this was a necessary prerequisite to estimating the construct validity of DCCM via the principle of convergent validation -- if the relationship
between the CUF-RS and CUF-SG methodologies which was reported in the published literature (as reviewed in Chapter IV) could be replicated in the present study, then the "comparators" for the validation of DCCM would be established. The third section addresses the issue whether the construct validity of DCCM as a measure of health status preferences was supported by the results of the present study, according to the principle of convergent validation.

(a) The Comparison of Preference Scores

As noted above, the first section of this discussion pertains to comparing the preference scores obtained using the three methodologies investigated in this study. Each of the five bases of comparison will be discussed in turn.

(i) Preferential Ordering

The first aspect on which the different preference measurement methodologies were compared was in terms of the order in which they ranked the alternative health states under evaluation. The results of the analysis at the level of all eight multiattribute health states will be discussed first. At this level, there was significant agreement among the three methodologies regarding the preferential ordering of alternative health states -- taken all three at a time (as measured by Kendall's W) and taken two methodologies at a time (as measured by Kendall's τ). A closer inspection of Tables 34 through 38 shows that there was not uniform agreement between each set of methodologies, taken two at a time.
Specifically, there was perfect agreement in preferential rank ordering of alternative health states between the CUF-RS and CUF-SG methodologies (as indicated by a value of 1.00 for Kendall's $\tau$) but less than perfect agreement in preferential rank ordering of alternative health states between the DCCM and each of the other preference measurement methodologies (as indicated by the appropriate values of Kendall's $\tau$).

Taken together, these results suggest that although there is significant overall agreement in preferential rank ordering of alternative health states (i.e., corner and interior states together), there is more perfect agreement between the CUF-RS and CUF-SG methodologies as opposed to between the DCCM and other methodologies investigated in this dissertation. This observation has implications for the assumption of procedural invariance and the phenomenon of preference reversal. In this regard, it should be noted that both the CUF-RS and CUF-SG methodologies use judgments as the response mode, while the DCCM uses choices as the response mode. At this level of the analysis, therefore, the distinction between judgmental and choice responses does not seem to have empirical support, since the preferential rank ordering of alternative health states by both response modes was not significantly different. Still, it should be noted that there does seem to be some basis to the judgment versus choice distinction in response modes, since the agreement in preferential rank ordering of alternative health states was greater between the methodologies which used judgmental response modes (i.e., CUF-RS and CUF-SG methodologies) as opposed to between the methodologies which used different response modes (i.e., CUF-RS and DCCM as well as CUF-SG and DCCM methodologies). This is
probably indicative of the nature of the alternative health states which were evaluated; therefore, a clearer picture might emerge when the eight alternative health states are split into their components -- i.e., corner and interior states -- and analyzed separately.

The analysis at the level of the five corner states yields some interesting results. As in the case of the analysis at the level of all eight multiattribute health states, there is significant agreement in preferential rank ordering of the five corner states when all three methodologies are considered together (as indicated by Kendall's W). However, when the agreement between each pair of methodologies is considered, Tables 36 through 38 show that there is significant agreement in preferential rank ordering of the five corner states between only the CUF-RS and CUF-SG methodologies. As noted above, these are the two methodologies which used judgments as the response mode. Therefore, the results at this level of the analysis support the distinction between judgmental and choice response modes, since there was significant agreement in preferential rank ordering of the five corner states between the methodologies which used judgments as the response mode (i.e., CUF-RS and CUF-SG methodologies) but not between the methodologies which used different response modes (i.e., CUF-RS and DCCM as well as CUF-SG and DCCM methodologies).

The analysis at the level of the three interior states only showed that there was perfect agreement in preferential rank ordering of the three interior states among all three methodologies -- taken together or two at a time. This result may again not support the distinction between judgment and choice response modes, since the methodologies using different
response modes agreed in terms of the preferential ordering of alternative health states (or the assumption of procedural invariance was upheld).

In the face of such seemingly divergent results, it is instructive to dwell upon the specifics of the corner and interior states. The corner states used in this study used the disvalue and disutility formulations of the preference scales; therefore, as explained in Chapter V, they were constructed so that each corner state consisted of one health status dimension on its worst level of functioning and all other dimensions on their best levels of functioning. In other words, the corner states had only one health status dimension which was not on the best level of functioning. On the other hand, the interior states had several health status dimensions which were not on the best level of functioning.

These structural differences in the corner versus interior states may be reflective of the different preference scores (in a relative sense) given to these alternative health states. In this regard, it is possible that if the lowest levels of functioning of each of the five health status dimensions were all equally undesirable and the highest levels of functioning of each of the five health status dimensions were all equally desirable, then there would not be substantial differences in the preference scores given to the five corner states. On the basis of the preference scores presented in Table 34, this may indeed be the case for the corner states in this study. It can be seen that with the exception of the corner state for mobility, the preference scores for all the other corner states obtained using any particular methodology are quite similar to each other -- in the case of the CUF-RS methodology, the difference between the highest and lowest preference scores for the
remaining four corner states is 0.12 units; the corresponding difference for the CUF-SG and DCCM methodologies is 0.08 and 0.10 units. In this regard, it should be remembered that a difference of 0.10 units has been recommended as being of practical significance (Furlong, Feeny, Torrance, et al., 1990). Therefore, it is possible that the differences observed in the preference scores of the four corner states apart from the corner state for mobility obtained using each methodology may not be of practical significance. On the other hand, the preference scores for the interior states are quite well-spaced apart from one another.

Keeping in mind that it was in the analysis at the level of the corner states only -- and not the interior states only -- in which the judgment versus choice distinction had some empirical support in this study, it is possible that response mode effects may be mediated by the similarity of alternatives in terms of preference scores. In other words, the assumption of procedural invariance may breakdown in practice only when the alternatives under evaluation are perceived as similar or equally (un)attractive. It should be noted that this is not tantamount to saying that the selection decision is random when the choice is between equally attractive alternatives; indeed, as Slovic (1975) demonstrated (and as reviewed in Chapter IV), this is certainly not the case.

As Slovic (1975) showed, there is a systematic component in operation when selecting between equally attractive alternatives. However, this systematic component is usually operative only when choice is used as the response mode. The systematic component stems from the issue of accountability or justification, which usually accompanies the decision making process when
choice is used as the response mode. As noted in Chapter IV, this component is usually not in operation when judgments are used as the response mode [however, see Hagafors and Brehmer (1983) for an exception]. Therefore, on account of the differential operation of this systematic component in judgment and choice response modes, it is possible that the breakdown of procedural invariance may lead to the phenomenon of preference reversal when evaluating similar or equally attractive alternatives.

This explanation can be viewed as complimentary to the prominence hypothesis, which in turn is consistent with contingent weighting theory (Tversky, Sattath, and Slovic, 1988), which was reviewed in Chapter IV. However, it should be noted that the other theoretical expositions for response mode effects reviewed in Chapter IV -- i.e., expression theory (Goldstein and Einhorn, 1987) and change of process theory (Mellers, Ordonez, and Birnbaum, 1992) -- cannot be ruled out as possible explanations of the results obtained in this study. In order to provide a falsification test for expression theory, it is necessary to collect data on decision making processes (via process tracing methods as outlined in Chapter IV). On the other hand, although the general model form determined by all three measurement methodologies was multiplicative in nature, the strict non-comparability of the conditional utility function-based procedure and discrete choice modeling in terms of the specific interaction components does not allow for a falsification test of change of process theory. Therefore, although the pattern of results obtained in this study is consistent with the postulates of contingent weighting theory, it is possible
that it could also be accommodated by the other theoretical expositions of response mode effects.

The potential explanation provided above adequately accounts for the results obtained in this study. In the case of the corner states only, the preference scores were quite similar within (but not across) each methodology for the four corner states not including the corner state for mobility. However, the preference scores for the interior states are well-separated. Therefore, there may be a breakdown of the assumption of procedural invariance in the case of the analysis of the corner states only, but not in the case of the interior states only. This could explain why there was less agreement in preferential rank ordering between the DCCM and other methodologies in case of the analysis on corner states only compared to the analysis on interior states only. Regarding the analysis on all eight multiattribute health states (i.e., the five corner and three interior states), significant agreement in preferential rank ordering between each pair of methodologies might be due to the “cancellation” effect, wherein the procedural invariance effect observed in the analysis of the interior states only was more influential than the breakdown of the procedural invariance effect observed in the analysis of the corner states only.

A few points need to be noted before moving on to the discussion of the magnitude of preference scores. First, at all levels of the analysis in this study, there was perfect agreement in rank ordering between the CUF-RS and CUF-SG methodologies. Strictly speaking, the response modes used by both these methodologies were different; the CUF-RS methodology used judgmental ratings, while the CUF-SG methodology used indifference
judgments. In this regard, no empirical support was found in this study for response mode effects on account of these two different types of judgmental responses (this is why the discussion above referred to both types of response modes generally as judgmental responses). Therefore, the results of this study do not support Montgomery, Selart, Garling, et al.'s (1994) argument of preference reversal between judgmental ratings and indifference judgments on account of the prominence effect when the former but not the latter are used as the response mode (as reviewed in Chapter IV).

Second, it should be noted that although some consistency checks were built in into the interview schedule (see Appendix C), these were only with respect to single methodologies taken separately; no attempt was made to influence consistency (either for or against) across measurement methodologies. As noted earlier, these consistency checks were necessary for the proper implementation of the conditional utility function-based procedure (Keeney and Raiffa, 1976).

Summarizing the results of the comparison of the preferential ordering of alternative health states by the different preference measurement methodologies investigated in this study, there was significant agreement among all three methodologies when all eight multiattribute health states were considered together. When analyzed by the component corner and interior states, it was found that the analysis on the five corner states showed significant agreement between only the CUF-RS and CUF-SG methodologies; on the other hand, the analysis on the three interior states showed perfect agreement among all three methodologies. These results
were explained on the basis of the moderating effect of similarity or equal (un)attractiveness of alternatives on the response mode effect. Specifically, when the alternatives under evaluation are similar or equally (un)attractive, the assumption of procedural invariance breaks down in practice, leading to the phenomenon of preference reversal; however, when the alternatives under evaluation are dissimilar and easily distinguishable from one another, the assumption of procedural invariance holds in practice, leading to perfect agreement in preferential rank ordering of alternatives by different measurement methodologies. It was argued that this pattern of results was consistent with contingent weighting theory (Tversky, Sattath, and Slovic, 1988), although expression theory (Goldstein and Einhorn, 1987) and change of process theory (Mellers, Ordonez, and Birnbaum, 1992) could not be ruled out as possible explanations for the results obtained.

(ii) Magnitude of Preference Scores

The second aspect on which the different preference measurement methodologies were compared was in terms of the magnitude of the preference scores they ascribed to the alternative health states under evaluation. Before discussing the results of the repeated measures analyses of variance which were conducted in order to address the research question pertaining to this aspect of the comparison of preference scores, it is instructive to note the general pattern of the magnitude of preference scores, as reported in Table 34. In this regard, it can be seen that the CUF-SG methodology always provided the highest preference scores to the alternative health states, followed by the DCCM, and finally by the CUF-RS
methodology. The only exception to this observed pattern of the magnitude of preference scores was in the case of the health state E8 (for a description of this state, see Appendix B), in which the DCCM provided a lower preference score than the CUF-RS methodology.

The analysis at the level of all eight multiattribute health states showed that there were significant differences between the preference scores of the alternative health states obtained using the CUF-SG and CUF-RS as well as the CUF-SG and DCCM methodologies. Although the analysis at the level of the five corner states also found a significant difference between the preference scores of the alternative health states obtained using the CUF-SG and CUF-RS methodologies, no significant differences were found between the preference scores obtained using the CUF-SG and DCCM methodologies; rather, the difference between the preference scores obtained using the CUF-RS and DCCM methodologies was found to be significant. The analysis at the level of the three interior states found no significant differences among the preference scores of the alternative health states obtained using the three different preference measurement methodologies.

In explaining this pattern of results, emphasis will be placed on the analysis at the levels of all eight multiattribute health states and the five corner states. This is because three health states is an extremely small number upon which to base the results of inferential statistics, since the small sample size would consequently lead to less statistical power to detect significant differences among the groups. This can be seen by the fact that the preference scores obtained using the CUF-SG methodology appear to be clearly higher than the other preference scores (as seen in Table 34).
The fact that the preference scores obtained using the CUF-SG methodology were always significantly higher than those obtained using the CUF-RS methodology is consistent with the body of theoretical and empirical evidence in both the decision theoretic as well as health status preference measurement literature reviewed in Chapter IV. As explained in that chapter, this is attributable to the effects of attitude toward risk of respondents, which is encoded in the utility function produced by the CUF-SG methodology but not in the value function produced by the CUF-RS methodology.

The other pairwise comparisons yielded divergent results at the two levels of analysis. Considering first the comparison between the preference scores of alternative health states obtained using the CUF-SG and DCCM methodologies, it was noted earlier that the published literature did not offer much guidance in terms of the relative magnitudes of preference scores obtained using these methodologies. The discrepancy between the results of the analyses at the levels of the corner states only and all multiattribute health states could be attributed to the nature of the corner and interior states. As noted earlier, the corner states had only one health status dimension which was not on the best level of functioning, while the interior states had several health status dimensions which were not on the best level of functioning.

These structural differences between the corner and interior states combined with the attitude toward risk of the respondents of this study may explain the pattern of observed results with respect to this aspect of the comparison among the three preference measurement methodologies.
Specifically, since the value of the interaction parameter $k$ for the multiplicative model form of the health status preference function obtained using the CUF-SG methodology was -0.1196, the respondents of this study were strictly multiattribute risk-averse (Richard, 1975). The determination of respondents' attitude toward risk from the value of the interaction parameter $k$ was discussed in Chapter IV, and will not be repeated here.

Given that the respondents of this study were strictly multiattribute risk-averse, it is understandable why the CUF-SG scores were generally higher than the DCCM scores -- they would opt for the sure-thing alternative in the standard gamble on account of their risk-aversion, thus leading to a higher preference score as determined by an indifference judgment between the sure-thing alternative and the lottery alternative. In order to explain how respondents' risk-aversion could account for the observation that the CUF-SG scores were significantly higher than the DCCM scores when considering both corner as well as interior states but not when considering corner states only, it is necessary to understand how risk-aversion may differentially influence the evaluation of corner versus interior states.

As noted above, the effect of risk-aversion on the evaluation of health states would be to provide higher preference scores via indifference judgments using the standard gamble than the preference scores that would be obtained with a preference measure which did not incorporate the effect of risk-aversion. Since the interior states consist of several health status dimensions which are not on the highest level of functioning while the corner states consist of only one health status dimension which is not on the highest level of functioning, it is possible that the effect of risk-aversion may
be exerted to a greater extent for interior states as opposed to corner states. In other words, the upward adjustment of preference scores due to risk aversion may be greater for interior states than corner states. It should be noted that this does not mean that the preference scores for the interior states will be greater than those for the corner states; what is of relevance here is the adjustment on the “base” score (i.e., the score which does not incorporate the effects of attitude toward risk).

Viewed in this way, the analysis at the level of the corner states only may not show significant differences between the preference scores of alternative health states obtained using the CUF-SG and DCCM methodologies, since the upward adjustment of the CUF-SG score (on account of the effect of risk aversion) to the “base” score -- which, in this case, would be the DCCM score -- may not be large in magnitude since four out of the five health status dimensions are on their highest levels of functioning. This provides an explanation of the results obtained for this level of the analysis in this study.

On the other hand, the analysis at the level of the interior states only may show significant differences between the preference scores of alternative health states obtained using the CUF-SG and DCCM methodologies, since the upward adjustment of the CUF-SG score (on account of the effect of risk aversion) to the “base” score -- which again would be the DCCM score -- may be large enough in magnitude since several health status dimensions are not on their highest levels of functioning. As noted earlier, significance testing was not performed at this level of the analysis; still, Table 34 indicates that the logic outlined above
might explain the possible results if the preference scores of more interior states were measured.

Regarding the analysis at the level of all eight multiattribute health states, the significance of the differences in preference scores between the CUF-SG and DCCM methodologies would depend on not only the proportion of corner and interior states but also the relative adjustments on the "base" scores for each of the states. In this study, a significant difference was observed between the preference scores of the alternative health states obtained using the CUF-SG and DCCM methodologies. Therefore, it is possible that the relative adjustment on the "base" (i.e., DCCM) scores of the interior states was large enough so that when considered along with the relative adjustments on the "base" scores of the corner states, the overall adjustment due to risk-aversion on all eight health states was sufficient to produce statistically significant differences between the CUF-SG and DCCM scores.

It should be noted that although a similar line of argument could be applied to the comparison between preference scores obtained using the CUF-RS and CUF-SG methodologies, it is not sufficient to explain the pattern of results obtained for that comparison. This is because of the difference in "base" preference scores of the DCCM and CUF-RS methodologies. The reason for this difference in "base" scores -- as explained in Chapter IV -- is the fact that the DCCM methodology, using a top-down strategy, inherently accounts for interactions among the health status dimensions. Since the CUF-RS methodology uses a bottom-up strategy, it does not inherently account for such interactions. Therefore, the "base"
scores of the DCCM would be higher than those of the CUF-RS methodology. Therefore, the same relative adjustment to “base” scores which may not be large enough to produce statistically significant differences between the preference scores of the CUF-SG and DCCM methodologies may be large enough to produce statistically significant differences between the preference scores of the CUF-SG and CUF-RS methodologies.

Moving on to the comparison between the preference scores of alternative health states obtained using the CUF-RS and DCCM methodologies, the results of the analysis at the level of all eight multiattribute health states did not find any statistically significant differences between the preference scores of alternative health states obtained using these methodologies. However, the results of the analysis at the level of the five corner states found statistically significant differences between the preference scores of alternative health states obtained using these methodologies. Since neither of these methodologies measures preferences under risky conditions, the effect of attitude toward risk cannot explain these results.

As noted earlier, it was hypothesized that the DCCM scores would be higher than the CUF-RS scores since they inherently account for interactions among health status dimensions. This hypothesis was supported in the analysis at the level of the five corner states only, but not in the analysis at the level of all eight multiattribute health states. However, when the analysis on corner and interior states was run without including the state E8 (the description of which is provided in Appendix B), it was found that the
DCCM scores were significantly higher than the CUF-RS scores (in accordance with the hypothesized relationship).

It is important to note the reason why the analysis was run without the state E8. This was simply because this particular health state was the only one for which the CUF-RS score was greater in magnitude than the DCCM score. Therefore, it is possible that there may have been some error in measurement of the preferences for that particular health state.

In summary, the results of this aspect of the comparison of preference scores showed that there were significant differences in the magnitudes of preference scores of alternative health states obtained using the three different measurement methodologies investigated in this study. The preference scores obtained using the CUF-SG methodology were the highest, followed by the preference scores obtained using the DCCM, and finally the CUF-RS scores. The differences between the CUF-SG and other scores were explained on the basis of attitude toward risk of the respondents. Specifically, it was shown that the respondents of this study were strictly multiattribute risk-averse. Further, it was argued that this risk-aversion differently influenced the evaluation of corner and interior states. Finally, the difference between DCCM and CUF-RS scores was explained on the basis of the fact that the DCCM scores inherently account for interactions among health status dimensions, while CUF-RS scores do not.
(iii) Correlations

The third aspect on which the different preference measurement methodologies were compared was in terms of the correlations between the preference scores of alternative health states obtained using the different methodologies. Two types of statistical measures were used to indicate the extent of correlation between the preference scores obtained using the three preference measurement methodologies -- Spearman’s $\rho$ and Pearson’s $r$. The results in both cases were very similar to those obtained using Kendall’s $\tau$ (which were discussed earlier).

Considering the results of the rank-order correlation first (measured using Spearman’s $\rho$), it is not surprising that these were similar to those obtained using Kendall’s $\tau$. This is because both these statistics provide insight into the nature of the ordinal relationship between two variables; where these statistics differ is in the way they measure this ordinal relationship. Kendall’s $\tau$ provides a measure of the agreement in preferential rank ordering of alternatives by two measures, while Spearman’s $\rho$ provides a measure of the rank-order correlation between two measures. As explained in Chapter II, correlation between measures is not tantamount to agreement between the measures. The fact that the results of the analyses using both these statistics led to the same conclusions is an indication of the consistency of the nature of ordinal relationship between the preference measurement methodologies. This quite reassuring from the viewpoint of an “internal validation” of the measurement methodologies investigated in this study, since it suggests that the ordinal relationship between the preference measurement methodologies is not sensitive to the statistical measure used.
to provide an indication of the relationship. The explanation provided in the section which discussed the pattern of results of the analysis using Kendall's $\tau$ is applicable in the case of the results of the analysis using Spearman's $\rho$ also; therefore, it will not be repeated here.

In the face of such similar results between the analyses using Kendall's $\tau$ and Spearman's $\rho$, the question that may arise is why use both measures to describe the ordinal relationship between the preference measurement methodologies investigated in this study. In this regard, it should be stressed that both statistics provide different information about the nature of the ordinal relationship between two measures, and are not strictly substitutable for one another. Indeed, when there are no ties (as was the case in this study), the absolute value of Spearman's $\rho$ exceeds that of Kendall's $\tau$ by as much as 50% (Gibbons, 1993). Therefore, it is possible that inferences based on Spearman's $\rho$ may be different than those based on Kendall's $\tau$. As explained earlier, it was decided to use both statistics because they provide different information about the nature of the ordinal relationship between two measures. The fact that the conclusions based on both statistics were the same is indicative of the robustness of the ordinal relationship between the preference measurement methodologies investigated in this study.

Before discussing the results of the analysis using Pearson's $r$ as a measure of the correlation between the preference scores obtained using the methodologies, it is instructive to compare the estimates of Spearman's $\rho$ obtained in this study with those obtained in the published health status preference measurement literature. For the purpose of this comparison, only...
the estimate at the level of all eight multiattribute health states is appropriate, since the studies in the published health status preference measurement literature have reported results at this level only. Further, the results of the published studies only pertain to comparisons of rating scale and standard gamble scaling methods, since there have been no published applications of DCCM to the measurement of health status preferences. Table 77 summarizes the results of this comparison.

The estimates of rank-order correlation between the CUF-RS and CUF-SG methodologies obtained in this study were greater than those obtained between holistic RS and SG methodologies in the published literature. There are two possible reasons for this: (a) both the published studies reported in Table 77 used respondents' own health status as the health state to be evaluated -- it is possible that differential attitudes toward risk might differentially affect evaluations of respondents' own health status by the SG, and (b) there may not be strict comparability between the results of this study and those of the published studies reported in Table 77, since the present study used the CUF measurement strategy, while the published studies reported in Table 77 used a holistic preference measurement strategy. It should also be noted that both the studies presented in Table 77 used only one health state, while the present study used eight health states for evaluation. Still, it can be concluded that the rank-order correlation between the CUF-RS and CUF-SG methodologies in this study was acceptably high.
Table 77: Estimates of Spearman’s ρ Between Rating Scale and Standard Gamble Scaling Methods Reported in the Published Health Status Preference Measurement Literature

<table>
<thead>
<tr>
<th>Subjects</th>
<th>n</th>
<th>Health States</th>
<th>Spearman’s ρ</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chronic Stable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Angina(^a)</td>
<td>220</td>
<td>1</td>
<td>0.31</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Renal Failure(^b)</td>
<td>58</td>
<td>1</td>
<td>0.14</td>
<td>nr(^c)</td>
</tr>
<tr>
<td><strong>Hypertension(^d)</strong></td>
<td>139</td>
<td>8</td>
<td>1.00</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Legend:

b: Hornberger, Redelmeier, and Peterson (1992)
c: not reported
d: present study
Considering the results of the linear correlation between the actual preference scores obtained using the three preference measurement methodologies (measured using Pearson's $r$), the results of the analyses at all three levels generally supported the conclusions of the analyses using Kendall's $\tau$ and Spearman's $\rho$, as described above. Therefore, it may be concluded that the relationship between the preference scores obtained using the three methodologies investigated in this study is invariant to the level of measurement at which the data are considered -- ordinal or interval. The explanation provided in the section which discussed the pattern of results of the analysis using Kendall's $\tau$ is applicable in the case of the results of the analysis using Pearson's $r$ also; therefore, it will not be repeated here. It should, however, be noted that in all cases, the value of the Pearson $r$ was above 0.90 -- therefore, it is possible that the small sample size of the analysis at the level of the five corner states only may have been influential in determining the (non)significance (in a statistical sense) of the observed results.

As in the case of Spearman's $\rho$, it is instructive to compare the estimates of Pearson's $r$ obtained in this study with those obtained in the published health status preference measurement literature. For the purpose of this comparison, only the estimate at the level of all eight multiattribute health states is appropriate, since the studies in the published health status preference measurement literature have reported results at this level only. Further, the results of the published studies only pertain to comparisons of rating scale and standard gamble scaling methods, since there have been no published applications of DCCM to the measurement of health status.
preference measurement. Table 78 summarizes the results of this comparison.

It can be seen that the estimates of linear correlation between the CUF-RS and CUF-SG methodologies obtained in this study were greater than those obtained between holistic RS and SG methodologies in the published literature. Again, it should be noted that there might not be strict comparability between the results of the present study and those of the published studies reported in Table 78, since the present study used the CUF measurement strategy, while the published studies reported in Table 78 used a holistic preference measurement strategy. Still, it can be concluded that the linear correlation between the CUF-RS and CUF-SG methodologies in this study was acceptably high.

In summary, the results of this aspect of the comparison of the preference scores obtained using the three preference measurement methodologies investigated in this study were generally supportive of the previous results discussed in this section. Specifically, the analysis at the level of all eight multiattribute health states showed that there was significant correlation between the preference scores obtained using the three measurement methodologies investigated in this study, both when measured using Spearman's \( \rho \) as well as Pearson's \( r \). Moreover, the occurrence of the phenomenon of preference reversal was associated with a lower value of Spearman's \( \rho \) between the CUF-RS and DCCM as well as the CUF-SG and DCCM methodologies as opposed to the CUF-RS and CUF-SG methodologies in the level of the analysis of the five corner states only. This
provides one with some degree of confidence that the observed results are not dependent on the specific statistical measure used to address the research question. With respect to the particular research hypotheses forwarded, the pattern of results was similar to that obtained using Kendall's $\tau$; it was argued that the same explanation forwarded when those results were discussed was applicable in these cases also.

(iv) Concordances

The fourth aspect on which the different preference measurement methodologies were compared was in terms of the concordance among the preference scores of multiattribute health states obtained using the different methodologies. Considering all three methodologies together, it was found that there was significant concordance among the preference scores obtained using the methodologies (as measured by the overall ICC). Further, this result held at both the levels of the analysis on which statistical significance testing was conducted, i.e., at the levels of all eight multiattribute health states and the five corner states only. When the methodologies were considered two at a time, it was found that there was significant concordance between each pair of methodologies in the analysis at the level of all eight multiattribute health states. However, in the analysis at the level of the five corner states only, it was found that there was significant concordance between only the CUF-RS and CUF-SG as well as between the CUF-RS and DCCM methodologies.
Table 78: Estimates of Pearson's r Between Rating Scale and Standard Gamble Scaling Methods Reported in the Published Health Status Preference Measurement Literature

<table>
<thead>
<tr>
<th>Subjects</th>
<th>n</th>
<th>Health States</th>
<th>Pearson's r</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physicians*</td>
<td>49</td>
<td>5</td>
<td>0.56</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Volunteers*</td>
<td>63</td>
<td>4</td>
<td>0.52</td>
<td>ns'</td>
</tr>
<tr>
<td>Hypertension*</td>
<td>139</td>
<td>8</td>
<td>0.99</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Legend:
a: Read, Quinn, Berwick, et al. (1984)
b: Richardson, Hall, and Salkeld (1990)
c: not significant at the 0.05 level
d: present study
In order to explain this pattern of results, it is instructive to compare them to those comparing the magnitude of the preference scores obtained using each of the three methodologies. Referring to Tables 39 and 47 through 49, it can be seen that there is close correspondence between these sets of results for all levels of the analysis. The only exception to this observation is the comparison between the CUF-RS and DCCM methodologies for the analysis at the level of all eight multiattribute health states -- the results of the repeated measures analysis of variance found no significant differences between the preference scores obtained using the two methodologies at this level of the analysis, while there was significant concordance between the methodologies at this level of the analysis. It should be remembered that this was the same result which was not hypothesized for the repeated measures analysis of variance (as explained earlier in this chapter).

Taken together, the results of the repeated measures analysis of variance and ICC provide synergistic information about the sources of variance in preference scores obtained using the different measurement methodologies investigated in this study. On account of the mixed model (or Model III) formulation of the ICC that was adopted for the purpose of this study, it is possible to identify more sources of variance than would be possible in the regular one-way model (or Model I) formulation of the ICC (as explained in Chapter V). Specifically, it is possible to parcel out the effect of between-rater (between-methodology, in the case of this study) variance. In other words, the ICC would provide a measure of concordance between the measurement methodologies of the preference scores for the alternative health states after removing any systematic differences (due to theoretical
considerations, such as utility versus value, or judgment versus choice) between the preference scores obtained using the methodologies in question. It should be noted that the existence of these systematic differences is tested by the repeated measures analysis of variance. Thus, it can be seen that the repeated measures analysis of variance and the ICC provide complimentary information -- the repeated measures analysis of variance tests whether there are systematic differences among the preference scores for alternative health states obtained using the different preference measurement methodologies, while the ICC measures the concordance among the preference measurement methodologies for preference scores of alternative health states after accounting for any systematic differences among the preference scores.

This line of argument is in accordance with Bartko’s (1966) conceptualization of the ICC as a measure of inter-rater reliability, where the ICC adjusts for the inherent differences in ratings provided by raters (e.g., if one rater is consistently lower than others simply because he or she is more conservative in his or her estimates; therefore, a 40% score provided by this rater would be equivalent to a 80% score provided by a more liberal rater). This logic is applied to the case of health status preference measurement, where the utilities would be systematically higher than the values; still, after adjusting for these systematic differences, the ICC can measure the extent to which the different methodologies are concordant in terms of the scores provided to alternative states of health.

Adopting this line of argument, the results of the repeated measures analysis of variance and the ICC provide empirical evidence that although the hypothesized theoretical differences among the preference scores for
alternative health states are -- for the most part, with the one exception of the difference between the CUF-RS and DCCM methodologies at the level of all eight multiattribute health states, as explained earlier -- supported by the data, the three preference measurement methodologies are still measuring the same construct, i.e., health status preferences.

In this regard, the ICC can be viewed as an indicator of the reliability of the methodologies as measures of health status preferences. Once again, Bartko’s (1966) discussion of the ICC as a measure of inter-rater reliability is applicable to the present situation; the only difference in this study is that the ICC would be used as a measure of inter-methodology reliability. Further, since the ICC is a measure of the correlation among variables after accounting for systematic differences among the variables, it provides a measure of the consistency of measurement of a particular construct (health status preferences in the present case) across measurement situations (the three methodologies in the present case). Since the results show that the overall ICC is significant, it may be concluded that the three methodologies are measuring the same construct.

The theoretical linkages between this line of argument and the framework of subjective measurement presented in Figure 3 should be noted. The line of argument presented in this section assumes that there is one underlying psychological continuum of health status preferences, and that each of the different preference measurement methodologies are reflective of this continuum. Further, the different scores obtained as a result of the different measurement methodologies are due to the different assumptions about other influences on health status preferences that are
encoded in the preference scores obtained using the different measurement methodologies, e.g., the preference scores obtained using the CUF-SG methodology are assumed to incorporate the respondent's attitude toward risk. However, since each of the measurement methodologies are intended to measure the same construct, they fit into the framework of subjective measurement presented in Figure 3. In the case of the ICC, the test of significance involves determining the probability that each of the measures being tested quantifies the same construct. Thus, the ICC can be seen as a measure of the extent to which the preference measurement methodologies are all measuring the same construct i.e. inter-methodology reliability.

In summary, the results of this aspect of the comparison of the preference measurement methodologies showed that all three methodologies investigated in this dissertation were measuring the same construct. This was interpreted as empirical evidence of inter-methodology reliability. These results, coupled with those of the repeated measures analysis of variance (discussed earlier), were interpreted as providing synergistic information about the sources of variance in preference scores of alternative health states. Specifically, it was argued that although the theoretically expected (and hypothesized) differences among the preference measurement methodologies existed in practice, there was still concordance among the methodologies in terms of the measurement of the same construct. The next set of research questions addressed the question of what this construct was that the three methodologies were measuring.
As noted earlier, functional relationships were hypothesized and tested between each of the preference measurement methodologies. The data supported the hypotheses in the case of the analysis at the level of all eight multiattribute health states, i.e., the relationship between value and utility functions -- i.e., the CUF-RS and CUF-SG as well as the DCCM and CUF-SG methodologies -- was well-represented by a power curve and the relationship between the value functions -- i.e., the CUF-RS and DCCM methodologies -- was well-represented by a linear function (NOTE: the term “well-represented” is used on the basis of the observed p-values and adjusted $R^2$ values for each of the regression equations in question). However, in the case of the analysis at the level of the five corner states only, only the hypothesized functional relationships between the value and utility functions -- i.e., the CUF-RS and CUF-SG as well as the DCCM and CUF-SG methodologies -- were supported by the data.

The fact that the functional relationships between the value and utility functions were supported at both the levels of the analysis provides further empirical support for the veracity of the value versus utility distinction in the context of health status preference measurement. This is consistent with the literature reviewed in Chapter IV. Therefore, the results of this study do not lend empirical support to von Winterfeldt and Edwards’ (1986) hypothesis that the distinction between value and utility preference functions may be spurious -- at least in the context of measuring preferences for health status.
Although the hypothesized functional relationship between the CUF-RS and DCCM methodologies for the analysis at the level of the five corner states only was not statistically significant, it should be noted that the adjusted $R^2$ value was still above 0.75, indicating a reasonably good fit. It is possible that the small sample size for this level of the analysis rather than a poor fit of the hypothesized functional relationship might have been responsible for the lack of statistical significance -- especially since the analysis at the level of all eight multiattribute health states was statistically significant ($p<0.001$). Therefore, in spite of the observed lack of statistical significance, it is argued that the functional relationship between the CUF-RS and DCCM methodologies can be represented by a linear function.

It is also instructive to compare the functional relationship between the CUF-RS and CUF-SG methodologies to that obtained by similar studies in the published health status preference measurement literature (which were reviewed in Chapter IV). In this regard, the only level of interest for the analysis is that of all eight multiattribute health states, since none of the previous published studies reviewed here reported the results separately for the corner and interior states; indeed, the published studies only reported the results for all states together. It should be noted that only the functional relationship between the CUF-RS and CUF-SG methodologies are compared to those obtained in similar studies in the published health status preference measurement literature since there have been no published applications of DCCM to the measurement of health status preferences.
Table 79: Functional Relationships Between Rating Scale and Standard Gamble Scaling Methods Reported in the Published Health Status Preference Measurement Literature

<table>
<thead>
<tr>
<th>Subjects</th>
<th>n</th>
<th>Health States</th>
<th>$x^a$</th>
<th>$R^2b$</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parents of Cancer Patients</td>
<td>194</td>
<td>4</td>
<td>2.29</td>
<td>0.70</td>
<td>nr$^a$</td>
</tr>
<tr>
<td>Population$^e$</td>
<td>43</td>
<td>6</td>
<td>1.61</td>
<td>nr$^d$</td>
<td>nr$^d$</td>
</tr>
<tr>
<td>Population$^f$</td>
<td>63</td>
<td>4</td>
<td>0.40</td>
<td>0.29</td>
<td>nr$^d$</td>
</tr>
<tr>
<td>Stroke$^g$</td>
<td>52</td>
<td>35$^b$</td>
<td>2.16</td>
<td>nr$^d$</td>
<td>nr$^d$</td>
</tr>
<tr>
<td>Hypertension$^i$</td>
<td>139</td>
<td>8</td>
<td>1.82</td>
<td>0.98</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Legend:

a: parameter to be estimated, according to the model
(1 - Rating Scale) = (1 - Standard Gamble)$^i$
b: only the present study has reported adjusted $R^2$ values; the other $R^2$ values shown are not adjusted for the small sample sizes
c: Torrance, Zhang, Feeny, et al. (1992)
d: not reported
e: Torrance (1976a)
f: Richardson, Hall, and Salkeld (1990)
g: Wolfson, Sinclair, Bombardier, et al. (1982)
h: used health status attribute levels, not health states for the analysis
i: present study
A few points need to be noted about Table 79. First, none of the published studies shown in Table 79 reported both the R² values and the p-values. Regarding the R² values, this non-reporting is possibly due to the fact that the functional relationships between the value and utility functions were not estimated as part of the original articles in both cases (as explained in Chapter IV). The subsequent articles on these studies which did report the functional relationships were mainly concerned with the parameter estimates, for comparison purposes with the parameter estimate in other studies. In other words, it is possible that the R² values were not reported for the functional relationships in these two studies simply because the goodness of fit of the functional relationship was not of immediate concern to the authors of the article. Still, it should be noted that it does little good to compare parameter estimates of different functional relationships if the relationships themselves show poor fit to the data. Regarding the non-reporting of the p-values, it is possible that this was because of the small sample sizes (in terms of the number of health states used as data points for the estimation of the functional relationship) for the studies. This would also suggest that more emphasis should be placed on the adjusted R² value in the present study for the interpretation of the results of the analysis at the level of the five corner states only, as explained earlier.

Second, it should be noted that although the Wolfson, Sinclair, Bombardier, et al. (1982) study used 35 data points to estimate the functional relationship between value and utility functions, this was based on the preference scores of levels of health status attributes, and not health states. As explained in Chapter IV, the use of levels for analysis purposes might be
misleading since the levels might combine differently in producing an overall preference score for an alternative [as explained by change of process theory (Mellers, Ordonez, and Birnbaum, 1992)].

Third, with the exception of the Richardson, Hall, and Salkeld (1990) study -- which had a very poor fit of the functional relationship -- the parameter estimates of the functional relationship between value and utility functions range between 1.6 and 2.3. As explained in Chapter IV, these parameter estimates can be interpreted as measuring the attitude toward risk of the respondents. Specifically, for the studies shown in Table 79, this parameter can be interpreted as the extent of risk-aversiveness of the respondents of the study in question. The differences in the parameter estimates can be explained on account of different sample characteristics, measurement methodologies, or operationalizations of health status.

For instance, it can be seen in Table 79 that the published studies used different subjects as respondents. In this regard, it is possible that the differences in the parameter estimates are simply due the differences in risk-aversiveness of the respondents per se. Further, all the published studies shown in Table 79 with the exception of the Torrance, Zhang, Feeny, et al. (1992) study used the holistic preference measurement strategy to measure health status preferences, unlike the present study, which used the CUF strategy for this purpose [the Torrance, Zhang, Feeny, et al. (1992) study also used the CUF strategy]. In this regard, it is possible that the specifics of different preference measurement methodologies interact with respondents' attitude toward risk to produce different parameter estimates. For instance, the act of forcing respondents to consider each attribute separately in the
CUF strategy might be associated with different preference scores compared to the holistic strategy simply because respondents may think more deliberately about each attribute in question. Moreover, each of the studies shown in Table 79 used a different operationalization of health status. In this regard, it is possible that the different operationalizations of health status might interact with respondents' attitude toward risk to produce different parameter estimates. For instance, the selection and phrasing of attribute levels might produce positive or negative frames, which have been shown to produce different preference scores for normatively equivalent alternatives (see Kahneman and Tversky, 1979).

The parameter estimate of the present study falls in the middle of the range of the estimates in the published health status preference measurement literature. Thus, it may be concluded that the risk-aversiveness of respondents of the present study for health status as operationalized by the EuroQol and measured using the CUF strategy was less than that of respondents whose health status preferences were operationalized using the McMaster Mark II Health Utilities Index and measured using the CUF strategy (Torrance, Zhang, Feeny, et al., 1992), but more than that of respondents whose health status preferences were operationalized using holistic scenarios which were not based on any specific health status measurement instrument reviewed in Chapter I and measured using a holistic preference measurement strategy (Torrance, 1976a). Still, it is noteworthy that there was agreement among all the studies shown in Table 79 in terms of the general nature of the functional relationship between the value and utility functions, i.e., a power-curve relationship. Further, it can be
seen that the fit of the functional relationship between the CUF-RS and CUF-SG methodologies was the most satisfactory in the present study compared to the other published studies in the health status preference measurement literature, in terms of adjusted $R^2$ values and p-values.

In summary, the data supported the existence of a power curve relationship between value and utility functions — i.e., the CUF-RS and CUF-SG as well as the DCCM and CUF-SG methodologies — and a linear relationship between the value functions — i.e., the CUF-RS and DCCM methodologies. This is consistent with the theoretical and empirical research reviewed in Chapter IV, and is therefore evidence of the construct validity of DCCM as a measure of health status preferences, according to the principle of convergent validation (it was explained earlier — in Chapter II — that the estimation of functional relationships comes closest in spirit to true convergence of preference scores among the methodologies since the functional relationships can be used as conversion formulae to convert scores obtained using one methodology to those using another).

(b) The Issue of Replication of Results Comparing the CUF-RS and CUF-SG

As noted earlier, a necessary prerequisite for the construct validation of DCCM as a measure of health status preferences was the replication of the relationship between the methodologies used as comparators according to the principle of convergent validation. In this regard, it was necessary to replicate the relationship reported in the literature between the CUF-RS and
CUF-SG methodologies. The issue of replication can be addressed on the basis of the results discussed in the previous section.

The results discussed in the previous section showed: (a) the CUF-RS and CUF-SG methodologies showed perfect agreement in terms of preferential rank ordering of alternative health states, (b) the preference scores obtained using the CUF-SG methodology were significantly higher than those obtained using the CUF-RS methodology, (c) the preference scores obtained using the CUF-RS and CUF-SG methodologies were significantly correlated in terms of both rank-order as well as linear correlations, (d) there was significant concordance between the preference scores obtained using the CUF-RS and CUF-SG methodologies, and (e) a power-curve function adequately (in terms of adjusted $R^2$ values and p-values) described the functional relationship between the preference scores obtained using the CUF-RS and CUF-SG methodologies.

As discussed in Chapter IV and shown in Tables 77 through 79, these results are consistent with those reported in the published health status preference measurement literature. Indeed, in some cases, the results of this study are suggestive of greater strengths of some hypothesized relationships than those reported in the published health status preference measurement literature, e.g., correlations and functional relationships. Therefore, it may be concluded that the relationship between the CUF-RS and CUF-SG methodologies reported in the published literature was replicated in the present study. Thus, the comparators which were used for the purpose of the convergent validation of DCCM were "established."
(c) The Issue of the Construct Validity of DCCM

It was noted at the beginning of this section that the need for comparing preference scores of the alternative health states obtained as a result of the different preference measurement methodologies was prompted by the principle of convergent validation as a means for validating the DCCM as a measure of health status preferences. Therefore, it is instructive to consider how each of the five bases of comparison contributes to this overall goal of the comparison of preference scores. This issue is discussed with respect to the analyses at the levels of all eight multiattribute health states and the five corner states. It was decided against discussing the results of the analysis at the level of the three interior states only on account of the small sample size at this level of the analysis.

The results of the analysis at the level of all eight multiattribute health states were generally supportive of the hypotheses forwarded in Chapter II. There was significant agreement among the three preference measurement methodologies in terms of the order in which they ranked the eight multiattribute health states -- this result held true when all three methodologies were considered together and also when the methodologies were considered two at a time. Therefore, at this level of the analysis, the assumption of procedural invariance appeared to hold in practice, since there was no significant occurrence of the phenomenon of preference reversal.

As hypothesized, there were significant differences in the magnitude of preference scores ascribed to the eight multiattribute health states by the
three preference measurement methodologies investigated in this study. Specifically, the preference scores obtained using the CUF-SG methodology were significantly higher than the preference scores obtained using both the CUF-RS and the DCCM methodologies at this level of the analysis. This result reinforces the veracity of the value versus utility distinction in the context of health status preference measurement, as reviewed in Chapter IV.

However, the hypothesis that the preference scores for the eight multiattribute health states obtained using the DCCM would be higher than those obtained using the CUF-RS methodology was not supported by the data. It is interesting to note that when the repeated measures analysis of variance was run without including the multiattribute health state E8 (see Appendix B for a description of this health state), a significant difference was found between these preferences scores in the hypothesized direction. The other results of the repeated measures analysis of variance on all eight multiattribute health states remained unchanged when the analysis was run without including the health state E8. It is important to note the reason why the analysis was run without the health state E8. This was simply because this particular health state was the only one for which the CUF-RS score was greater in magnitude than the DCCM score (as seen in Table 34). Therefore, it is possible that there may have been some error in measurement of the preferences for that particular health state.

The other hypotheses forwarded in Chapter II were supported by the results of the analysis at the level of all eight multiattribute health states. The preference scores of the eight multiattribute health states obtained using the three methodologies were significantly correlated with each other, both in
terms of rank-order correlations as well as in terms of the magnitude of the preference scores. As explained in Chapter II, these results were expected since the three methodologies were intended to measure the same construct, i.e., health status preferences. Further, there was a significant degree of concordance among the preference scores of the eight multiattribute health states obtained using the three methodologies -- when considering all three methodologies together as well as when considering the methodologies two at a time. This is an indication that after adjusting for any theoretical differences that might be expected between the preference scores, there is still agreement in terms of the measurement of the same construct. This conclusion was further supported by the results of the functional relationships which were fit to the data. In this regard, the hypothesized functional relationships between the different preference functions were supported by the data -- i.e., a power curve described the relationship between the CUF-RS and CUF-SG as well as between the DCCM and CUF-SG methodologies, while a linear function described the relationship between the DCCM and CUF-RS methodologies. This is consistent with the theoretical and empirical literature on the expected relationships between value and utility functions intended to measure the same construct (as reviewed in Chapter IV). Taken as a whole, the results of the analysis at the level of the eight multiattribute health states of the comparison of the preference scores of the eight multiattribute health states provide empirical support for the construct validity of DCCM as a measure of health status preferences, according to the principle of convergent validation.
The results of the analysis at the level of the five corner states only were not as unequivocal as those of the analysis at the level of all eight multiattribute health states, as discussed above. Although there was significant agreement in terms of preferential rank ordering of the five corner states by the three methodologies when all the methodologies were considered together, there were significant differences in rank ordering of the five corner states when the methodologies were considered two at a time. Therefore, at this level of the analysis, the assumption of procedural invariance was violated in practice, since the phenomenon of preference reversal was observed between the rank ordering of the five corner states between the CUF-RS and DCCM and between the CUF-SG and DCCM methodologies. Thus, the judgment versus choice distinction appeared to have some empirical support at this level of the analysis.

As hypothesized, there were significant differences in the magnitude of preference scores ascribed to the eight multiattribute health states by the three preference measurement methodologies investigated in this study. Specifically, the preference scores obtained using the CUF-SG methodology were significantly higher than the preference scores obtained using the CUF-RS methodology at this level of the analysis. This result is consistent with the literature reviewed in Chapter IV, and also with the results of the analysis at the level of all eight multiattribute health states, as discussed earlier.

However, unlike the results of the analysis at the level of all eight multiattribute health states, there were no significant differences found between the preference scores of the five corner states obtained using the CUF-SG and DCCM methodologies, while there were significant differences
found between the preference scores of the five corner states obtained using
the CUF-RS and DCCM methodologies. Although the latter result was
hypothesized, the former was not. The differences between the results at the
levels of all eight multiattribute health states and the five corner states only
were explained in terms of the structural differences between the corner and
interior states.

Although there was significant correlation between the CUF-RS and
CUF-SG methodologies, this was not so for the case between the CUF-RS
and DCCM as well as the CUF-SG and DCCM methodologies. These results
are consistent with those of the preferential rank ordering of the five corner
states by the preference measurement methodologies taken two at a time. In
terms of concordance coefficients, there was significant concordance among
all three methodologies when they were considered together, but only
between the CUF-RS and CUF-SG as well as between the CUF-RS and
DCCM methodologies when the methodologies were considered two at a
time. Finally, the hypothesized functional relationships were supported
between the CUF-RS and CUF-SG as well as between the CUF-SG and
DCCM methodologies, but not between the CUF-RS and DCCM
methodologies. It should be noted that the linear relationship that was
hypothesized between the CUF-RS and DCCM methodologies had a
reasonably high adjusted R^2 value (over 0.75), and that the lack of statistical
significance might have been on account to the small sample size in terms of
the number of data points used for the regression analysis. Taken as a whole,
these results provide mixed support for the construct validity of the DCCM
as a measure of health status by the principle of convergent validation at this level of the analysis.

The results of the analysis at these two levels, as discussed above, appear to be in conflict with each other. In the analysis at the level of all eight multiattribute health states, the construct validity of the DCCM as a measure of health status preferences was strongly supported; however in the analysis at the level of the five corner states only, the construct validity of the DCCM as a measure of health status preferences received mixed support. In the face of such equivocal results, it is necessary to consider the nature of the corner and interior states, and understand why the different nature of these health states may explain the apparently conflicting results at the two levels of analyses.

In this regard, the explanations provided in the discussion of the results of the analytical procedures using Kendall's \( \tau \), the repeated measures analysis of variance, and the ICC provide insight into how the differences in the nature of corner and interior states could lead to results which run counter to the hypothesized relationships between preference measurement methodologies. The important question that needs to be addressed is to what extent do these "unhypothesized" results argue against the construct validity of DCCM as a measure of health status preferences.

It should be remembered that the hypotheses forwarded in Chapter II were based on a review of the published theoretical and empirical literature. As noted earlier in this chapter, the published literature has not maintained
a distinction between corner and interior states, at least in terms of inferences that may be drawn from the data. Thus, most of the empirical evidence published in the literature is usually based on analyses conducted using interior states as the alternatives to be evaluated. Further, when corner states have been included in the analyses, they are usually combined with the interior states, i.e., no separate results are reported for the analyses of the corner versus interior states [see, for e.g., Torrance, Zhang, Feeny, et al. (1992)]. Therefore, strictly speaking, the hypotheses forwarded in Chapter II may be applicable only to the analysis at the level of all eight multiattribute health states. It should be noted that the results of the analysis at this level generally supported the hypotheses forwarded in Chapter II, and were therefore generally consistent with the published theoretical and empirical evidence.

Although the hypotheses forwarded in Chapter II may not strictly apply to the analysis at the level of the five corner states only, it is still noteworthy that the overall ICC (among the three preference measurement methodologies all considered together) was significant, and that the hypothesized functional relationships were supported by the data. As noted earlier, the overall ICC can be interpreted as a measure of inter-methodology reliability, while the functional relationships come closest in spirit to the notion of true convergence of preference scores obtained using the different measurement methodologies. Therefore, the results at this level of the analysis show that although the phenomenon of preference reversal may be operative in practice, the three different preference measurement
methodologies investigated in this study still measured the same construct, i.e., health status preferences.

Moreover, the small number of data points (i.e., five) for the analysis at the level of the five corner states only deserves comment. In this regard, it should be noted that it is possible that the observed nonsignificance (in a statistical sense) of some of the results of this level of the analysis might have been a function of the small number of data points, and consequently of less amount of statistical power to test the research hypotheses. This would suggest that greater emphasis be placed on the value of the test statistic, as opposed to the observed p-value. Viewed in this way, it can be seen that the Pearson’s r values were all above 0.90, and that the pairwise ICC values were above 0.80. Further, the nonsignificant (in a statistical sense) functional relationship between the CUF-RS and DCCM methodologies had an adjusted $R^2$ value of above 0.75. Theses values are all quite high, and suggest that the observed nonsignificance (in a statistical sense) of the results of the analysis at the level of the five corner states only might reflect the effect of the small number of data points as opposed to a practically significant effect.

In this regard, the role of the Bonferroni adjustment should also be noted. For instance, the observed p-value of the Spearman’s ρ statistic between the preference scores obtained using the CUF-RS and CUF-SG methodologies (which had a value of 1.00, indicating perfect agreement) was 0.008, which — strictly speaking — is not statistically significant at the corrected p-value of 0.007 adopted after the Bonferroni correction was made to adjust for multiple comparisons. Therefore, in this case even perfect agreement between measures will not be statistically significant. This again
suggests caution in using observed p-values as the sole criterion for interpretation of the results of the analysis at the level of the five corner states only.

Therefore, it can be seen that although the results of the analysis at the level of the five corner states only did not support all of the hypotheses forwarded in Chapter II, this does not necessarily mean that they did not support the construct validity of DCCM as a measure of health status preferences, according to the principle of convergent validation. Indeed, the results of the overall ICC and functional relationships along with the effects of the small number of data points and the Bonferroni adjustment suggest that the construct validity of the DCCM as a measure of health status preferences may be supported at this level of the analysis, according to the principle of convergent validation.

In summary, the results of the comparison of preference scores obtained using the different preference measurement methodologies investigated in this study generally supported the construct validity of the DCCM as a measure of health status preferences when the analysis was conducted using all eight multiattribute health states together. Although more tentative than the analysis at the level of all eight multiattribute health states, the results of the analysis at the level of the five corner states only also supported this conclusion.
Comparing Respondent Evaluations

The last four research questions addressed the comparison of respondent evaluations of the three preference measurement methodologies investigated in this dissertation. It was necessary to determine the psychometric properties of the data collected on respondent evaluations before comparing the evaluations. This section will discuss the results of psychometric properties of the data collected as well as the comparison of the respondent evaluations of the preference measurement methodologies.

(a) Reliability

The reliability of each set of respondent evaluations was estimated using Cronbach's alpha (Cronbach, 1951). The alpha coefficients for the CUF-RS, CUF-SG, and DCCM datasets were 0.84, 0.84, and 0.89 respectively (see Table 53). In discussing the interpretation of these coefficients, the published guidelines and empirical evidence will be reviewed; accordingly, a decision will be made regarding the acceptability of each of these coefficients as estimates of the reliability of the data collected on respondent evaluations of the three preference measurement methodologies investigated in this dissertation. This discussion draws heavily from Peterson's (1994) meta-analysis of Cronbach's alpha.

There have been several guidelines published on recommended minimally acceptable reliability coefficients (see, for e.g., Davis, 1964; Kaplan and Saccuzzo, 1982; Murphy and Davidshofer, 1988; Nunnally, 1967; Nunnally 1978). Table 80 presents some selected recommendations of minimally acceptable reliability coefficients in the published literature.
These recommendations were selected because they have been cited frequently in studies and are considered by Peterson (1994) to be "illustrative" of the recommendations regarding minimally acceptable reliability coefficients.

It can be seen from Table 80 that the minimally acceptable reliability coefficient is a function of the purpose of the research, with lower values of reliability coefficients being deemed acceptable for research in the developmental stages. It should be noted that the particular purpose of this research can be classified as "preliminary," since it was concerned with the development of an instrument to measure respondent evaluations of preference measurement methodologies. In this regard, the minimally acceptable reliability coefficient across most of the recommendations shown in Table 80 is 0.70. Since each of the reliability coefficients estimated in this study was above this value, they can be considered acceptable according to these recommended criteria.

However, Peterson (1994) has noted that none of these recommendations has "... an empirical basis, a theoretical justification, or an analytical rationale. Rather, they appear to reflect either 'experience' or intuition" (p.381). Moreover, Guilford (1965) has noted that different reliability coefficients are not conceptually or mathematically comparable -- therefore, the usefulness of blanket recommendations regarding minimally acceptable reliability coefficients may be questionable. Therefore, it may be more instructive to compare the estimated reliability coefficients obtained for the data sets in this study to those reported specifically for Cronbach's alpha in similar studies in the published literature.
Table 80: Selected Recommended Minimally Acceptable Reliability Coefficients

<table>
<thead>
<tr>
<th>Source</th>
<th>Situation</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaplan and Saccuzzo</td>
<td>Basic Research</td>
<td>0.7-0.8</td>
</tr>
<tr>
<td>(1982, p.106)</td>
<td>Applied Research</td>
<td>0.95</td>
</tr>
<tr>
<td>Murphy and Davidshofer</td>
<td>Unacceptable Level</td>
<td>&lt; 0.6</td>
</tr>
<tr>
<td>(1988, p.89)</td>
<td>Low Level</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>Moderate to High Level</td>
<td>0.8-0.9</td>
</tr>
<tr>
<td></td>
<td>High Level</td>
<td>0.9</td>
</tr>
<tr>
<td>Nunnally</td>
<td>Preliminary Research</td>
<td>0.5-0.6</td>
</tr>
<tr>
<td>(1967, p.226)</td>
<td>Basic Research</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Applied Research</td>
<td>0.9-0.95</td>
</tr>
<tr>
<td>Nunnally*</td>
<td>Preliminary Research</td>
<td>0.7</td>
</tr>
<tr>
<td>(1978, pp.245-246)</td>
<td>Basic Research</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>Applied Research</td>
<td>0.9-0.95</td>
</tr>
</tbody>
</table>

Legend:
a: it is interesting to note that Nunnally changed the recommended minimally acceptable reliability coefficient in the second edition of his classic text *Psychometric Theory*. No explanation was offered for this change.
In this regard, Peterson's (1994) meta-analysis of Cronbach's alpha provides very useful information for comparative purposes. Peterson (1994) reported a meta-analysis of Cronbach's alpha on the basis of studies reported in the marketing and psychology literature. The literature reviewed in this meta-analysis was quite extensive -- a census of eight psychology- and marketing-related journals (from 1960 to 1992), a convenience sample of 16 other journals and conference proceedings, and a sample of unpublished manuscripts.

Peterson (1994) classified his results on the basis of eleven research design characteristics: (a) sample size -- operationalized as less than 100, 100-199, 200-299, 300 or more, and not given, (b) type of sample -- operationalized as college students, consumers, businesspersons, mixed, and cannot tell, (c) number of response categories -- operationalized in increments of one, up to a maximum of eight and as more than eight thereafter, (d) number of items -- operationalized in increments of one, up to a maximum of 11 and as more than 11 thereafter, (e) scale type -- operationalized as Likert-type or semantic differential, (f) scale format -- operationalized as only endpoints labeled, numerical values on inner categories, verbal values on inner categories, and cannot tell, (g) nature of scale -- operationalized as odd number of item categories, even number of item categories, or cannot tell, (h) administration mode -- operationalized as self, interviewer, or not given, (i) scale orientation -- operationalized as respondent-centered, stimulus-centered, or both, (j) nature of construct -- operationalized as dependent, independent, or cannot tell/both, and (k) type
of research -- operationalized as scale development, scale application, or cannot tell.

For the purpose of comparing the results of the meta-analysis to those obtained in this study, the following research design characteristics were appropriate as comparators: (a) sample size -- 100-199, (b) type of sample -- since patients were not included in Peterson's (1994) operationalization as a separate category, it was decided to use both consumers as well as the "cannot tell" category as comparators, (c) number of response categories -- seven (see Chapter V for the details on the instrumentation), (d) number of items -- eight (see Chapter V for the details on the instrumentation), (e) scale type -- semantic differential, (f) scale format -- verbal values on inner categories (see Chapter V for details on the instrumentation), (g) nature of scale -- since an exploratory factor analysis was conducted, the decision on the number of item categories could not be made beforehand; therefore, it was decided to use both the "even number of item categories" (since there were two factors extracted in each case) as well as "cannot tell" categories as comparators, (h) administration mode -- interviewer, (i) scale orientation -- stimulus-centered, since the evaluations of preference measurement methodologies were the subject of the instrumentation, (j) nature of construct -- cannot tell/both, since the respondent evaluations were not hypothesized to be either independent or dependent variables, and (k) type of research -- scale development, since the instrumentation was developed specifically for this dissertation.
Table 81: Selected Results from Peterson's (1994) Meta-analysis

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n'</th>
<th>Mean $\alpha$</th>
<th>Median $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100-199</td>
<td>1,169</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>Type of Sample:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consumers</td>
<td>879</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>Cannot Tell</td>
<td>86</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>Response Categories:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seven</td>
<td>991</td>
<td>0.78</td>
<td>0.82</td>
</tr>
<tr>
<td>Number of Items:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eight</td>
<td>183</td>
<td>0.73</td>
<td>0.77</td>
</tr>
<tr>
<td>Scale Type:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Differential</td>
<td>372</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>Scale Format:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal Values on</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inner Categories</td>
<td>1,869</td>
<td>0.77</td>
<td>0.80</td>
</tr>
<tr>
<td>Nature of Scale:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Even Number</td>
<td>863</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>Cannot Tell</td>
<td>667</td>
<td>0.76</td>
<td>0.78</td>
</tr>
</tbody>
</table>
Table 81 (continued)

<table>
<thead>
<tr>
<th>Administration Mode: Interviewer</th>
<th>153</th>
<th>0.72</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale Orientation: Stimulus-centered</td>
<td>1,206</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td>Nature of Construct: Cannot Tell/Both</td>
<td>1,470</td>
<td>0.75</td>
<td>0.78</td>
</tr>
<tr>
<td>Type of Research: Scale Development</td>
<td>1,270</td>
<td>0.77</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Legend:

a: number of studies upon which the results are based

[based on data reported in Peterson (1994)]
The number of studies upon which the results were based, and the mean and median Cronbach’s alpha values reported by Peterson (1994) are presented in Table 81 for the research design characteristics relevant to this dissertation. It can be seen that in all cases, the estimated reliability coefficients of this study were higher than those for the appropriate comparators reported by Peterson (1994). Therefore, even on the basis of the comparison of the estimated reliability coefficients obtained this study to Cronbach’s alpha values reported in similar studies in the literature, it can be concluded that each of the sets of data on respondent evaluations of the preference measurement methodologies was acceptable in terms of reliability.

Table 53 (see Chapter VI) also provides estimates of what the value of Cronbach’s alpha would be if each of the items of the eight-item semantic differential scale were deleted. It can be seen that in almost all cases, deleting an item is associated with a decrease in the value of Cronbach’s alpha for each of the three data sets. The only exceptions are the items of length for the CUF-RS and CUF-SG methodologies, and possibly difficulty in doing for the CUF-SG methodology. Although other items may be associated with an increase in the value of Cronbach’s alpha for the data sets, these increases are rather small. In this regard, it should be noted that the items of length and difficulty in doing were selected for consideration above simply because their omission would increase the alpha value by more than 0.015 units (an admittedly arbitrary cutoff point). However, before actually deleting any items, it is important to consider the role of the items in the underlying factor structure of respondent evaluations of the preference
measurement methodologies. This is because decisions about deletion of any items should not be made solely on the basis of mechanical rules of thumb (especially arbitrary ones, as noted above). This is particularly important in this case since the associated increase in the Cronbach's alpha value is only 0.02 units, and the "base" (i.e., without deleting any items) value of Cronbach's alpha is acceptable to begin with, as discussed earlier in this section. Therefore, the decision about deletion of any items from the instrument to measure respondent evaluations of the preference measurement methodologies will be deferred until the end of the next section, which discusses the results of the exploratory factor analyses conducted on each of the data sets of respondent evaluations of preference measurement methodologies.

In summary, the results of the estimation of the reliability of the data collected on respondent evaluations of the different preference measurement methodologies was acceptable with respect to the proposed guidelines in the published literature as well as compared to values of Cronbach's alpha reported in similar studies in the published literature. The issue of retaining versus deleting items of the instrument used to measure respondent evaluations of the preference measurement methodologies was deferred until the end of the next section.

Exploratory Factor Analyses

As explained earlier, maximum likelihood exploratory factor analysis was conducted on each of the three data sets of respondent evaluations of
the preference measurement methodologies in order to determine the nature of the underlying factor structure of respondent evaluations of preference measurement methodologies. The results of each of these exploratory factor analyses will be discussed in turn in this section.

(a) Respondent Evaluations of the CUF-RS Methodology

Table 57 (see Chapter VI) presents the final rotated factor loading matrix for the respondent evaluations of the CUF-RS methodology. Although this matrix does not exhibit true simple structure, it is quite amenable to interpretation. The items on interest, sensibility, and usefulness clearly load on only the first factor. On the other hand, the items on difficulty in doing and length clearly load on only the second factor. Adopting Hair, Anderson, and Tatham's (1987) cutoff of ± 0.19 as an indication of statistical significance at the 0.05 level, the remaining three items load on both the factors. Although the inclusion of items which loaded on multiple factors was not precluded, the major emphasis in interpretation of the factor loading matrix was in terms of the pattern of factor loadings and the consequent meaning of the matrix (as noted in the previous chapter). In other words, statistical significance of individual items was only treated as a first (but necessary) prerequisite for further investigation. The final decision was made on the basis of the relation of the items under consideration to the items already decided as loading on the two factors.

In this regard, the item on success had a much higher loading on the first factor than on the second; further, it is closer in spirit to the other items which loaded on the first factor (i.e., interest, sensibility, and usefulness) in
that it is indicative of an affective-based feeling about the preference measurement methodology under evaluation. On the other hand, the item on difficulty in understanding had a higher loading on the second factor than on the first; moreover, it is closer in spirit to the other items which loaded the second factor (i.e., difficulty in doing and length) in that it is indicative of an action-based performance of or doing the tasks required of the preference measurement methodology under evaluation. The remaining item, i.e., clarity, can be interpreted as loading on both factors since it could reflect a feeling about the methodology (i.e., I-thought-it-was-very-clear) as well as an indicator of the difficulty in doing the tasks required of the methodology under evaluation (i.e., It-was-so-clear-that-I-had-no-difficulty-in-doing-it).

It should be noted that this particular interpretation of factor loadings was adopted mainly on account of its meaningfulness -- both the items of clarity and difficulty in understanding had similar factor loadings on the factors; the classification of these items as loading on a particular factor was guided by the logic outlined above. Further, it was reassuring to note that a similar interpretation was apparent in the evaluation of the other two preference measurement methodologies investigated in this study.

In summary, the results of the factor analysis of respondent evaluations of the CUF-RS methodology yielded two factors as underlying the structure of respondent evaluations, the first related to an affective-based feeling about the methodology and the second related to an action-based performance of or doing the tasks required of the methodology. The specific
items which loaded on the "feeling" factor were the items on interest, sensibility, usefulness, and success of the preference measurement methodology under evaluation, while the items which loaded on the "doing" factor were the items on length, difficulty in doing, and difficulty in understanding the preference measurement methodology under evaluation. The item on clarity loaded on both the factors.

(b) Respondent Evaluations of the CUF-SG Methodology

Tables 61 and 62 (see Chapter VI) present the rotated factor loading matrices for respondent evaluations of the two- and three-factor solutions of the exploratory factor analysis of respondent evaluations of the CUF-SG methodology. Although these rotated factor loading matrices do not truly exhibit simple structure, they are still amenable to meaningful interpretation. More importantly, they serve the purpose of deciding how many factors to extract for the final solution. In this regard, both the two- as well as three-factor solutions share some similarities in terms of the loadings of different items on factors. Specifically, the items on interest and sensibility on the one hand and the items on difficulty in doing, length, and difficulty in understanding on the other hand loaded on the same factors. Moreover, the item on clarity loaded on two factors in both solutions. It is interesting to note that the order of extraction of the factors was different in the two solutions, in that the first factor for the two-factor solution was the second factor for the three-factor solution.

However, there were also some differences between the two- and three-factor solutions. Specifically, the item on usefulness loaded on only
one factor in the two-factor solution, whereas it had statistically significant loadings on two factors in the three-factor solution. Further, the item on success loaded on the same factor as the items on interest and sensibility in the two-factor solution, but it loaded on a separate factor (along with the item on usefulness) in the three-factor solution. It should, however, be noted that the interpretation of the item on success in the three-factor solution should be mediated by the recognition that this was the item which was involved in the Heywood case.

Comparing the results of the factor loadings obtained with the two- and three-factor solutions, it can be seen that the three-factor solution does not in any way enhance the meaningfulness of the two-factor solution. On the contrary, it has worse simple structure than the two-factor solution. Further, the occurrence of the Heywood case may be responsible for the observed loading of the item on success, in which case caution should be exercised in the interpretation of the loading of this item. Most importantly, in terms of the unequivocal loadings of items on factors (i.e., items which load on only one factor), the two-factor solution is better than the three-factor solution. Therefore, it was decided to adopt the two factor solution for the purpose of interpretation of the underlying structure of respondent evaluations of the CUF-SG methodology. It should be noted that this was the same conclusion reached on the basis of the test of close fit, but not on the basis of the test of exact fit. In terms of naming the two factors, the conclusions reached for the two-factor solution of the respondent evaluations of the CUF-RS methodology are also applicable here; therefore, they will not be repeated here.
In summary, the results of the factor analysis of respondent evaluations of the CUF-SG methodology yielded two factors as underlying the structure of respondent evaluations, the first related to an affective-based feeling about the methodology and the second related to an action-based performance of or doing the tasks required of the methodology. The specific items which loaded on the “feeling” factor were the items on interest, sensibility, usefulness, and success of the preference measurement methodology under evaluation, while the items which loaded on the “doing” factor were the items on length, difficulty in doing, and difficulty in understanding the preference measurement methodology under evaluation. The item on clarity loaded on both the factors. It is noteworthy that this is exactly the same pattern of factor loadings as was observed for the respondent evaluations of the CUF-RS methodology. However, the factor loading matrix of the respondent evaluations of the CUF-SG methodology exhibited better simple structure than that of the respondent evaluations of the CUF-RS methodology.

(c) Respondent Evaluations of the DCCM Methodology

Tables 67 through 69 (see Chapter VI) present the rotated factor loading matrices of the two-, three-, and four-factor solutions of the exploratory factor analysis of respondent evaluations of the DCCM methodology. Compared to the results of the other factor analyses, the factor loading matrices of this factor analysis exhibit noticeably less simple structure. Still, they are amenable to meaningful interpretation, especially in deciding how many factors to extract for the final solution. In the discussion
which follows, the two-factor solution will first be discussed, followed by a comparison of the two-factor solution with the other solutions. On the basis of this discussion, a decision will be made regarding how many factors to extract for the final solution.

In terms of the loadings of individual items on factors, the two-factor solution of this analysis shows remarkable similarity to the other two-factor solutions discussed earlier. Like the other two-factor solutions discussed earlier, the two-factor solution of this analysis has the items of interest, sensibility, and success loading on one factor and the items of difficulty in doing and difficulty in understanding loading on the other factor. Further, in all three two-factor solutions, the item on clarity loads on both factors.

There is some similarity regarding the loading pattern of the item on usefulness also -- although this item had statistically significant loadings on both factors in this analysis, it had a considerably higher loading on the factor which had the items of interest, sensibility, and success. Therefore, this item can be thought of as loading on this particular factor -- this is again in agreement with the results of the two-factor solutions discussed earlier. The only difference between this and the previous two-factor solutions is in terms of the item on length -- in this analysis, this item loads higher on the factor which has the items of interest, sensibility, and success (also usefulness, if one classifies it in this manner, as discussed above); in the previous analyses, this item load higher on the factor which has the items of difficulty in doing and difficulty in understanding. Still, it should be noted that this particular item had statistically significant loadings on both the factors in this solution.
It is instructive to consider the reasons for this difference in loadings for the item of length in the respondent evaluations of the different preference measurement methodologies investigated in this dissertation. In this regard, it is possible that the actual time taken by respondents to complete the tasks required of them to implement the DCCM was so short that they did not remember the act of actually doing it; rather, they only remembered the fact that they did it and their feeling about it. On the other hand, the actual time taken by respondents to complete the tasks required of them to implement the CUF-RS and CUF-SG methodologies was so long that their evaluation of these methodologies on the item of length was based on the actual act of doing them rather than the feeling about how long these tasks were.

A final point to note about the two-factor solution presented in Table 67 is that the factor loading for the item on sensibility is greater than 1.0. This is, of course, not possible in theory and might be reflective of a potential problem in the direct quartimin approach used by FCAP. This notion is bolstered by the observation (presented in Table 70 in Chapter VI) that such an anomaly disappears when the promax approach to oblique rotation is applied to the analysis.

Comparing the three-factor solution to the two-factor solution, it can be seen that it does not offer much improvement in terms of interpretability. Although the item on usefulness does distinctly load on one factor only, it should be noted that the Heywood case; therefore any conclusion regarding its loading pattern needs to be treated with caution. The other major difference between the two- and three-
factor solutions is in terms of the item on success -- this item unequivocally loaded on only one factor in the two-factor solution, whereas it loaded on two factors in the three-factor solution. Apart from these two differences, the results obtained with the three-factor solution are similar to those obtained with the two-factor solution.

The comparison between the two-factor solution and the four-factor solution yields similar results. The major advantage of the four-factor solution over the two-factor solution is that the item on usefulness clearly on only one factor -- without being involved in a Heywood case. However, the item on sensibility loads on two different factors in the four-factor solution, while the item on length loads on three different factors. Further, the item on difficulty in understanding is the sole item which loads on the fourth factor -- however, this item was the one which was involved in the Heywood case in this particular solution.

Therefore, it can be seen that both the three- as well as the four-factor solutions do not in any way enhance the meaningfulness of the two-factor solution. On the contrary, they exhibit worse simple structure than the two-factor solution. Therefore, it was decided to adopt the two factor solution for the purpose of interpretation of the underlying structure of respondent evaluations of the DCCM methodology. It should be noted that this conclusion is not in keeping with the results of either the test of exact fit or the test of close fit. These tests respectively indicated that a four- and three-factor solution were most appropriate. However, as noted in Chapter V [as well as by Joreskog (1971)], the ultimate criteria on which to evaluate models are in terms of interpretability and usefulness. As explained above, adoption
of these criteria lead to the selection of the two-factor solution as the most appropriate, notwithstanding the fact that it may run the risk of not fitting closely in the population. In terms of naming the two factors, the conclusions reached for the two-factor solution of the respondent evaluations of the CUF-RS and CUF-SG methodologies are also applicable here; therefore, they will not be repeated here.

In summary, the results of the factor analysis of respondent evaluations of the DCCM methodology yielded two factors as underlying the structure of respondent evaluations, the first related to an affective-based feeling about the methodology and the second related to an action-based performance of or doing the tasks required of the methodology. The specific items which loaded on the "feeling" factor were the items on interest, sensibility, usefulness, length and success of the preference measurement methodology under evaluation, while the items which loaded on the "doing" factor were the items on difficulty in doing and difficulty in understanding the preference measurement methodology under evaluation. The item on clarity loaded on both the factors.

It is noteworthy that this is quite similar to the pattern of factor loadings as was observed for the respondent evaluations of the CUF-RS and CUF-SG methodologies. The only difference between this pattern of factor loadings and that of the respondent evaluations of the CUF-RS and CUF-SG methodologies was in terms of the item on length, as explained above. Further, it should be noted that the factor loading matrices of the respondent evaluations of the CUF-RS and CUF-SG methodologies exhibited better
simple structure than the factor loading matrix of the respondent evaluations of the DCCM.

Summarizing the results of the three exploratory factor analyses taken together, this study supported the existence of two factors as representative of the structure of respondent evaluations of health status preference measurement methodologies. It is noteworthy that each separate exploratory factor analysis on each data set supported the two-factor solution. This is suggestive of the stability of the two-factor solution across different preference measurement methodologies, and is therefore suggestive of a "global" underlying factor structure of respondent evaluations of preference measurement methodologies. Specifically, this two-factor global structure consists of a factor related to an affective-based feeling about the methodology and a second factor related to an action-based performance of or doing the tasks required of the methodology.

The correspondence between this particular interpretation and that suggested by Nutt (1980) is noteworthy, since the results of Nutt's (1980) study -- which was reviewed in Chapter IV -- suggested that the construct of respondents' overall evaluation of preference measurement methodologies consists of at least two underlying factors, one relating to a feeling about the specific purpose of the methodology and the second relating to the operational aspects of the methodology -- i.e., "feeling" and "doing" respectively.

However, the specific differences between the present study and that of Nutt (1980) also need to be noted. The major difference between these two
studies lies in the level of detail in which the construct of overall respondent evaluations of preference measurement methodologies was investigated. Specifically, the Nutt (1980) study was not designed to determine the nature of the underlying factor structure of respondent evaluations of preference measurement methodologies; consequently, only two attributes of respondent evaluations were explored in the Nutt (1980) study -- perceived accuracy and ease of use.

On the other hand, the present study was designed to determine the nature of the underlying factor structure of respondent evaluations of preference measurement methodologies; consequently, data was collected on eight different attributes hypothesized to contribute to overall respondent evaluations of preference measurement methodologies -- interest, sensibility, difficulty in doing, usefulness, length, success, clarity, and difficulty in understanding. Of these eight attributes, the results of the exploratory factor analyses indicated that the attributes of interest, sensibility, usefulness, and success loaded on the "feeling" factor only, while the attributes of difficulty in doing and difficulty in understanding loaded on the "doing" factor only. The attribute of clarity loaded on both factors. Finally, the attribute of length loaded on the "doing" factor for the respondent evaluations of the CUF-RS and CUF-SG methodologies, but on the "feeling" factor for the DCCM. This last result was explained on the basis of the relative amount of time taken by respondents to complete the tasks required of them in the case of the DCCM as opposed to the other methodologies investigated in this study (as explained above).
Having discussed the nature of the underlying factor structure of the respondent evaluations of the three different preference measurement methodologies investigated in this study, it is important to address the issue of keeping versus deleting items from the instrument used to measure respondent evaluations of the preference measurement methodologies. In the section of this chapter which discussed the reliability coefficients of the three data sets of respondent evaluations of preference measurement methodologies, it was noted that the only two items that warrant consideration for deletion were the items on difficulty in doing for the CUF-SG methodology and length for the CUF-RS and CUF-SG methodologies. The case of each of these two items will be discussed in turn.

Regarding the item of difficulty in doing, this is an important concern with the standard gamble scaling method (as reviewed in Chapter IV) and therefore needs to be included in any evaluation of a preference measurement methodology which is operationalized using the standard gamble scaling method. Further, this particular item had a high factor loading for the respondent evaluations of the CUF-SG methodology (as shown in Table 63). Therefore, it was decided to retain this item in the instrument used to measure respondent evaluations of preference measurement methodologies.

Regarding the item of length, it can be seen in Tables 57, 63, and 70 that it had moderately high factor loadings in each of the respondent evaluations. Further, it was the only item which loaded on different factors in the respondent evaluations -- it loaded on the same factor in the case of the respondent evaluations of the CUF-RS and CUF-SG methodologies,
while it loaded on the other factor in the case of the respondent evaluations of the DCCM (a possible explanation for this observed phenomenon was provided earlier in this chapter). Moreover, this was the same set of bipolar adjectives which was found to load primarily on the potency dimension -- not the evaluation dimension, like the other sets of bipolar adjectives used in this dissertation -- in the pioneering research on the semantic differential by Osgood, Suci, and Tannenbaum (1957). Still, it should be remembered that the particular respondents of this study might have had different attitudes toward the length of time taken to complete the tasks required to implement preference measurement methodologies than other respondents. This is because of the fact the most respondents of this study were veterans who were interviewed while waiting for their prescriptions to be filled in the clinic. Therefore, it is possible that they did not mind spending a long amount of time performing the tasks required of them to implement a particular preference measurement methodology. Therefore, before actually deleting the item on length from the instrument measuring respondent evaluations of the preference measurement methodologies, it is necessary to administer the instrument to another group of respondents who have different attitudes toward the length of time taken to complete the tasks required to implement preference measurement methodologies and see if the results of this study can be replicated in this other group of respondents. If the results are replicated, the item on length may be deleted from the instrument. Thus, although the item on length shows a variable pattern of loadings across the respondent evaluations of the different preference measurement methodologies investigated in this dissertation, it was decided
against deleting this item from the instrument used to measure respondent evaluations of preference measurement methodologies.

In summary, none of the eight items of the semantic differential used to measure respondent evaluations of the preference measurement methodologies were deleted. However, as noted above, the final decision on the item on length needs to be made after collecting data from another group of respondents who have different attitudes toward the length of time taken to complete the tasks required to implement preference measurement methodologies.

Comparing Respondent Evaluations

The data measuring respondent evaluations of the three preference measurement methodologies investigated in this dissertation were collected using judgmental rating as well as direct comparison response modes. The results of the comparison of respondent evaluations of the preference measurement methodologies using both these response modes will be discussed in this section. The pattern of results obtained in this study will also be compared to those obtained in similar studies in the published literature. Table 82 summarizes the results of the comparisons of respondent evaluations of preference measurement methodologies reported in selected studies in the published literature (each of which was reviewed in Chapter IV).
Table 82: Selected Results of Comparisons of Respondent Evaluations of Preference Measurement Methodologies in the Literature

<table>
<thead>
<tr>
<th>Study</th>
<th>Variable</th>
<th>Result&lt;sup&gt;be&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adelman, Sticha, and Donnell (1984)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Confidence</td>
<td>ARS &gt; SCJ</td>
</tr>
<tr>
<td></td>
<td>Confidence</td>
<td>RS &gt; SG</td>
</tr>
<tr>
<td></td>
<td>Ease in Use</td>
<td>ARS &gt; SCJ</td>
</tr>
<tr>
<td></td>
<td>Ease in Use</td>
<td>RS &gt; SG</td>
</tr>
<tr>
<td>Bass, Steinberg, Pitt, et al. (1994)</td>
<td>Total Time</td>
<td>RS = SG&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>Busschbach, Hessing, and de Charro (1993)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Ease in Doing</td>
<td>RS &gt; SG</td>
</tr>
<tr>
<td></td>
<td>Total Time</td>
<td>RS &gt; SG&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
<tr>
<td>Lyness and Cornelius (1982)&lt;sup&gt;d&lt;/sup&gt;</td>
<td>Confidence</td>
<td>HOL &gt; ARS</td>
</tr>
<tr>
<td>Nutt (1980)</td>
<td>Ease in Use</td>
<td>ARS &gt; SCJ&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Perceived Accuracy</td>
<td>SCJ&lt;sup&gt;a&lt;/sup&gt; &gt; ARS</td>
</tr>
<tr>
<td>Nutt (1981)</td>
<td>Acceptability</td>
<td>SCJ &gt; ARS</td>
</tr>
<tr>
<td></td>
<td>Accuracy</td>
<td>SCJ &gt; ARS</td>
</tr>
<tr>
<td></td>
<td>Total Time</td>
<td>ARS &gt; SCJ</td>
</tr>
<tr>
<td></td>
<td>Understandability</td>
<td>SCJ &gt; ARS</td>
</tr>
<tr>
<td>Torrance (1976)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Ease in Doing</td>
<td>SG &gt; RS</td>
</tr>
<tr>
<td>Usher (1991)</td>
<td>Confidence</td>
<td>ARS &gt; ROC</td>
</tr>
<tr>
<td></td>
<td>Ease in Doing</td>
<td>ARS &gt; ROC</td>
</tr>
</tbody>
</table>
Table 82 (continued)

Legend:
   a: details of all studies shown are reviewed in Chapter IV
   b: ARS = additive rating scale procedure;
      HOL = holistic preference measurement strategy;
      ROC = rank-order judgment conjoint methodology;
      RS = rating scale scaling method;
      SCJ = social/clinical judgment analysis;
      SG = standard gamble scaling method;
   c: ">" means "was preferred to" for all variables except Total Time,
      for which ">" means "took more time than"
   d: other methodologies were also compared in this study; only the
      results relevant to the methodologies investigated in this study
      are presented
   e: although the RS scaling method took slightly longer than the
      SG scaling method, the difference was not statistically
      significant at the 0.05 level
   f: the RS scaling method in this study had some additional questions
      which were not asked in the SG scaling method; therefore, caution
      should be exercised in interpreting the results of the total time
   g: a modified version of SCJ was used in this study
Rather than discuss the results of the respondent evaluations using each response mode separately and then comment on the differences, it was decided to organize the discussion on respondent evaluations of the preference measurement methodologies around the variables pertinent to the evaluation. Given the pattern of results obtained and the fact that response mode effects were rather negligible for this aspect of the study, this decision appears to be justified. Further, since the results of the exploratory factor analyses indicated that there were two factors underlying the structure of respondent evaluations of each preference measurement methodology, the discussion is sub-divided into two major sections, one on each of the factors of interest.

First, however, a brief discussion on the most noteworthy result of the respondent evaluations of the three preference measurement methodologies investigated in this dissertation is presented -- the strong preference for the DCCM among the three preference measurement methodologies investigated in this study in terms of each of the variables pertinent to the evaluation. After this, the results of the respondent evaluations in terms of each of the two factors are discussed. The section concludes with a discussion on the role of response mode effects in the respondent evaluations of preference measurement methodologies investigated in this study.

(a) The Strong Preference for DCCM

The most notable result of the respondent evaluations of the three preference measurement methodologies investigated in this dissertation was the fact that the DCCM was evaluated as the best of the three methodologies
in terms of all the variables pertinent to the evaluation. Indeed, the results of the direct comparison of the respondent evaluations of the three methodologies indicated that for each of the variables pertinent to the evaluation, the majority of the respondents evaluated the DCCM as the best of the three methodologies. Further, it is remarkable that this result was invariant to any response mode effects which might have been operative in this study, since the respondent evaluations using both judgmental ratings as well as direct comparisons indicated that the DCCM was the best of the three methodologies investigated in this dissertation with respect to all the variables pertinent to the evaluation. As noted above, the role of response mode effects on the respondent evaluations will be discussed later in this section.

The role of a pro-computer bias cannot be ruled out in explaining this pattern of results of the respondent evaluations of the different preference measurement methodologies investigated in this study. In this regard, it should be remembered that the DCCM was the only methodology which required computer support for its implementation -- the other methodologies were implemented using other visual aids (as described in Chapter V). However, it is also possible -- given the nature of the study population, i.e., veterans -- that an anti-computer bias might be operative, since most veterans (and indeed most elderly respondents) are not computer literate. It would be an interesting extension to the present study to compare the three preference measurement methodologies in terms of respondent evaluations when each is implemented with the help of computer support,
so that any computer-related bias -- for or against -- would be distributed equally among the methodologies.

Indeed, some preliminary work in this very area has been reported by Goldstein, Clarke, Michelson, et al. (1994), who found a computerized display of information to be associated with better recall and recognition of information compared to a written display of information. It should, however, be noted that the Goldstein, Clarke, Michelson, et al. (1994) study did not directly measure respondent evaluations of the methodologies; further, unlike the present study, the Goldstein, Clarke, Michelson, et al. (1994) study did not use visual aids to enhance understanding of the written material. The significance of the Goldstein, Clarke, Michelson, et al. (1994) study, however, lies in the fact that it demonstrated the feasibility of developing computer-supported methods of data collection in the context of health status preference measurement.

(b) The “Feeling” Factor

The first factor which was interpreted as underlying the structure of respondent evaluations of the preference measurement methodologies pertained to an affective-based feeling about the methodology. In this regard, the specific variables which loaded on this factor were interest, sensibility, usefulness, and success. Further, although the variable “confidence” was not included in the semantic differential used to measure respondent evaluations, it was decided to include this variable as part of the “feeling” factor because confidence may relate more to a feeling about any construct rather than an operational aspect of the construct. It is, however,
acknowledged that the true factor on which the variable “confidence” loads on can only be determined by means of factor analysis on another data set which includes this variable as part of the semantic differential used to measure respondent evaluations of preference measurement methodologies. This is, naturally, an area which requires further research.

It is interesting to note that of all the variables included for discussion under this section (i.e., the “feeling” factor), the one which has received the most attention in the published studies measuring respondent evaluations of preference measurement methodologies is that of confidence. The published literature has found mixed results in terms of the confidence respondents say they have in top-down versus bottom-up measurement strategies. As shown in Table 82, Adelman, Sticha, and Donnell (1984) and Usher (1991) reported that their respondents had more confidence in bottom-up measurement strategies as compared to top-down measurement strategies, while Lyness and Cornelius (1982) reported the opposite. It should be noted that although the Lyness and Cornelius (1982) study did not use a statistically-inferred measurement strategy -- and therefore, not strictly a top-down measurement strategy -- the respondents’ task in a holistic preference measurement strategy is the same as that in a top-down measurement strategy, i.e., to evaluate multiattribute alternatives. The respondents of the present study said they had greater confidence in the methodology using a top-down measurement strategy (i.e., DCCM) as opposed to the methodologies using bottom-up measurement strategies (i.e., the CUF-RS and CUF-SG methodologies).
It is possible that these differences in respondent evaluations may be explained by the different response modes used in the different studies. In this regard, the present study is the only one of the ones shown in Table 82 which used choice as the response mode; the other studies used some form of judgmental responses. As explained in Chapter IV, choice is a more natural form of response for most subjects, since it most closely approximates the task respondents might perform if they were actually making a decision among alternatives. Further, the act of choosing an alternative from a competing set of alternatives has an inherent aspect of justification or accountability about it, as explained in Chapter IV. Such an aspect is not usually present when judgments are used as the response mode. Both these reasons suggest that respondents would be more confident of their evaluations if the evaluations were collected using choices as opposed to judgments as the response mode.

With respect to the comparison between the CUF-RS and CUF-SG methodologies in terms of confidence, the judgmental ratings showed that the CUF-RS methodology was preferred to the CUF-SG methodology, while the direct comparisons showed the opposite. In this regard, it should be noted that the difference between the judgmental ratings of the CUF-RS and CUF-SG methodologies was not found to be statistically significant at the 0.05 level (see Table 72). Therefore, it appears that the respondents of this study were more confident of the CUF-SG methodology as a measure of their preferences for health status, compared to the CUF-RS methodology.

This result is in opposition to that of Adelman, Sticha, and Donnell (1984), whose respondents were more confident of a methodology using the
rating scale scaling method as opposed to the standard gamble scaling method. These differences can be explained by the fact that the CUF-SG methodology was always implemented after the CUF-RS methodology in the present study. The reasons for this were explained in Chapter V, and will not be repeated here. However, the possibility of a learning effect and consequent greater confidence in the CUF-SG methodology as opposed to the CUF-RS methodology cannot be ruled out as a possible explanation for these results.

The remaining four variables -- interest, sensibility, usefulness, and success -- show a very similar pattern of results, and can therefore be discussed together. In each of these cases, the DCCM is the most preferred of the three methodologies. Comparing this result to those reported in the published literature, it can be seen from Table 82 that none of the studies in the table specifically measured any one of these variables. However, the studies by Nutt (1980, 1981) compared SCJ analysis and the ARS procedure in terms of accuracy and acceptability. Both these variables can be thought of as indicating a "feeling" about the methodologies in question, possibly success and usefulness respectively, since the accuracy of a methodology can be viewed as the degree of success with which it measures what it is intended to measure while the acceptability of a methodology can be viewed as related to the degree to which it will be considered useful. Viewed in this way, there is close correspondence between the results of both of Nutt's (1980, 1981) studies and the present one, since all the studies found top-down strategies to be preferred to bottom-up strategies in terms of "feelings" about preference measurement methodologies.
Further, for each of the four variables -- interest, sensibility, usefulness, and success -- both the judgmental ratings as well as the direct comparisons show that the CUF-SG methodology was preferred to the CUF-RS methodology by the respondents of the present study. Further, it is interesting to note that there were no significant differences at the 0.05 level between the judgmental ratings of the CUF-RS and CUF-SG methodologies on any of these variables (as shown in Table 72). Although this observation might seem to run counter to the explanation forwarded earlier in this chapter on the breakdown of the assumption of procedural invariance only when the alternatives under evaluation are spaced close to each other along the continuum of preference, it should be remembered that the CUF-SG methodology was always implemented after the CUF-RS methodology in this study. In this regard, the respondents of this study may have had a better "feeling" about the CUF-SG methodology as opposed to the CUF-RS methodology simply on account of learning effects. Therefore, it is possible that although the CUF-RS and CUF-SG methodologies might have been rated to be approximately equal in attractiveness, there might have been a systematic component in operation when respondents were asked to choose between the two methodologies. The similarity between this conclusion and that reached by Slovic (1975) is remarkable; moreover, when asked to rank the three preference measurement methodologies investigated in this study in terms of an overall evaluation, the respondents indicated they preferred the CUF-SG methodology over the CUF-RS methodology. This is also suggestive of the fact that the CUF-SG methodology might have been chosen
over the CUF-RS methodology even though both were rated about equal on account of an accountability or justification effect.

(c) The "Doing" Factor

The second factor which was interpreted as underlying the structure of respondent evaluations of the preference measurement methodologies pertained to an action-based performance of or doing the tasks required of the methodology. In this regard, the specific variables which loaded on this factor were difficulty in doing, difficulty in understanding and length. While it is acknowledged that the respondent evaluations of the DCCM found length to load on the other factor, it was decided to include the variable of length in the "doing" factor for the present discussion since this was the factor it loaded on in the case of the other two methodologies investigated in this dissertation. Further, this would facilitate the discussion of the total time taken to complete the tasks required of each methodology, which was also measured and would be more in line with "doing" the methodology as opposed to a "feeling" about the methodology. Moreover, although the variable "clarity" loaded on both factors in each of the exploratory factor analyses conducted in the present study, it was decided to include it in the present discussion since it could relate closely to the variable "difficulty in understanding."

As shown in Table 71, the DCCM took the shortest time to complete, followed by the CUF-RS methodology, and finally by the CUF-SG methodology. These differences were all statistically significant at a familywise alpha level of 0.05, as seen in Table 72. It is interesting to note
that the "subjective" respondent evaluations of the length of time taken to complete the tasks required of each methodology showed close correspondence with the "objective" estimates of total time (i.e., as measured by a watch). Indeed, Tables 71 and 72 show that the respondent evaluations of the variable "length" had the same ordering as those of the total time taken to complete the tasks required of each methodology. Further, the differences in respondent evaluations of the variable "length" were all statistically significant at a familywise alpha level of 0.05, as seen in Table 72. Therefore, it may be concluded that respondents had a good idea of the relative time commitments of each of the preference measurement methodologies investigated in this dissertation.

This conclusion brings up the issue of the need for including both "objective" as well as "subjective" measures of time commitments in comparative evaluations of preference measurement methodologies. The argument for including both types of measures is to check for consistency between the measures. In this regard, it may be argued that if there is consistency between the "objective" and "subjective" measures of time commitments, the similar conclusion may apply to the other variables pertinent to the respondent evaluations, for which there are no "objective" measures. Therefore, one may have more faith in the respondent evaluations of the other variables pertinent to the respondent evaluations. The argument against including the "objective" measure of time commitment is simply that it may prompt respondents to be pressured into finishing the tasks required of each methodology quickly, i.e., a sort of "under-the-gun" effect. Such an effect would naturally have consequences for the quality of the data
collected. It should, however, be noted that there was no indication of such an effect being operative in the present study (on the basis of informal and formal comments of the respondents). The argument against including the "subjective" measure of time commitment is it adds another item to the semantic differential used to measure respondent evaluations of the different preference measurement methodologies. Further, the results of the exploratory factor analyses discussed earlier in this chapter indicate that the item of "length" may be considered for deletion from the semantic differential (as explained earlier in this chapter).

As explained in Chapter IV, it was decided to include both measures in this study for the purpose of checking for consistency between the two measures. This was considered important since there have been no published applications of detailed investigations of factor structures of respondent evaluations of preference measurement methodologies. However, before making a decision to delete either one of the measures for use in future research, it is important to replicate the pattern of results observed in this study. This is important in order to ensure that the results obtained in this study were not a function of capitalization on the chance characteristics of the sample used for this study. This is, of course, the same conclusion that was reached at the end of the previous section, i.e., to not delete any items from the semantic differential until further research has been conducted with the instrument in different populations. Naturally, this is a subject for future research.

The result that the DCCM was the shortest to implement of the three preference measurement methodologies investigated in this dissertation is
consistent with the hypothesis forwarded in Chapter II and also with the literature reviewed in Chapter IV. As noted in Chapter IV, respondents usually focus on a few attributes when making choices among alternatives. Although there was no explicit test of the number of attributes which served as determinants of the evaluation, it should be noted that several respondents made informal comments which suggested that they might have used noncompensatory decision making strategies in making choices among the alternative health states.

Further, the respondents indicated in the respondent evaluations that the DCCM was the easiest of the three methodologies to do as well as understand. This could explain why respondents took much shorter time to complete the tasks required of them in the implementation of the DCCM as opposed to the other methodologies investigated in this dissertation. Moreover, it should be noted that the DCCM uses a top-down measurement strategy, while the CUF-RS and CUF-SG methodologies use bottom-up measurement strategies. In this regard, the pattern of results of the total time taken to complete the tasks required of each methodology replicate those of Nutt (1981), who found respondents took less time to complete the tasks required of social/clinical judgment or SCJ analysis (which uses a top-down measurement strategy) compared to the tasks required of the additive rating scale or ARS procedure (which uses a bottom-up measurement strategy). Further, Nutt (1981) also reported that the respondents of his study preferred SCJ analysis over the ARS procedure in terms of understandability, which is also consistent with the results of the present study in terms of the variable “difficulty in understanding.”
However, it should be noted that the results of the present study with respect to the variable "difficulty in doing" are not in agreement with some reports in the published literature. In this regard, Adelman, Sticha, and Donnell (1984), Nutt (1980), and Usher (1991) reported that their respective respondents found the ARS procedure (which uses a bottom-up measurement strategy) to be easier to do -- some studies operationalized this variable as easy in "use" -- than either SCJ analysis (see Adelman, Sticha, and Donnell, 1984; Nutt, 1980) or rank-order conjoint methodology (see Usher, 1991), both of which use a top-down measurement strategy. These discrepancies can be explained by either the pro-computer bias explained earlier or by the fact that the top-down measurement strategy used in this study (i.e., DCCM) used choice as the response mode. The other studies referred to above used some form of judgments as the response mode; therefore, it is possible that respondents generally find making choices among alternative to be easier to do as compared to making a judgment about each alternative under evaluation. Indeed, as reviewed in Chapter IV, this is consistent with the published literature on response mode effects.

Regarding the differences between the CUF-RS and the CUF-SG methodologies in terms of total time taken to complete the tasks required of each methodology, this is possibly a function of the fact that respondents found the CUF-RS methodology much easier to do compared to the CUF-SG methodology. Indeed, the respondent evaluations of the preference measurement methodologies using judgmental ratings indicated that the respondents judged the CUF-RS methodology to be significantly easier to do (familywise p<0.05) compared to the CUF-SG methodology (as shown in
Further, this particular result appeared to be invariant to any response mode effects that might have been operative in this study, since the rank ordering of the three methodologies on the variable "difficulty in doing" was the same whether judgmental ratings or direct comparisons were used as the response mode.

The total time taken by respondents to complete the tasks required of them for each methodology could also be a function of how clear they felt each methodology was and/or how difficult they thought each methodology was to understand. It should be noted that both these variables were involved in the cases where the phenomenon of preference reversal was shown to exist in the context of respondent evaluations of preference measurement methodologies. Specifically, when respondent evaluations were measured using judgmental ratings, it was found that the CUF-RS methodology was preferred over the CUF-SG methodology with respect to clarity and difficulty in understanding, but the opposite was true when respondent evaluations were measured using direct comparisons. However, the case of the variable "clarity" needs to be treated with some caution, since the absolute difference in mean judgmental ratings between the CUF-RS and CUF-SG methodologies on clarity was only 0.01 units; further, as shown in Table 72, this difference was not found to be statistically significant (at the 0.05 level).

The case of the variable "difficulty in understanding," however, deserves further explanation. When respondent evaluations were measured using judgmental ratings, it was found that respondents found the CUF-RS methodology to be more easy to understand than the CUF-SG methodology
(familywise p<0.05), while the opposite result held when respondent evaluations were measured using direct comparisons. This result may not be an instance of true preference reversal, because the CUF-SG methodology was always implemented after the CUF-RS methodology in this study. The reasons for this were explained in Chapter V, and will not be repeated here. Therefore, it is possible that the respondents of this study found the CUF-SG methodology more difficult to understand than the CUF-RS methodology in terms of absolute judgmental ratings, which did not have any aspect of comparison involved in the response. However, when asked to compare the methodologies, it is possible that respondents might have indicated they found the CUF-SG methodology easier to understand than the CUF-RS methodology on account of learning effects.

Other explanations may also account for the respondent evaluations of the CUF-RS and CUF-SG methodologies on the variable "difficulty in understanding." For instance, there is the issue of the degree of cognitive complexity involved in performing the tasks required to implement each methodology. In this regard, although both the scaling methods require each alternative health state to be evaluated with respect to two anchors, the rating scale scaling method -- as implemented in the present study -- involved the evaluation of eight multiattribute health states simultaneously (since the arrows corresponding to each health state were not replaced after each individual health state was rated) while the standard gamble scaling method involved the evaluation of multiattribute health states one at a time. Therefore, it is possible that the degree of cognitive burden imposed on
respondents was less in the case of the CUF-SG as opposed to the CUF-RS methodology.

This difference in cognitive burden between the two methodologies would have been manifest when respondents were asked to provide a direct comparison of the methodologies as opposed to when respondents were asked to rate each methodology separately. This is because information-processing in a direct comparison response mode is usually attribute-based, while information-processing in a judgmental rating response mode is usually alternative-based (see Narayana, 1977).

It is instructive to compare the pattern of results between the CUF-RS and CUF-SG methodologies obtained in this study with those reported in the published literature. In terms of total time, Bass, Steinberg, Pitt, et al. (1994) found no statistical differences between the rating scale and standard gamble scaling methods in their study, while Busschbach, Hessing, and de Charro (1993a) reported that their respondents took longer to complete the rating scale scaling method compared to the standard gamble scaling method. However, the time estimates of the latter study need to be viewed in recognition of the fact that the respondents answered some additional questions in the rating scale scaling method as opposed to the standard gamble scaling method -- therefore, the additional time taken for completing the tasks required of the rating scale scaling method might have been reflective of these additional questions.

With respect to the respondent-evaluated difficulty in doing the tasks required of each methodology, the published literature shows mixed results. For instance, Torrance (1976a) has reported that the respondents of this
study found the standard gamble scaling method to be easier to do compared to the rating scale scaling method, while Adelman, Sticha, and Donnell (1984) and Busschbach, Hessing, and de Charro (1993a) have reported the opposite. It should be noted that Torrance’s (1976a) conclusions were largely based on informal feedback from his respondents; the only formal feedback in terms of questionnaires was for a sample of highly educated respondents — even then, the scores given by respondents for each scaling method were not reported. Therefore, one cannot make any conclusions regarding the practical or statistical significance of Torrance’s (1976a) conclusions on respondent evaluations of difficulty in doing the rating scale and standard gamble scaling methods. In light of this observation, the observed results of the present study in terms of comparing respondent evaluations of the rating scale and standard gamble scaling methods in terms of difficulty in doing are largely consistent with those reported in the published literature.

(d) Response Mode Effects in Respondent Evaluations

It was noted earlier that the DCCM was the most preferred of the three methodologies investigated in this dissertation with respect to each of the variables pertinent to the respondent evaluations. In terms of response mode effects, this result is consistent with those of the comparison of preference scores obtained using the preference measurement methodologies, where it was found that the phenomenon of preference reversal does not occur when the alternatives under evaluation are well-spaced out along the continuum of preferences. The fact that the DCCM was
well-separated from the other two methodologies investigated in this study can be seen from the respondent evaluations of the three methodologies using judgmental ratings, shown in Table 71. Further, as shown in Table 72, the results of the pairwise comparisons indicate that the DCCM is significantly better than at least one of the other methodologies in terms of all but three of the variables pertinent to the evaluation.

Considering the preferential rank ordering of the CUF-RS and CUF-SG methodologies on each of the variables pertinent to the respondent evaluations, Table 74 shows that there were three instances of preference reversal, i.e., for the variables of clarity, difficulty in understanding, and confidence. It can be seen from Tables 71 and 72 that the CUF-RS and CUF-SG methodologies were rated as being close to each other on the continuum of preference for the variables of clarity and confidence, as indicated by the lack of statistical significance of the pairwise comparison between the CUF-RS and CUF-SG methodologies on these variables (as reported in Table 72).

However, there was a significant difference in the ratings of the CUF-RS and CUF-SG methodologies for the variable of difficulty in understanding, which runs counter to the above explanation. Further, there are other instances in which the CUF-RS and CUF-SG methodologies were rated as being close to each other along the continuum of preference (as indicated by the lack of statistical significance between their respective judgmental ratings, shown in Table 72) but there was no corresponding occurrence of preference reversal, e.g., usefulness and success. These results can be explained on account of the fact that the CUF-SG methodology was always implemented after the CUF-RS methodology in the present study.
and that in terms of the overall evaluation, respondents preferred the CUF-SG methodology to the CUF-RS methodology -- therefore, it is possible that the effect of accountability or justification may have led to the CUF-SG methodology being chosen over the CUF-RS methodology in the direct comparison response mode.

Taken together, the findings of the comparison of preference scores and respondent evaluations suggest that the breakdown of the assumption of procedural invariance might be mediated by the similarity of the alternatives under evaluation. Further, the effect of justification or accountability in choice as opposed to judgment may also be influential in the phenomenon of preference reversal.

In summary, the results of the respondent evaluations of the three preference measurement methodologies investigated in this study revealed a clear preference for the DCCM over the CUF-RS and CUF-SG methodologies. Although the effect of a pro-computer bias cannot be ruled out, this result is noteworthy in that it was robust to any response mode effects which might have been operative in this study. The comparison between the CUF-RS and CUF-SG methodologies appeared to be split between the two factors -- the CUF-RS methodology was found to be preferred over the CUF-SG methodology on most of the variables included in the "doing" factor, while the opposite was true in the case of the "feeling" factor. However, when asked to provide an overall ranking of the three methodologies, respondents indicated that the CUF-SG methodology was preferred over the CUF-RS methodology. Although the role of learning
effects cannot be ruled out in explaining this preference of the CUF-SG methodology over the CUF-RS methodology, the fact that these results were also largely free from any response mode effects is noteworthy. The implications of these results are discussed along with the implications of the other results of this study in the next section of this chapter.

Implications of the Results

The results of this dissertation have some implications for health status preference measurement, health status measurement, and the outcomes movement in health care. Each of these implications are discussed in this section.

Implications for Health Status Preference Measurement

The purpose of this dissertation was to validate discrete choice conjoint methodology (DCCM) as a measure of health status preferences. The approach adopted for validation was construct validation, specifically convergent validation against a conditional utility function-based measurement strategy with a rating scale scaling method (or CUF-RS methodology) and a conditional utility function-based measurement strategy with a standard gamble scaling method (or CUF-SG methodology). The results of the study generally supported the construct validity of DCCM as a measure of health status preferences, according to the principle of convergent validation.
Therefore, this study offers a viable new methodology for the measurement of health status preferences. The question that arises is what are the relative advantages and disadvantages of this new methodology compared to the existing health status preference measurement methodologies. Possibly the most salient advantage of DCCM over the existing health status preference measurement methodologies is that it uses choice as the response mode, and therefore provides theoretical justification for explaining and/or predicting respondents' choices on the basis of their preferences inferred from choice data. As explained in Chapter II, preferences inferred from judgmental data might be different than those inferred from choice data on account of the possible breakdown of the assumption of procedural invariance in practice.

Moreover, preferences inferred from judgmental data do not have a theoretical justification for explaining and/or predicting respondents' choices among alternatives under evaluation. Since the ultimate purpose of measuring patients' preferences for health status is for use in public policy and/or clinical decision making, and given the fact that making a decision involves choosing among alternatives under evaluation, it is important to have an adequate theoretical basis for linking health status preferences to choices. At the present state of the art of health status measurement, DCCM is the only health status preference measurement methodology which offers this theoretical basis.

The pattern of results obtained in this study indicate that the assumption of procedural invariance is likely to breakdown in practice when the alternatives are closely spaced along the subjective continuum of
preference. In light of this observation, it may be tempting to conclude that the only instance in which one might want to consider using DCCM to measure health status preferences is when the alternative health states under evaluation are closely spaced along the subjective continuum of preference. Unfortunately, this information is not available prior to the measurement of health status preferences -- after all, that is the purpose of measuring health status preferences. Further, it is important for the researcher to not make a priori guesses about the relative spacing of alternative health states along the subjective continuum of preference, since it is possible that such guesses might be different from those of the patient -- as noted in Chapters I and II, it is the patient's preferences which need to be considered for use in clinical decision making.

Given this, it may be tempting to adopt the extreme position that when can't the DCCM be used as a health status preference measurement methodology, since it will rarely -- if ever -- be known a priori how close the alternatives are along the subjective continuum of preference. This is because the DCCM as a measure of health status preferences is not without its own drawbacks. The major drawback is that given the current state-of-the-art of DCCM, one cannot analyze data at the individual level. The reasons for this were explained in Chapter III and will not be repeated here. It may be considered important to analyze the data at the individual patient level for use in individual -- as opposed to group -- clinical decision making. Further, the use of the individual as the unit of analysis has obvious advantages in segmentation studies, since segments of respondents can be formed by
aggregation of respondents with similar preference functions as opposed to a priori specification of segments (as used in the current study).

Other drawbacks of the DCCM are related to the complexity of the study design, which -- as explained in Chapter III -- requires the construction of interlocking double conditional designs for implementation. However, it should be noted that the availability of computer programs (such as the CBC System used in the present study) may nullify this particular drawback.

It should be noted that the results of this study supported the reliability and construct validity of DCCM as a measure of health status preferences; however, this study was not designed as a comparative assessment of the reliability or validity of the three preference measurement methodologies investigated in this study. Therefore, no conclusions can be drawn on the basis of the results of this study on which of the three preference measurement methodologies investigated in this study has greater reliability or validity as a measure of health status preferences. Indeed, in the absence of a gold standard as a “true” measure of health status preferences, one cannot know which one of several alternative preference measurement methodologies provides an accurate measure of respondents’ health status preferences.

However, the present study was designed as a head-to-head comparison of the CUF-RS, CUF-SG, and DCCM methodologies in terms of respondent evaluations. As noted earlier, respondent evaluations are an important aspect to consider in selecting among alternative preference measurement methodologies, since they may be indicative of the quality of
the data collected using the preference measurement methodologies. In this regard, the results of this study showed that the DCCM was the most preferred in terms of each of the different variables pertinent to the respondent evaluations of the three preference measurement methodologies investigated in this study. Although the role of a pro-computer bias cannot be ruled out in explaining the results of the respondent evaluations, it is still remarkable that the results were not sensitive to any response mode effects which might have been operative in this study. Further, the fact that the DCCM was most preferred in each of the variables pertinent to the evaluation -- by the majority of respondents as measured by direct comparisons of the three preference measurement methodologies -- is also noteworthy.

In summary, although no conclusion may be drawn regarding the relative reliability and validity of the different preference measurement methodologies investigated in this study, it should be noted that this study provided empirical evidence for the reliability and construct validity of DCCM as a measure of health status preferences. Further, it was shown that the DCCM was the most preferred of the three preference measurement methodologies investigated in this study in terms of respondent evaluations. Therefore, it is concluded that DCCM is a viable choice-based methodology for measuring health status preferences which has an adequate theoretical basis for explaining and/or predicting respondents' choices among health-care alternatives.
On another front, this study offers an instrument for measuring respondent evaluations of health status preference measurement methodologies, the psychometric properties of which have been subject to investigation and found to be acceptable. This instrument may prove useful in future research comparing preference measurement methodologies and in deciding which methodology or methodologies to use in any particular instance. Further, by providing some understanding of the underlying factor structure of respondent evaluations of preference measurement methodologies, this study addresses an issue which has not been completely addressed in the psychometric, decision theoretic, business, as well as health-care literature on preference measurement.

Implications for Health Status Measurement

The importance of measuring health status preferences for the purpose of providing true meaning to health status scores was discussed in Chapter II. By measuring EuroQol health status preferences using three different preference measurement methodologies, this study provides insight into the relationship between the different determinants of true meaning of health status scores, such as value versus utility, and judgmental versus choice-based preferences. Although the DCCM which was validated in this study as a measure of health status preferences does not incorporate all the factors that need to be considered in providing true meaning to health status scores (for example, the current state of the art in DCCM does not allow for the incorporation of respondents' attitude toward risk as part of the preference function), it does provide a theoretical linkage between health
status preferences and choices made on the basis of those preferences. In other words, the DCCM offers theoretical support for making choices among health-care alternatives on the basis of health status preferences obtained as a result of its implementation. As noted earlier, the DCCM is the only existing health status preference measurement methodology which offers such a theoretical linkage between preferences and choices.

The study also has implications for the EuroQol health status measurement instrument. Specifically, by modeling preferences for health status (as operationalized by the EuroQol) using three different preference measurement methodologies, this study provides valuable information regarding the nature of the model form of the EuroQol health status preference function. In this regard, it is notable that there was agreement among the three preference measurement methodologies in terms of the general model form of the EuroQol health status preference function (i.e., a multiplicative model form). As discussed in Chapter II, this is generally consistent with most studies which have attempted to verify the existence of interactions among health status dimensions, and provides support to the notion that an additive model form is overly restrictive to represent the complexity of the construct of health status.

It was noted earlier that the current focus of research of the EuroQol group of researchers was methodological, in terms of measuring preferences for health status (operationalized by the EuroQol) using different measurement methodologies and in different patient populations. By measuring EuroQol health status preferences using two existing and one new health status preference measurement methodology, this study may
prove useful in terms of the former goal of the EuroQol group of researchers. Further, with respect to the second goal -- i.e., measuring preferences for health status in different patient populations -- there have been no published applications of the EuroQol in a U.S. population. Therefore, the present study provides useful information about the health status preferences of a group of U.S. patients, and may be useful for the purpose of cross-national comparisons of EuroQol health status preferences.

Moreover, the studies reviewed in Chapter IV (some of which were summarized in Table 75) indicate that very few studies in the published health status measurement literature have measured the health status preferences of hypertensive patients. This is in spite of the fact that there have been several studies which have measured the health status of hypertensive patients and several hypertension-specific health status measurement instruments have been developed (see Bulpitt and Fletcher, 1992, 1994 for a review). Therefore, the present study fills a much-needed gap in the published literature on health status measurement in hypertension.

Implications for the Outcomes Movement in Health Care

It is important to note the broader implications of this study for the outcomes movement in health care. It was noted in Chapter I that a key step in the realization of the true potential of health status measures is that these measures should be easily interpretable and understood by all interested stakeholders. In this regard, the implications of this study for the major stakeholders in the outcomes movement in health care need to be noted --
pharmacists and other health-care providers, patients, pharmaceutical manufacturers and marketers, and payors.

As Ellwood (1988) has noted, the centerpiece of the outcomes movement is the measurement of patients' quality of life, of which health status is an important component (as shown in Figure 1). Further, it was explained in Chapter I that patients' evaluation of their own health is usually operationalized in terms of health status measures as opposed to the more traditional clinical measures with which most clinicians are comfortable (see Milio, 1983). Given this, it becomes important for pharmacists and other health-care providers to understand the meaning of health status scores if there is to be true communication between patients and clinicians.

Specifically in the case of pharmacists, if pharmacists are expected to adopt and fulfill the mission of pharmaceutical care, it is necessary for them to understand the role of health status evaluation and measurement. Further, in order to be effective counselors of drug therapy for patients, pharmacists need to understand patients' evaluation of their own health. This is especially true while dealing with patients with chronic conditions, in whom improvement of health status may be the primary goal of therapy. Moreover, clinicians and other health-care decision makers need to understand patients' preferences for health status so that they may be able to make their decisions be accordingly more responsive to patients' preferences. As explained in Chapter I, this may lead to better patient outcomes, such as increased patient satisfaction and improved health-related quality of life.

The importance of measuring health status preferences for the purpose of providing true meaning to health status scores was discussed in
Chapter II. By providing a theoretical linkage between health status preferences and choices among health-care alternatives, the DCCM which was validated in this study shows promise as a methodology for measuring health status preferences and consequently providing meaning to health status scores so that these scores may then be used for public policy or clinical decision making in health care. Since these scores would reflect the preferences of the patient, it may be expected that the decisions made on the basis of these scores may be more responsive to patients' preferences. Therefore, health-care decisions made on the basis of health status scores may lead to better patient outcomes, as discussed earlier.

This might facilitate the use of such scores in routine clinical decision making. This in turn will present patients with the opportunity to let their preferences be known to the clinicians making decisions on their behalf. As noted in Chapter I, the literature suggests that patients would like to let clinicians know what they want and discuss alternatives with clinicians but leave the final decision (or choice of therapy) to clinicians. By letting clinicians know patients' preferences for health states, it is possible that the therapeutic decisions made by these clinicians would be more responsive to patient preferences. This in turn would lead to positive patient outcomes, such as better compliance with therapy, increased patient satisfaction, and improved health-related quality of life.

It was noted in Chapter I that a major driver behind the development of health status scores was to let payors know what they were getting for the money they spent on health care. By providing meaning to health status scores, this study may facilitate their use in making “patient-centered”
reimbursement decisions, since patients' preferences will be reflected in the health status scores used for the purpose of health-care decision making.

It was also noted in Chapter 1 that pharmaceutical manufacturers and marketers are increasingly sponsoring and conducting pharmacoeconomic and health status studies. The measures of preference provided for different dimensions and dimension levels of health status might help pharmaceutical marketers in positioning their products so as to effectively compete in a crowded antihypertensive market.

On a more methodological front, one of the issues in pharmacoeconomic analyses today is the use of "cost-utility league tables." These league tables are nothing but a listing of the cost and consequences of alternative health-care interventions under evaluation in any specific health-care decision context. However, there have been repeated warnings in the published literature on the dangers of blindly using such league tables (see, for e.g., Drummond, Torrance, and Mason, 1993; Petrou, Malek, and Davey, 1993). The reason for this is simply because -- as explained in various chapters in this dissertation -- different health status preference measures incorporate different assumptions (i.e., values versus utilities, judgments versus choices) and therefore may not be strictly comparable. If, however, there is the need to determine what (for example) the equivalent of a particular score of a standard gamble would be on a rating scale continuum of preference, the functional relationships estimated in this study could be used as a conversion formula. In this regard, each of the preference scores on the CUF-RS, CUF-SG, and DCCM scales could be converted to another of the
three scales used in this study. This may be useful for the purpose of estimating parameters in decision analysis or meta-analysis.

Limitations

It is important to consider the following potential limitations of this study. First, the study population consisted of veterans treated on an outpatient basis for hypertension at a Midwestern Department of Veterans Affairs Outpatient Clinic. The reason for this constraint was primarily logistical and resource-driven. However, it is acknowledged that the results of this study may not be generalizable to patients enrolled in other health plans in other geographic locations.

Second, only hypertensive patients were surveyed for this study. The reason for this is to avoid any variation in health status preferences that may be due to existing diseases or conditions. Since the emphasis of this study is primarily methodological, it was felt that it was necessary to control as many sources of variance as possible in order to be sure that the methodology "worked." However, it is acknowledged that the study results may not be generalizable to patients suffering from another condition or disease.

Third, it was decided to use patients as subjects because it was felt that they would be the best judges of their preferences for health states. Therefore, the results of this study are only indicative of the health status preferences of hypertensive patients, not their family members or their friends, or their health-care professionals.
Fourth, it was decided to conceptualize health status on the basis of the conceptualization inherent in the EuroQol -- therefore, the results of the study may not be generalizable to health status preference functions using other operationalizations of health status.

Fifth, following Torrance, Boyle, and Horwood (1982) and Torrance, Zhang, Feeny, et al. (1992), it was decided to hold the expected duration of stay in the health state under evaluation as a constant to be introduced as part of the instructions to the respondents. This was fixed at one year (in accordance with the instructions of the EuroQol). Therefore, the health status preferences obtained in this study will be with respect to one fixed time duration, and may not be the same for the same health state under other durations of time.

Sixth, given the problems of DCCM in handling data at the individual level, only aggregate-level data were reported. This does, of course, obscure any individual differences that might exist in the data. However, given the methodological focus of this study and the fact the most published applications of health status preference measurement reported the data at the aggregate level, this was considered acceptable. Still, at the present state of the art of DCCM, this is a drawback in terms of the application of DCCM to the measurement of individual health status preferences.

Finally, given the focus on investigating the construct validation of DCCM as a measure of health status preferences using the approach of convergent validation, it was decided to be consistent with the published literature in the field and use immediate death as the bottom anchor of the health status preference scale. This, however, does not preclude the existence
of health states considered worse than death -- since the focus of this dissertation was more methodological in terms of comparing different preference measurement methodologies, it was considered acceptable.

**Avenues for Future Research**

Having discussed the implications and limitations of the present study, it is important to discuss some avenues for future research that might help the field of health status measurement realize its potential. One of the major avenues for future research is in the area of choice-based modeling of health status preferences. The present study has only taken the first step in this area. At a minimum, the test-retest reliability of the parameter estimates obtained using the DCCM in the present study needs to be estimated. Further, the area of choice-based modeling of preferences is an exciting field of ongoing research, as witnessed by recent advances in the design of hierarchical choice-based designs (for e.g., see Oppewal, Louviere, and Timmermans, 1994), the incorporation of latent variables in choice models (see Batsell and Louviere, 1991), and the use of multiple correspondence analysis to analyze choice data (for e.g., see Kaciak and Louviere, 1990). Each of these extensions would be interesting to apply to the measurement of health status preferences.

Further, as noted in Chapter III, there are several different types of choice models which may be used to describe respondents' decision making processes. It would be instructive to compare the merits and demerits of each as a measure of health status preferences. This is important because --
as noted in Chapter III -- the selection of a particular choice model represents a tradeoff between analytical tractability and veracity of underlying assumptions.

For instance, the Elimination By Aspects model of choice (Tversky, 1972) is considered to be a valid descriptive representation of respondents' decision making processes, but has thus far proven to be quite intractable in an analytical sense. On the other hand, the multinomial logit (MNL) choice model used in this study offers a compromise between analytical tractability and the veracity of its underlying assumptions. However, the assumption of Independence from Irrelevant Alternatives (or IIA) of the MNL choice model has been shown to not hold in practice in some studies. In order to account for this violation of IIA, one may use a Generalized Extreme Value or GEV choice model (McFadden, 1979); however, the drawback of the GEV choice model is that it is less analytically tractable than the MNL choice model.

Therefore, it can be seen that there are several possible choice models that may be assumed to represent respondents' decision making processes. It is possible that these models may not be equivalent in terms of their resulting parameter estimates of health status levels and attributes. Further, since each has its own strengths and weaknesses in terms of veracity of assumptions and analytical tractability, it would be instructive to compare each as an alternative representation of respondents' health status decision making processes.

Although it was decided to use the conditional utility function-based procedure as the comparator in terms of measurement strategy for the purpose of this study, other measurement strategies could have been used
for this purpose. In this regard, an interesting extension to this study would be to compare the DCCM with a rank-order and/or a rating scale conjoint judgment strategy as a measure of health status preferences. Since similar comparisons have been reported in the marketing literature (as reviewed in Chapter III), it would be interesting to see if their results could be replicated in the context of health status preference measurement.

Along the same lines, research focused on comparing the DCCM to health status preference measurement methodologies using other scaling methods such as the time tradeoff and person tradeoff methods needs to be conducted. Further, the effect of data collection method needs further investigation, in order to determine if there was a pro- or anti-computer bias in operation in the present study.

The results of the present study suggested that the phenomenon of preference reversal might occur when the alternatives under evaluation are closely spaced along the continuum of preference. This is an interesting result, since it suggests the role of similarity of alternatives as a mediator for any possible response mode effects. However, as an isolated result, it is in need of replication. It would be possible to design a study for this purpose -- the study in question may consist of two groups, the first of which evaluates alternatives which are equal in attractiveness (as previously matched) and the second of which evaluates alternatives which are well-spaced along the continuum of preference. Both groups could be made to evaluate the alternatives using judgment and choice response modes (in random order). If the pattern of results observed in the present study were to be replicated,
the phenomenon of preference reversal would occur in the group which evaluated equally attractive alternatives, but not in the other group.

Another interesting avenue for future research is in respondent decision making processes in health status preference measurement. There is a paucity of studies in the published health status literature which have used process tracing methods to determine respondents' decision making processes. The usual approach has been to assume that a particular decision making process applies, usually compensatory decision making. However, the informal feedback obtained from respondents in this study indicated that at least some of the respondents in this study used noncompensatory decision making processes while evaluating the alternative health states presented to them. The implications of such decision making processes on health status preferences have largely been unexplored in the health status literature.

Further, the instrument developed for measuring respondent evaluations of preference measurement methodologies needs to be used in other studies with different patient populations. As noted earlier, this will permit a decision to be made regarding retaining or deleting the item on length in the instrument. Further, it will also test whether the particular factor structure of respondent evaluations of preference measurement methodologies determined in this study is replicated in other studies with different patients.

Finally, there are some other areas for future research which need to be mentioned. These areas share the common theme that they were discovered during the decision making stages of the present study, and it
was felt that it would be useful if there were some guidance offered by the published literature in these areas. These areas include the relative performance of fixed versus randomly constructed choice sets in DCCM and finely-divided versus coarsely-divided rating scales in health status preference measurement.

Although these avenues are by no means an exhaustive listing of all possible avenues for future research, they do serve as a starting point for building upon the results of this study and moving onto the next level of health status preference measurement. As noted in Chapter I, this is necessary in order to provide true meaning to health status scores, so that they may realize their full potential in public policy and clinical decision making. In this way, the decisions made on the basis of health status scores may be more responsive to patients' preferences and reflect a reasoned approach to decision making in health care. This is, as noted in the opening section of Chapter I, consistent with the guiding principles of outcomes management. Indeed, it might serve to propel the health-care system into a fourth era, *The Era of Responsive and Reasoned Decision Making.*
APPENDIX A

CORRESPONDENCE WITH EUROQOL GROUP AND EUROQOL INSTRUMENT USED IN THE PRESENT STUDY
October 3, 1994

Claire Gudex  
Centre for Health Economics  
University of York  
Heslington  
York Y01 5DD

Dear Dr. Gudex,

I am a Ph.D. Candidate in Pharmaceutical Administration at the Ohio State University. My dissertation research involves measuring and comparing health status preferences via a relatively new technique called Choice-Based Conjoint Analysis and the classical Standard Gamble method.

Having found the conceptualization of health status (i.e., the dimensions and levels of health status) adopted by the EuroQol® instrument to be particularly suited for my dissertation research, I would like to request the permission of the EuroQol® group to use these specific dimensions and levels for my research.

If you have any questions pertaining to my request, please do not hesitate to contact me at (614)-292-1716. If you would prefer to speak with my advisor, Dr. Dev S. Pathak, you may reach him at (614)-292-6415. If it is more convenient for you to communicate with us via FAX, please contact us at (614)-292-1335. Thank you very much for your consideration.

Sincerely,

Zafar Hakim
Zafar Hakim  
Pharmacy, Practice & Administration  
College of Pharmacy  
The Ohio State University  
217 LLoyd M. Parks Hall  
500 West 12th Ave  
Columbus  
OHIO OH 43210  

12th October 1994

Dear Mr Hakim,

Thank you for your enquiry regarding the EuroQol instrument.

Enclosed is a EuroQol questionnaire and some accompanying information, including a leaflet which you might find helpful. The EuroQol Group has decided that the instrument may be used without cost provided that the research is not being undertaken for commercial uses.

Please could you provide us with a brief description of your project for our records (see form enclosed). Meanwhile please do not hesitate to contact me again if you have any further enquiries.

Yours sincerely,

Dr Claire Gudex  
Research Fellow

NB: The enclosure regarding our TTD & SG manuals may also be of interest.
MEASURING PREFERENCES FOR HEALTH OUTCOMES

EuroQol® Instrument

Start Time

End Time
By placing a check (√) in one box in each group below, please indicate which statements best describe your own health state *today*.

**Mobility**

- I have no problems in walking about
- I have some problems in walking about
- I am confined to bed

**Self-care**

- I have no problems with self-care
- I have some problems washing or dressing myself
- I am unable to wash or dress myself

**Usual Activities**

- I have no problems with performing my usual activities (e.g., work, study, housework, family or leisure activities)
- I have some problems with performing my usual activities
- I am unable to perform my usual activities

**Pain/Discomfort**

- I have no pain or discomfort
- I have moderate pain or discomfort
- I have extreme pain or discomfort

**Anxiety/Depression**

- I am not anxious or depressed
- I am moderately anxious or depressed
- I am extremely anxious or depressed

Compared with my general level of health over the past 12 months, my health state *today* is (please check one box):

- Better
- Much the same
- Worse
APPENDIX B

CARDS USED TO DESCRIBE HEALTH STATES
Operation Works: Walk and Run Normally

Operation Does Not Work: Use Crutches

Leg Heals by itself: Limp
M1

no problems walking about

M2

some problems walking about

M3

confined to bed
S1

no problems with self-care

S2

some problems washing/dressing self

S3

unable to wash/dress self
no problems performing usual activities

some problems performing usual activities

unable to perform usual activities
no pain or discomfort

moderate pain or discomfort

extreme pain or discomfort
A1
not anxious or depressed

A2
moderately anxious or depressed

A3
extremely anxious or depressed
<table>
<thead>
<tr>
<th></th>
<th>AB</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no problems walking about</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no problems with self care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no pain or discomfort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>not anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AW</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>confined to bed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>unable to wash/dress self</td>
<td></td>
</tr>
<tr>
<td></td>
<td>unable to perform usual activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>extreme pain or discomfort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>extremely anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>confined to bed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no problems with self-care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no pain or discomfort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>not anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>
CS
no problems walking about
unable to wash/dress self
no problems performing usual activities
no pain or discomfort
not anxious or depressed

CU
no problems walking about
no problems with self-care
unable to perform usual activities
no pain or discomfort
not anxious or depressed

CP
no problems walking about
no problems with self-care
no problems performing usual activities
extreme pain or discomfort
not anxious or depressed
CA

no problems walking about
no problems with self-care
no problems performing usual activities
no pain or discomfort
extremely anxious or depressed

BO

confined to bed
some problems washing/dressing self
no problems performing usual activities
no pain or discomfort
not anxious or depressed

E3

some problems walking about
no problems with self-care
some problems performing usual activities
extreme pain or discomfort
moderately anxious or depressed
confined to bed
unable to wash/dress self
unable to perform usual activities
moderate pain or discomfort
not anxious or depressed

IMMEDIATE DEATH
APPENDIX C

INTERVIEWER SCHEDULE

NOTE:

1. This schedule is written as if the attribute of mobility was the first to be evaluated -- in the actual interview, the order of presentation of attributes was randomized
2. This schedule is written as if the order of presentation was feeling thermometer tasks, followed by choice tasks, followed by standard gamble tasks -- in the actual interview, the order of presentation of tasks was partially randomized (as explained in the text)
Good Mr./Ms. ____________________.

My name is ZAFAR HAKIM, and I am an investigator for this study. I'm also the person who spoke to you over the telephone on ______________. Thank you so much for taking the time to help us in our research.

As I had mentioned on the telephone, I am a graduate student in Pharmacy at The Ohio State University. I'm in the final stages of my research here, and I am currently working with the VA outpatient clinic on my dissertation project, which is the reason for this interview.

---

TAKE OUT RESEARCH CONSENT FORMS

---

Before I ask you any questions, I would like to show you this research consent form. What this form does is give me permission to ask you any questions. Please read through it carefully, and if you have any questions please do not hesitate to ask me. After reading through the form, please sign in the appropriate place. Then I shall sign in the place next to where you sign.

---

WAIT FOR PATIENT TO SIGN INFORMED CONSENT FORMS

SIGN IN APPROPRIATE PLACE

GIVE PHARMACIST IN NEIGHBORING BOOTH FORMS TO SIGN AS WITNESS

FILL IN SUMMARY INFORMATION SHEET

NOTE DOWN OVERALL START TIME
Thank you. We are now ready to begin.

**TAKE OUT BACKGROUND INFORMATION INSTRUMENT**

I would like to begin by asking you some background information about yourself. Since we will not be noting your name or social security number at any point in this interview, we will use this background information to describe patients similar to yourself *in very general terms*. In this way, we can be sure of representing your personal preferences without running the risk of specifically identifying you in any way. *Remember, all information you provide will be maintained in confidence.*

Please note that you are not obliged in any way to answer these questions. If you feel awkward for any reason whatsoever, please feel free to refuse to answer the question and ask me to move on to the next question. Shall we begin?

**ADMINISTER EACH QUESTION IN TURN**

You are probably wondering what is the purpose of this interview and what this study is all about. Well, as I had stated on the telephone, the purpose of this study is to understand the preferences of patients such as yourself for the outcomes of drug therapy. Now, let me explain this in a little more detail.
Whenever you take any drug, you experience certain outcomes. Some of these outcomes are good, while others are not. An example of a good outcome of drug therapy may be a decrease in the level of pain experienced by an individual, while an example of a not-so-good outcome of the same drug therapy may be a side-effect of confusion or trouble in concentrating, which could lead to some problems in performing your usual activities. While it would be nice to totally separate the good from the not-so-good outcomes of drug therapy, unfortunately this is not possible since all these outcomes may occur at the same time.

However, different drug therapies have different outcomes. Therefore, while choosing between alternative drug therapies, attention must be paid to these different outcomes. Further, and of special relevance to this study, there is the issue of the importance of these different outcomes to the person who is taking the drug therapy -- the patient. For instance, one patient may not be bothered by the side-effect of confusion, while another may be extremely bothered by this same side-effect.

Therefore, in order to make a choice between alternative drug therapies, two pieces of information are needed. First, we need information about the different outcomes of the alternative drug therapies. Second, we need information about the degree of importance patients taking those drug therapies attach to the outcomes in question, i.e., we need information about patients' preferences for the outcomes of drug therapy.

Medical science has provided us with much of the first piece of information. However, little is known about the second piece of information, i.e., patients' preferences for the outcomes of drug therapy. This study aims to fill that gap by providing health-care professionals with such information.

Patients such as yourself are in a unique position to provide such information, because in the end it is your quality of life which is affected by the outcomes of the drug therapy you take.
That reminds me of another point I want to stress at the outset of this interview, that is that *there is no such thing as a right or wrong answer to any of the questions you will be asked in this interview*. The reason for this is straightforward -- since it is your preferences which are of central interest to this study, whatever answer you give is the correct one. Before I proceed any further, do you have any questions about the purpose of this study?

---

**WAIT FOR ANY QUESTIONS AND ANSWER THEM**

---

First, could you please indicate on this sheet the state of your own health *today*?

---

**GIVE PATIENT EUROQOL INSTRUMENT**

**WAIT TILL PATIENT FILLS OUT EUROQOL INSTRUMENT AND HANDS BACK**

---

Thank you. We are now ready to begin to measure your preferences for these outcomes. In this interview, we shall be using three different techniques to measure your preferences for these outcomes -- a "feeling thermometer," a series of "choice tasks," and a "chance board." Shall we begin?
“FEELING THERMOMETER TASKS”

This is a visual aid called a "feeling thermometer."

DISPLAY FEELING THERMOMETER

We call this a feeling thermometer because it helps us to measure people's feelings about different things. In this interview, we are going to use the thermometer to measure your preferences for different outcomes of drug therapy.

This is how it works. The thermometer scale will indicate the way you feel about different outcomes of drug therapy. I will be showing you a series of cards, each of which describes different outcomes of drug therapy. The more preferable you feel an outcome to be, the closer it should be toward the top end of the thermometer.

RUN FINGER UP SCALE FROM 50 TO 100

At the top of the thermometer is a mark saying "best imaginable health state," which has a score of 100 -- this represents the evaluation of the most preferable outcome.

POINT TO TOP OF SCALE AT "BEST IMAGINABLE HEALTH STATE"
The less preferable you feel an outcome to be, the closer it should be toward the bottom end of the thermometer.

RUN FINGER DOWN SCALE FROM 50 TO 0

At the bottom of the thermometer is a mark saying "worst imaginable health state," which has a score of 0 -- this represents the evaluation of the least preferable outcome.

POINT TO BOTTOM OF SCALE AT "WORST IMAGINABLE HEALTH STATE"

Let's begin with a trial run. I'd like you to imagine that you have been involved in an accident and have hurt your leg. When you see the doctor, he explains that two options -- to undergo an operation or not. Before I proceed any further, let me give you these three cards and arrows.

GIVE PATIENT THE THREE CARDS & ARROWS ON OPERATION

These are the three cards which describe each outcome possible. For instance, you may undergo the operation and walk and run normally for the rest of your life, undergo the operation and use crutches for the rest of your life, or decide to let your leg heal by itself and walk with a limp for the rest of your life. The arrows correspond to each card.
Here's what you need to do. First, decide which of the three outcomes is most desirable to you. Place the arrow corresponding to that outcome on the point of the scale labeled "best imaginable health state." Then place the arrow corresponding to the outcome you feel is least desirable on the point of the scale labeled "worst imaginable health state." Finally, place the remaining arrow on any point along the scale so that it best represents your feelings as compared to the outcomes already on the scale.

WAIT FOR PATIENT TO COMPLETE TRIAL RUN

You are now ready to begin the study tasks. At first, please rate the different types of outcomes within each major aspect of health individually. For instance, for the aspect of health termed mobility, please rate each of the three different types of outcomes of drug therapy on mobility, i.e., no problems walking about, some problems walking about, and confined to bed. To help you in this rating task, we have prepared three cards describing each outcome and three arrows with labels corresponding to the labels on the cards.

GIVE PATIENT THE THREE CARDS & ARROWS ON MOBILITY

These are the three cards which describe each outcome of drug therapy on mobility (i.e., the ability to walk about). You will notice that each card has a label on the top right-hand side corner.

POINT TO LABEL ON CARDS
I have also given you three arrows with labels corresponding to the labels on the cards.

---

**POINT TO LABEL ON ARROWS**

**NOTE CORRESPONDENCE BETWEEN THE TWO SETS OF LABELS**

Notice that both labels are the same. This means that the arrow should be used to represent the description provided on the card. The reason for giving you both cards as well as arrows is that it is easier to stick the arrows on the board, since they take up less space than the cards. For rating each of these three outcomes, here is what you need to do.

1. Read the three cards carefully.
2. Arrange the cards in decreasing order of preference. In other words, place the outcome you feel is most preferable on the top, the second most preferable outcome in the middle, and the least preferable outcome at the bottom of the stack of cards.
3. Select the arrow with the label of the outcome you feel is the most preferable.
4. Place this arrow on the left-hand side of the thermometer at the mark saying "best imaginable health state," i.e., at the 100-mark. Please make sure that the tip of the arrow touches the scale.
5. Select the arrow with the label of the outcome you feel is the least preferable.
6. Place this arrow on the left-hand side of the thermometer at the mark saying "worst imaginable health state," i.e., at the 0-mark. Please make sure that the tip of the arrow touches the scale.
7. Select the remaining arrow, with the label of the remaining outcome.
8. Place this arrow on the left-hand side of the thermometer at the value which best represents your preference for this outcome as compared to the two outcomes already on the board. Please make sure that the tip of the arrow touches the scale.
9. **Please check that the third outcome you just rated is placed at such a point that is best represents your feelings as compared to the outcomes already on the thermometer.**
WAIT FOR PATIENT TO COMPLETE THE TASK

Are you satisfied that these evaluations reflect your preferences for these outcomes? If so, we are now ready to record the scores for each arrow.

NOTE DOWN SCORES ON DATA COLLECTION FORM

The next step is to repeat this task for all the remaining four aspects of health, i.e., self-care, usual activities, pain/discomfort, and anxiety/depression.

REPEAT THE ABOVE PROCEDURE FOR THE OTHER ATTRIBUTES
IN RANDOM ORDER

NOTE DOWN SCORES ON DATA COLLECTION FORM

That finishes the first part of the feeling thermometer task, which is measuring your preferences for individual outcomes of drug therapy. The next step is combining these different outcomes to generate a profile of hypothetical outcomes. *It is important to note that each profile of hypothetical outcomes shown consists of one type of outcome for each of the five major aspects of health.* In other words, each profile consists of one outcome for mobility, one outcome for self-care, one outcome for usual activities, one outcome for pain/discomfort, and one outcome for anxiety/depression.
I will hand you another series of cards. You will note that the aspects of health are listed in the same order on each card, but the particular outcome for one or more of the aspects of health will change on every card. Therefore, please study each card very carefully before making your evaluation.

When thinking about the profile of health described in each card, imagine that it will last for one year. What happens after that is not known and should not be taken into account. In order to make your task a little easier, instead of handing out all ten cards at the same time, I will give them to you in sets. Here is the first card. As before, each card has a label. I will also be giving you an arrow along with each card. You will notice that each arrow has a label corresponding to the labels on the cards.

GIVE PATIENT CARD AND ARROW FOR “ALL-BEST”

This first card describes a state of the best possible health. This is because it has the best outcomes for all the aspects of health. Therefore, it should be placed at the top of the thermometer, at the “best imaginable health state” or 100-mark.

Now, remember how you do this. Place the arrow on the left-hand side of the thermometer at the mark saying “best imaginable health state,” i.e., at the 100-mark. Please make sure that the tip of the arrow touches the scale.

WAIT FOR PATIENT TO PLACE ARROW ON THERMOMETER

Now, let me give you the next two cards and arrows.
GIVE PATIENT CARDS AND ARROWS FOR “ALL-WORST” AND “IMMEDIATE DEATH”

Both these cards describe states which are quite undesirable. This card here

POINT TO “ALL-WORST” CARD

has the worst outcomes for all the aspects of health. On the other hand, this card here

POINT TO “IMMEDIATE DEATH” CARD

represents a state of immediate death. Please note that your being asked to evaluate the state of immediate death has nothing whatsoever to do with your current state of health. We need this evaluation for the purpose of understanding how other states of health compare to immediate death in your opinion. If, however, you in any way feel uncomfortable with having to think about the state of immediate death, please tell me and we will stop this interview right then and there -- and further, this decision of yours will not affect the quality of care you receive at the VA.

Now, which of these two cards do you think is worse? Please place the arrow corresponding to this card on the bottom of the scale at the “worst imaginable health state” or 0-mark.
Now, please place the remaining arrow on any point along the scale so that it reflects your feelings as to how good or bad it is compared to the other two arrows already on the scale.

Thank you. I'm now going to give you eight different cards and arrows. Instead of giving you them altogether, I will give them to you one at a time. Please read each one carefully and follow exactly the same steps you took for rating the previous cards. If you have any questions as you are going through this procedure, please do not hesitate to ask me. Please remember to place these arrows at such points so they best represent your feelings as compared to the arrows already on the thermometer.
We are almost done with the feeling thermometer exercise. Before I note down your ratings, are there any changes that you would like to make?

PAUSE UNTIL PATIENT INDICATES SATISFACTORY COMPLETION OF ANY REVISIONS

THEN RECORD SCORES ON DATA COLLECTION FORM

Before we proceed to the next exercise, I would like to obtain some feedback from you regarding the feeling thermometer exercise. What I have here is a short questionnaire.

TAKE OUT EVALUATION FORM FOR FEELING THERMOMETER

Regarding the feeling thermometer tasks you just completed, did you find them boring or interesting?

WAIT FOR PATIENT TO ANSWER

Would you say they were “extremely,” “moderately,” or “slightly” ______?
Did you think they were sensible or senseless?

Would you say they were “extremely,” “moderately,” or “slightly” _____?

Did you think they were difficult to do or easy to do?

Would you say they were “extremely,” “moderately,” or “slightly” _____?
WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Do you think they will be useful or useless?

WAIT FOR PATIENT TO ANSWER

Would you say they will be “extremely,” “moderately,” or “slightly” _____?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Did you find them to be long or short?

WAIT FOR PATIENT TO ANSWER

Would you say they were “extremely,” “moderately,” or “slightly” ______?
The purpose of these tasks was to determine your preferences for drug therapy outcomes. In this regard, did you think these tasks were successful or unsuccessful?

Would you say they were "extremely," "moderately," or "slightly" _______?

Did you find these tasks were clear or vague?
Would you say they were “extremely,” “moderately,” or “slightly” ____?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Did you find these tasks difficult to understand or easy to understand?

WAIT FOR PATIENT TO ANSWER

Would you say they were “extremely,” “moderately,” or “slightly” ____?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Finally, how confident are you that the feeling thermometer tasks did a good job of measuring your preferences for drug therapy outcomes?
SHOW PATIENT RESPONSE CATEGORIES

LET PATIENT INDICATE APPROPRIATE RESPONSE CATEGORY

CHECK APPROPRIATE RESPONSE CATEGORY

Thank you! We are now ready to move onto the next exercise. Are you ready or do you want a breather?
"CHOICE TASKS"

In this exercise, you will go through a series of choice tasks, by means of which you will reveal your preferences for drug therapy outcomes. In order to make this exercise easier, I will be displaying each choice task on this computer.

SHOW PATIENT COMPUTER

TURN ON POWER OF COMPUTER

Each choice task consists of three alternative drug therapies and a fourth alternative of “immediate death.” The outcomes of these three drug therapies on the five major aspects of your health are described by descriptive phrases.

Please note that your being asked to evaluate the state of immediate death has nothing whatsoever to do with your current state of health. We need this evaluation for the purpose of understanding how other states of health compare to immediate death in your opinion. If, however, you in any way feel uncomfortable with having to think about the state of immediate death, please tell me and we will stop this interview right then and there -- and further, this decision of yours will not affect the quality of care you receive at the VA.

Please take as long as you want to study each alternative carefully. When thinking about the profile of health described in each alternative, imagine that it will last for one year. What happens after that is not known and should not be taken into account. After evaluating each alternative, make your selection. You will need to press the number of the alternative (which is given at the top of each profile) you pick in order to indicate your choice. You may enter your selection directly on the computer.
There are twelve such choice tasks that you will need to perform. Please follow the same instructions for the remaining eleven choice tasks. After completing all choice tasks, I will obtain feedback from you about your opinion of the exercise. This feedback is very important in making revisions for use in future extensions of this study.

At this point are there any questions? Just to make sure you are comfortable doing the choice tasks, let's do a practice run.

Here goes for the practice run. I will be giving you one practice choice task. This trial choice task will be structured in exactly the same manner as the "final" choice tasks -- that is, it will contain three alternatives to choose among, each described in terms of their outcomes on the five major aspects of your health and a fourth alternative of immediate death. Similar to your task for the "final" choice tasks, you have to choose ONE of these four alternatives.

ADMINISTER TRIAL RUN
ENTERTAIN ANY QUESTIONS

Let's see what we have here. The first drug therapy profile is _________.

HELP PATIENT IN INTERPRETING EACH ALTERNATIVE IF NEEDED
There (AFTER PATIENT MAKES A CHOICE). You are now ready for the "final" choice tasks. There is nothing different about these "final" choice tasks as compared to the trial choice task. Just do as you did for the trial task. Remember, the outcomes of the alternative drug therapies are expected for the next ONE year of your life, and what happens after that is not known and should not be taken into account. Do you have any questions before we begin?

WAIT FOR QUESTIONS

ANSWER THEM AND THEN PROCEED

ADMINISTER STUDY CONJOINT TASKS

You're almost done with the choice task exercise.

Before we proceed to the next exercise, I would like to obtain some feedback from you regarding the choice task exercise. What I have here is a short questionnaire.

TAKE OUT EVALUATION FORM FOR CHOICE TASKS

Regarding the choice tasks you just completed, did you find them boring or interesting?
WAIT FOR PATIENT TO ANSWER

Would you say they were "extremely," "moderately," or "slightly" ______?

WAIT FOR PATIENT TO ANSWER

CHECK APPROPRIATE RESPONSE CATEGORY

Did you think they were sensible or senseless?

WAIT FOR PATIENT TO ANSWER

Would you say they were "extremely," "moderately," or "slightly" ______?

WAIT FOR PATIENT TO ANSWER

CHECK APPROPRIATE RESPONSE CATEGORY

Did you think they were difficult to do or easy to do?
WAIT FOR PATIENT TO ANSWER

Would you say they were "extremely," "moderately," or "slightly" _____?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Do you think they will be useful or useless?

WAIT FOR PATIENT TO ANSWER

Would you say they will be "extremely," "moderately," or "slightly" _____?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Did you find them to be long or short?
Would you say they were “extremely,” “moderately,” or “slightly” ______?

The purpose of these tasks was to determine your preferences for drug therapy outcomes. In this regard, did you think these tasks were successful or unsuccessful?

Would you say they were “extremely,” “moderately,” or “slightly” ______?
Did you find these tasks were clear or vague?

WAIT FOR PATIENT TO ANSWER

Would you say they were "extremely," "moderately," or "slightly" _____?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Did you find these tasks difficult to understand or easy to understand?

WAIT FOR PATIENT TO ANSWER

Would you say they were "extremely," "moderately," or "slightly" _____?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY
Finally, how confident are you that the choice tasks did a good job of measuring your preferences for drug therapy outcomes?

SHOW PATIENT RESPONSE CATEGORIES

LET PATIENT INDICATE APPROPRIATE RESPONSE CATEGORY

CHECK APPROPRIATE RESPONSE CATEGORY

Thank you! We are now ready to move onto the next exercise. Are you ready or do you want a breather?
"CHANCE TASKS"

This exercise will help you reveal your preferences for drug therapy outcomes by evaluating different drug therapy alternatives on the basis of a certain chance that they will produce a given outcome.

This exercise will consist of a series of chance tasks. In each chance task, I will show you three different descriptions of the outcomes of drug therapy on the five major aspects of health. One of these descriptions will be certain to occur. The other two, however, can occur with only some chance.

Your task is to decide between two choices -- the first choice will be the certain description while the second will involve taking a chance that the other descriptions may occur. I will keep changing this amount of chance until we find how much chance you will take to avoid the certain choice.

As before, there are no right or wrong answers. All we are concerned with is what YOU think. In order to make this task easier for you, I will be using this device called a "chance board."

---

SHOW PATIENT CHANCE BOARD

---

We call this a chance board because it indicates the chance or probability of a particular event occurring. As you can see, the top part of the board is labeled Choice "A."

---

POINT FINGER AT 'CHOICE "A"'
The bottom part of the board is labeled Choice "B."

POINT FINGER AT 'CHOICE "B"

You will be asked to pick Choice A or Choice B. Choice B, at the bottom of the board, will describe an alternative drug therapy profile that is neither the best nor the worst of the three alternative descriptions.

Choice B is fairly simple since it describes only one form of health. If Choice B is chosen then this form of health is certain to occur. If something is certain, it is equal to 100% chance or probability. **Remember, this form of health described will last for one year. What happens after that is not known and should not be taken into account.**

However, Choice A is a little more complex. This is because if Choice A is chosen, there are two possible results -- usually an extremely desirable one and an extremely undesirable one. For instance, the two results could be perfect health for one year and immediate death. The chances of each of these two results occurring are indicated by the numbers appearing above each result and the size of the matching colors in the circle between the numbers.

To make these ideas a little clearer, let's run through a quick, imaginary example. Let us assume you have been involved in an accident and have hurt your leg. When you see the doctor, he explains that you have two choices. Here are the descriptions for this example.
Choice A is an operation and Choice B is to let your leg heal by itself. Let me explain each of these choices in a little more detail. If you choose Choice B, i.e., letting your leg heal by itself, it is certain that you will have a limp. You will be able to walk, but you will NOT be able to run.

On the other hand, you can choose the operation, i.e., Choice A. This operation is risky. It does not always work. If the operation does work, your leg will be fixed and you can walk and run normally. If the operation does not work, you will have to use crutches.

The chance of walking and running normally after the operation is shown above the description provided on the card.

POINT TO PINK CHANCE VALUE

and matches the proportion of pink color in this circle.

POINT TO PINK SECTION OF CIRCLE
The chance of having to use crutches after the operation is shown above the description provided on the card:

---

**POINT TO BLUE CHANCE VALUE**

---

and matches the proportion of blue color in this circle:

---

**POINT TO BLUE SECTION OF CIRCLE**

---

These chances will change during the course of the exercise. Your task is to choose Choice A or Choice B each time I change the chances like this:

---

**DEMONSTRATE BY TURNING WHEEL ON CHANCE BOARD**

**RE-ADJUST CHANCES TO 90/10**

---

Here, the chances of walking and running normally are 90%, while the chances of having to use crutches are 10%. What this means is that if I were to spin this arrow,
chances are that it will *probably* land on the pink shaded area because there is a very large amount of pink -- 90% -- showing in the circle. There is still a chance that the arrow might land on the blue shaded area, but it is a smaller chance since there is only a small amount of blue showing on the circle.

Another way to think of this is that, on average, for every 100 people who choose Choice A (i.e., the operation), 90

will walk and run normally

but 10
will have to use crutches after the operation.

POINT TO CARD ON RIGHT-HAND SIDE

Given this information, your task is to choose between Choices A and B.

SET CHANCES TO 20/80

I have now changed the chances of these results occurring. You may now note that the chances of walking and running normally are 20%, while the chances of having to use crutches are 80%. What this means is that if I were to spin this arrow,

POINT TO ARROW IN CIRCLE

chances are that it will probably land on the blue shaded area because there is a very large amount of blue -- 80% -- showing in the circle. There is still a chance that the arrow might land on the pink shaded area, but it is a smaller chance since there is only a small amount of pink showing on the circle.

Another way to think of this is that, on average, for every 100 people who choose Choice A (i.e., the operation), 20
will walk and run normally

but 80

will have to use crutches after the operation.

Given this information, your task is to choose between Choices A and B.
CHANGE CHANCES TO 50/50

I have again changed the chances. If I spin the arrow now, what color is it most likely to stop on?

ALLOW PATIENT TO ANSWER

There is an equal chance that it will stop on pink or blue, because there is an equal amount of pink and blue showing in the circle. So, in terms of the descriptions of health provided by the cards what does this mean?

ALLOW PATIENT TO ANSWER

Do you understand how the chance board works?
We are now ready to begin the first chance task.

These three cards describe the different outcomes of drug therapy on mobility, i.e., the ability to walk about. Arrange the cards in decreasing order of preference. In other words, place the outcome you feel is most preferable on the top, the second most preferable outcome in the middle, and the least preferable outcome at the bottom of the stack of cards. When you are finished, please let me know.

Thank you. Now, I'm going to place the topmost card -- i.e., the one you felt was the most preferable -- on this pocket of the chance board.
PLACE MOST PREFERABLE OUTCOME IN POCKET OVER LEFT-HAND SIDE OF CHOICE “A”

Next, I’m going to place the bottommost card -- i.e., the one you felt was the least preferable -- on this pocket of the chance board.

PLACE LEAST PREFERABLE OUTCOME OVER POCKET ON RIGHT-HAND SIDE OF CHOICE “A”

You may remember that I’d told you that Choice A has two possible results - an extremely desirable one and an extremely undesirable one. In this case, the extremely desirable result is this one,

POINT TO MOST PREFERABLE OUTCOME OVER POCKET ON LEFT-HAND SIDE OF CHOICE “A”

while the extremely undesirable one is this one.

POINT TO LEAST PREFERABLE OUTCOME OVER POCKET ON RIGHT-HAND SIDE OF CHOICE “A”
The chances of each of these two results occurring are indicated by the numbers appearing above each result and the size of the matching colors in the circle between the numbers. Do you understand what Choice A involves? I realize this may be somewhat confusing, so please do not hesitate to ask me any questions. It is very important for this study that you fully understand what Choice A involves.

ENTERTAIN ANY QUESTIONS

If you understand Choice A, then Choice B should be easy to understand. Choice B is nothing but this third outcome which is certain to occur.

PICK UP REMAINING OUTCOME

PUT IN POCKET ON CHANCE BOARD OVER CHOICE "B"

SET CHANCES TO 100/0

Your task is to tell me which option you would choose, Choice A or Choice B. You may also indicate if you are indifferent between the two options. As you can see on the chance board, Choice A is now a 100% chance of having no problems walking about or a 0% chance of being confined to bed. Choice B is a 100% chance of the description provided on the card, i.e., having some problems walking about. Please remember that each of these cards describes the outcomes for a period of one year. What happens after that is not known and should not be taken into account. Which option would you pick, Choice A or Choice B? Or are you indifferent between the two options?
WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “A,” SET CHANCES TO 10/90 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (1.00) THEN GO TO NEXT QUESTION

IF PATIENT PICKS “B,” CIRCLE RESPONSE (?) AND ASK PROMPT:
“Why did you choose a 100% chance of being confined to bed rather than a 100% chance of having no problems walking about?” RECORD VERBATIM RESPONSE THEN GO TO NEXT QUESTION

Now, I’ve changed Choice A to show that there is a 10% chance of having no problems walking about and a 90% chance of being confined to bed. Choice B is still a 100% chance of having some problems walking about. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?
WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “B,” SET CHANCES TO 90/10 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.10)
THEN GO TO NEXT QUESTION

IF PATIENT PICKS “A,” ASK PROMPT,
“Do you mean you would prefer to have a 90% chance of being confined to bed and a 10% chance of having no problems walking about rather than having some problems walking about?”
IF NO, REPEAT CHOICES ON BOARD
IF YES, ASK,
“If Choice A was certainly being confined to bed with no chance of having no problems walking about, do you think this would be better than Choice B, equal to Choice B, or worse than Choice B?”
IF BETTER, MARK RESPONSE (NEGATIVE)
IF EQUAL, MARK RESPONSE (0.00)
IF WORSE, MARK RESPONSE (0.05)
THEN GO TO NEXT QUESTION

The board now shows Choice A to be a 90% chance of having no problems walking about with a 10% chance of being confined to bed and Choice B still a 100% chance of having some problems walking about. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?
I have now changed Choice A to a 20% chance of having no problems
walking about with an 80% chance of being confined to bed. Choice B is still
the 100% chance of having some problems walking about. As before, please
remember that the outcomes shown are for a period of one year and what
happens after that is not known and should not be taken into account.
Would you pick Choice A or Choice B now? Or are you indifferent between
the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “A,” SET CHANCES TO 20/80 AND
CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.90)
THEN GO TO NEXT QUESTION

IF PATIENT PICKS “B,” MARK RESPONSE (0.95)
THEN GO TO NEXT QUESTION
Choice A is now an 80% chance of having no problems walking about with a 20% chance of being confined to bed, while Choice B remains the 100% chance of having some problems walking about. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "A," SET CHANCES TO 30/70 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.80) THEN GO TO NEXT QUESTION

IF PATIENT PICKS "B," MARK RESPONSE (0.85) THEN GO TO NEXT QUESTION

The choices have now been changed so that Choice A has a 30% chance of having no problems walking about but a 70% chance of being confined to bed. Choice B remains the 100% chance of having some problems walking about. Remember the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?
WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "B," SET CHANCES TO 70/30 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.30) THEN GO TO NEXT QUESTION

IF PATIENT PICKS "A," MARK RESPONSE (0.25) THEN GO TO NEXT QUESTION

I have now changed Choice A to a 70% chance of having no problems walking about and a 30% chance of being confined to bed. Choice B remains the 100% chance of having some problems walking about. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "A," SET CHANCES TO 40/60 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.70) THEN GO TO NEXT QUESTION

IF PATIENT PICKS "B," MARK RESPONSE (0.75) THEN GO TO NEXT QUESTION
Choice A has now been adjusted to indicate a 40% chance of having no problems walking about and a 60% chance of being confined to bed. Choice B remains the 100% chance of having some problems walking about. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “B,” SET CHANCES TO 60/40 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.40)
THEN GO TO NEXT QUESTION

IF PATIENT PICKS “A,” MARK RESPONSE (0.35)
THEN GO TO NEXT QUESTION

I have now changed Choice A to a 60% chance of having no problems walking about and a 40% chance of being confined to bed. Choice B remains the 100% chance of having some problems walking about. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?
WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "A," SET CHANCES TO 50/50 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.60)
THEN GO TO NEXT QUESTION

IF PATIENT PICKS "B," MARK RESPONSE (0.65)
THEN GO TO NEXT QUESTION

Now, I've changed Choice A to a 50% chance of having no problems walking about and a 50% chance of being confined to bed. Choice B remains the 100% chance of having some problems walking about. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "A," MARK RESPONSE (0.45)

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.50)

IF PATIENT PICKS "B," MARK RESPONSE (0.55)

THEN GO TO NEXT CHANCE TASK
Thank you. That ends this particular chance task. We will now do a similar chance task for each aspect of health, i.e., self-care, usual activities, pain/discomfort, and anxiety/depression.

**GIVE PATIENT CARDS REMAINING SINGLE ATTRIBUTES IN RANDOM ORDER**

**FOLLOW THE SAME PROCEDURE AS FOR MOBILITY**

The final part of the chance exercise is to evaluate profiles of drug therapies in the same way you have just evaluated single outcomes of drug therapy. For all the remaining evaluations, the alternatives of Choice A will remain the same. Over the left-hand pocket is this card,

**SHOW PATIENT ALL-BEST CARD**

while over the right-hand pocket is this card.

**SHOW PATIENT IMMEDIATE DEATH CARD**

As you remember, Choice A has two possible results -- an extremely desirable one and an extremely undesirable one. In this case, the extremely desirable result is this one,
while the extremely undesirable one is this one.

The chances of each of these two results occurring are indicated by the numbers appearing above each result and the size of the matching colors in the circle between the numbers. Do you understand what Choice A involves? I realize this may be somewhat confusing, so please do not hesitate to ask me any questions. It is very important for this study that you fully understand what Choice A involves.

Now, as I'd said, the alternatives of Choice A will remain the same for all the remaining chance tasks. The chance of them occurring, however, will be varied in each chance task. Further, in each remaining chance task there will be a different Choice B. As before, Choice B is a description of the outcomes of a drug therapy alternative which are certain (i.e., 100% chance) to occur.
We are now ready to begin. Let me give you the first drug therapy profile to study.

---

**GIVE PATIENT FIRST HEALTH STATE**

**SELECTED AT RANDOM**

---

Please read over the description and when you are finished I will put it on pocket B at the bottom of the chance board.

---

**PUT CARD OVER POCKET ON CHANCE BOARD IN CHOICE “B”**

**SET CHANCES TO 100/0**

---

Your task is to tell me which option you would choose, Choice A or Choice B. You may also indicate if you are indifferent between the two options. As you can see on the chance board, Choice A is now a 100% chance of having perfect health or a 0% chance of immediate death. Choice B is a 100% chance of the description provided on the you just finished reading. Please remember that each of these cards describes the outcomes for a period of one year. What happens after that is not known and should not be taken into account. Which option would you pick, Choice A or Choice B? Or are you indifferent between the two options?
WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "A," SET CHANCES TO 10/90 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (1.00)
THEN GO TO NEXT QUESTION

IF PATIENT PICKS "B," CIRCLE RESPONSE (?) AND ASK PROMPT:
"Why did you choose a 100% chance of immediate death rather than a 100% chance of perfect health?"
RECORD VERBATIM RESPONSE
THEN GO TO NEXT QUESTION

Now, I’ve changed Choice A to show that there is a 10% chance perfect health and a 90% chance of immediate death. Choice B is still a 100% chance of the state of health described on the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?
WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "B," SET CHANCES TO 90/10 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.10)
THEN GO TO NEXT QUESTION

IF PATIENT PICKS "A," ASK PROMPT,
"Do you mean you would prefer to have a 90% chance of dying immediately and a 10% chance of perfect health rather than living in the state of health described in the card in Choice B?"

IF NO, REPEAT CHOICES ON BOARD
IF YES, ASK,
"If Choice A was certain immediate death with no chance perfect health, do you think this would be better than living as Choice B, equal to living as Choice B, or worse than living as Choice B?"

IF BETTER, MARK RESPONSE (NEGATIVE)
IF EQUAL, MARK RESPONSE (0.00)
IF WORSE, MARK RESPONSE (0.05)
THEN GO TO NEXT QUESTION

The board now shows Choice A to be a 90% chance of perfect health with a 10% chance of immediate death and Choice B still a 100% chance living in the state of health described on the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?
I have now changed Choice A to a 20% chance of perfect health with an 80% chance of immediate death. Choice B is still the 100% chance living in the state of health described in the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

---

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "A," SET CHANCES TO 20/80 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.90) THEN GO TO NEXT QUESTION

IF PATIENT PICKS "B," MARK RESPONSE (0.95) THEN GO TO NEXT QUESTION

---

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "B," SET CHANCES TO 80/20 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.20) THEN GO TO NEXT QUESTION

IF PATIENT PICKS "A," MARK RESPONSE (0.15) THEN GO TO NEXT QUESTION
Choice A is now an 80% chance of perfect health with a 20% chance immediate death, while Choice B remains the 100% chance of living in the state of health described in the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS "A," SET CHANCES TO 30/70 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.80) THEN GO TO NEXT QUESTION

IF PATIENT PICKS "B," MARK RESPONSE (0.85) THEN GO TO NEXT QUESTION

The choices have now been changed so that Choice A has a 30% chance of perfect health but a 70% chance of immediate death. Choice B remains the 100% chance of living in the state of health described in the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?
I have now changed Choice A to a 70% chance of perfect health and a 30% chance of immediate death. Choice B remains the 100% chance living in the state of health described in the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “B,” SET CHANCES TO 70/30 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.30) THEN GO TO NEXT QUESTION

IF PATIENT PICKS “A,” MARK RESPONSE (0.25) THEN GO TO NEXT QUESTION

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “A,” SET CHANCES TO 40/60 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.70) THEN GO TO NEXT QUESTION

IF PATIENT PICKS “B,” MARK RESPONSE (0.75) THEN GO TO NEXT QUESTION
Choice A has now been adjusted to indicate a 40% chance of perfect health and a 60% chance of immediate death. Choice B remains the 100% chance of living in the state of health described in the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “B,” SET CHANCES TO 60/40 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.40) THEN GO TO NEXT QUESTION

IF PATIENT PICKS “A,” MARK RESPONSE (0.35) THEN GO TO NEXT QUESTION

I have now changed Choice A to a 60% chance of perfect health and a 40% chance of immediate death. Choice B remains the 100% chance of living in the state of health described in the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?
WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “A,” SET CHANCES TO 50/50 AND CONTINUE AS BELOW

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.60)
THEN GO TO NEXT QUESTION

IF PATIENT PICKS “B,” MARK RESPONSE (0.65)
THEN GO TO NEXT QUESTION

Now, I’ve changed Choice A to a 50% chance perfect health and a 50% chance of immediate death. Choice B remains the 100% chance of living in the state of health described in the card. As before, please remember that the outcomes shown are for a period of one year and what happens after that is not known and should not be taken into account. Would you pick Choice A or Choice B now? Or are you indifferent between the two options?

WAIT FOR PATIENT TO ANSWER

IF PATIENT PICKS “A,” MARK RESPONSE (0.45)

IF PATIENT IS INDIFFERENT, MARK RESPONSE (0.50)

IF PATIENT PICKS “B,” MARK RESPONSE (0.55)

THEN GO TO NEXT CHANCE TASK

Thank you. That ends this particular chance task. We will now do a similar chance task for the last seven cards.
GIVE PATIENT CARDS ONE AT A TIME AT RANDOM
FOR EACH, FOLLOW THE SAME PROCEDURE AS ABOVE

You're almost done with the chance task exercise.

Before we proceed to the next exercise, I would like to obtain some feedback from you regarding the chance task exercise. What I have here is a short questionnaire.

TAKE OUT EVALUATION FORM FOR CHANCE TASKS

Regarding the chance tasks you just completed, did you find them boring or interesting?

WAIT FOR PATIENT TO ANSWER

Would you say they were “extremely,” “moderately,” or “slightly” ______?
WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Did you think they were sensible or senseless?

WAIT FOR PATIENT TO ANSWER

Would you say they were “extremely,” “moderately,” or “slightly” ______?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Did you think they were difficult to do or easy to do?

WAIT FOR PATIENT TO ANSWER

Would you say they were “extremely,” “moderately,” or “slightly” ______?
WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Do you think they will be useful or useless?

WAIT FOR PATIENT TO ANSWER

Would you say they will be "extremely," "moderately," or "slightly" _____?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Did you find them to be long or short?

WAIT FOR PATIENT TO ANSWER

Would you say they were "extremely," "moderately," or "slightly" _____?
The purpose of these tasks was to determine your preferences for drug therapy outcomes. In this regard, did you think these tasks were successful or unsuccessful?

Would you say they were "extremely," "moderately," or "slightly" ______?

Did you find these tasks were clear or vague?
Would you say they were “extremely,” “moderately,” or “slightly” ______? 

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Did you find these tasks difficult to understand or easy to understand?

WAIT FOR PATIENT TO ANSWER

Would you say they were “extremely,” “moderately,” or “slightly” ______?

WAIT FOR PATIENT TO ANSWER
CHECK APPROPRIATE RESPONSE CATEGORY

Finally, how confident are you that the chance tasks did a good job of measuring your preferences for drug therapy outcomes?
SHOW PATIENT RESPONSE CATEGORIES

LET PATIENT INDICATE APPROPRIATE RESPONSE CATEGORY

CHECK APPROPRIATE RESPONSE CATEGORY

Thank you! We are now ready to move onto the next exercise. Are you ready or do you want a breather?
Thank you very much for your time. Finally, I'd like to get some feedback from you about some aspects of this study. Please note that your feedback is very important to us and will be used in future research. Please answer all questions as truthfully as possible and do not feel hesitant about making any negative comments about any aspect of the study.

First of all, I'd like you to think about the three tasks you just completed -- the feeling thermometer tasks, the choice tasks, and the chance tasks. Now, out of the three of these tasks, which did you think was the most interesting?

WAIT FOR PATIENT'S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM

And the second most interesting?

WAIT FOR PATIENT'S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM

Which of the three tasks did you think was the most senseless?

WAIT FOR PATIENT'S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM
And the second most senseless?

WAIT FOR PATIENT’S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM

Which of the three tasks did you think was the most difficult to do?

WAIT FOR PATIENT’S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM

And the second most difficult to do?

WAIT FOR PATIENT’S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM

Which of the three tasks do you think would be the most useful?
WAIT FOR PATIENT'S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM

And the second most useful?

WAIT FOR PATIENT'S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM

Which of the three tasks did you think was the longest?

WAIT FOR PATIENT'S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM

And the second longest?

WAIT FOR PATIENT'S RESPONSE
RECORD RESPONSE IN DATA COLLECTION FORM
Which of the three tasks did you think was the most successful?

WAIT FOR PATIENT'S RESPONSE

RECORD RESPONSE IN DATA COLLECTION FORM

And the second most successful?

WAIT FOR PATIENT'S RESPONSE

RECORD RESPONSE IN DATA COLLECTION FORM

Which of the three tasks did you think was the most clear?

WAIT FOR PATIENT'S RESPONSE

RECORD RESPONSE IN DATA COLLECTION FORM

And the second most clear?
Which of the three tasks did you think was the most difficult to understand?

And the second most difficult to understand?

Please rank the three tasks in terms of how confident you are that they did a good job of measuring your preferences for drug therapy outcomes.
Considering all of the above factors, what is your overall ranking of the three tasks?

Finally, we would like to get your feedback about some other general aspects of the interview.

The descriptions of the outcomes of drug therapy were presented in terms of five different aspects of health, i.e., mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Do you feel that there are any other aspects of health that we need to consider?
Finally, do you have any other comments about any aspect of this study?

WAIT FOR PATIENT’S RESPONSE

RECORD RESPONSE IN DATA COLLECTION FORM

Do you have any questions for me?

WAIT FOR ANY QUESTIONS AND ANSWER THEM

Thank you very much for your time!
APPENDIX D

DATA COLLECTION FORM FOR RATING SCALE SCALING METHOD
RESPONDENT NUMBER

MEASURING PREFERENCES FOR HEALTH OUTCOMES

Data Collection Form for Feeling Thermometer Tasks

Start Time __________
End Time __________
### 1. Single Attributes

<table>
<thead>
<tr>
<th>CARD</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td></td>
</tr>
<tr>
<td>U2</td>
<td></td>
</tr>
<tr>
<td>U3</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td></td>
</tr>
</tbody>
</table>
2. Multi-attribute States

<table>
<thead>
<tr>
<th>CARD</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td></td>
</tr>
<tr>
<td>. AW</td>
<td></td>
</tr>
<tr>
<td>CM</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td></td>
</tr>
<tr>
<td>CU</td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>BO</td>
<td></td>
</tr>
<tr>
<td>E3</td>
<td></td>
</tr>
<tr>
<td>E8</td>
<td></td>
</tr>
<tr>
<td>HT</td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX E

DATA COLLECTION FORM FOR
STANDARD GAMBLE SCALING METHOD
MEASURING PREFERENCES FOR HEALTH OUTCOMES

Data Collection Form for Chance Tasks

Start Time

End Time
1. Single Attributes

<table>
<thead>
<tr>
<th>CARD</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>S1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>U1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>U2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>U3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>P2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>P3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td></td>
</tr>
</tbody>
</table>
### 2. Multi-attribute States

<table>
<thead>
<tr>
<th>CARD</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td></td>
</tr>
<tr>
<td>AW</td>
<td></td>
</tr>
<tr>
<td>CM</td>
<td></td>
</tr>
<tr>
<td>CS</td>
<td></td>
</tr>
<tr>
<td>CU</td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td></td>
</tr>
<tr>
<td>BO</td>
<td></td>
</tr>
<tr>
<td>E3</td>
<td></td>
</tr>
<tr>
<td>E8</td>
<td></td>
</tr>
<tr>
<td>HT</td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td></td>
</tr>
</tbody>
</table>
Thank you for agreeing to participate in this study. The purpose of this study is to understand the preferences of patients such as yourself for the outcomes of drug therapy.

Over the next few screens, you will be shown profiles of hypothetical outcomes of several drug therapies. Each profile will be described in terms of the outcomes of the drug therapy on five major aspects of your health:

(a) mobility
(b) self-care
(c) usual activities
(d) pain/discomfort
(e) anxiety/depression

Press any key to continue.
Please note that the outcomes of each drug therapy on the five major aspects of your health are described by descriptive phrases. Please refer to the handout for the full versions of these phrases and the proper way to interpret each drug therapy profile.

Each screen reflects a choice task. Each choice task consists of four alternatives. For each screen, please indicate which ONE alternative you would choose from the group on the basis of the described outcomes. You will need to press the number of the alternative (given on the top of each profile) you pick in order to indicate your choice.

Press any key to continue.
While performing each choice task, please remember the following:

(a) there are no right or wrong answers. Since it is your preferences which are of central interest to this study, whatever answer you give is the correct one

(b) please take as long as you want to evaluate each alternative. The choice tasks are difficult and there is no need to hurry your decision.

(c) the outcomes described for the alternative drug therapies are those which would be expected for the next one year of your life

Press any key to begin the interview.
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>some problems walking about</td>
<td>confined to bed</td>
<td>no problems walking about</td>
<td>Immediate death</td>
</tr>
<tr>
<td></td>
<td>no problems with self-care</td>
<td>some problems washing/dressing self</td>
<td>unable to wash/dress self</td>
<td></td>
</tr>
<tr>
<td></td>
<td>some problems performing usual activities</td>
<td>unable to perform usual activities</td>
<td>no problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>moderate pain or discomfort</td>
<td>no pain or discomfort</td>
<td>extreme pain or discomfort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>not anxious or depressed</td>
<td>extremely anxious or depressed</td>
<td>moderately anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>confined to bed</td>
<td>no problems walking about</td>
<td>some problems walking about</td>
<td>Immediate death</td>
</tr>
<tr>
<td></td>
<td>unable to wash/dress self</td>
<td>some problems washing/dressing self</td>
<td>no problems with self-care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no problems performing usual activities</td>
<td>unable to perform usual activities</td>
<td>some problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>moderate pain or discomfort</td>
<td>extreme pain or discomfort</td>
<td>no pain or discomfort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>extremely anxious or depressed</td>
<td>not anxious or depressed</td>
<td>moderately anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>confined to bed</td>
<td>some problems walking about</td>
<td>confined to bed</td>
<td>Immediate death</td>
</tr>
<tr>
<td>no problems with self-care</td>
<td>no problems with self-care</td>
<td>unable to wash/dress self</td>
<td></td>
</tr>
<tr>
<td>no problems performing usual activities</td>
<td>some problems performing usual activities</td>
<td>unable to perform usual activities</td>
<td></td>
</tr>
<tr>
<td>no pain or discomfort</td>
<td>extreme pain or discomfort</td>
<td>moderate pain or discomfort</td>
<td></td>
</tr>
<tr>
<td>not anxious or depressed</td>
<td>moderately anxious or depressed</td>
<td>not anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>some problems</td>
<td>no problems</td>
<td>no problems</td>
<td>Immediate death</td>
</tr>
<tr>
<td></td>
<td>walking about</td>
<td>walking about</td>
<td>walking about</td>
<td>death</td>
</tr>
<tr>
<td></td>
<td>unable to wash/dress self</td>
<td>some problems washing/dressing self</td>
<td>some problems washing/dressing self</td>
<td></td>
</tr>
<tr>
<td></td>
<td>unable to perform usual activities</td>
<td>some problems performing usual activities</td>
<td>some problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no pain or discomfort</td>
<td>moderate pain or discomfort</td>
<td>extreme pain or discomfort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>moderately anxious or depressed</td>
<td>extremely anxious or depressed</td>
<td>extremely anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no problems walking about</td>
<td>confined to bed</td>
<td>some problems walking about</td>
<td>Immediate death</td>
</tr>
<tr>
<td></td>
<td>unable to wash/dress self</td>
<td>no problems with self-care</td>
<td>some problems washing/dressing self</td>
<td></td>
</tr>
<tr>
<td></td>
<td>some problems performing usual activities</td>
<td>unable to perform usual activities</td>
<td>no problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>no pain or discomfort</td>
<td>extreme pain or discomfort</td>
<td>moderate pain or discomfort</td>
<td></td>
</tr>
<tr>
<td></td>
<td>not anxious or depressed</td>
<td>extremely anxious or depressed</td>
<td>moderately anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>confined to bed</td>
<td>no problems walking about</td>
<td>some problems walking about</td>
<td>Immediate death</td>
</tr>
<tr>
<td>some problems washing/dressing self</td>
<td>no problems with self-care</td>
<td>unable to wash/dress self</td>
<td></td>
</tr>
<tr>
<td>some problems performing usual activities</td>
<td>unable to perform usual activities</td>
<td>no problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>moderate pain or discomfort</td>
<td>no pain or discomfort</td>
<td></td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td>not anxious or depressed</td>
<td>extremely anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>no problems walking about</td>
<td>some problems walking about</td>
<td>confined to bed</td>
<td>Immediate death</td>
</tr>
<tr>
<td>no problems with self-care</td>
<td>some problems washing/dressing self</td>
<td>unable to wash/dress self</td>
<td></td>
</tr>
<tr>
<td>unable to perform usual activities</td>
<td>no problems performing usual activities</td>
<td>some problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td>no pain or discomfort</td>
<td>extreme pain or discomfort</td>
<td>moderate pain or discomfort</td>
<td></td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td>not anxious or depressed</td>
<td>extremely anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>no problems walking about</td>
<td>confined to bed</td>
<td>some problems washing/dressing self</td>
<td>unable to wash/dress self</td>
</tr>
<tr>
<td></td>
<td>some problems washing/dressing self</td>
<td>unable to perform usual activities</td>
<td>unable to perform usual activities</td>
<td>no problems performing usual activities</td>
</tr>
<tr>
<td></td>
<td>unable to perform usual activities</td>
<td>no pain or discomfort</td>
<td>no pain or discomfort</td>
<td>moderate pain or discomfort</td>
</tr>
<tr>
<td></td>
<td>no pain or discomfort</td>
<td>not anxious or depressed</td>
<td>not anxious or depressed</td>
<td>extremely anxious or depressed</td>
</tr>
<tr>
<td></td>
<td>moderately anxious or depressed</td>
<td></td>
<td></td>
<td>Immediate death</td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>some problems walking about</td>
<td>confined to bed</td>
<td>no problems walking about</td>
<td>Immediate death</td>
</tr>
<tr>
<td>unable to wash/dress self</td>
<td>no problems with self-care</td>
<td>no problems with self-care</td>
<td></td>
</tr>
<tr>
<td>unable to perform usual activities</td>
<td>some problems performing usual activities</td>
<td>no problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>extreme pain or discomfort</td>
<td>moderate pain or discomfort</td>
<td></td>
</tr>
<tr>
<td>extremely anxious or depressed</td>
<td>not anxious or depressed</td>
<td>moderately anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>confined to bed</td>
<td>no problems walking about</td>
<td>some problems walking about</td>
<td>Immediate death</td>
</tr>
<tr>
<td>unable to wash/dress self</td>
<td>some problems washing/dressing self</td>
<td>no problems with self-care</td>
<td></td>
</tr>
<tr>
<td>no problems performing usual activities</td>
<td>some problems performing usual activities</td>
<td>unable to perform usual activities</td>
<td></td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>no pain or discomfort</td>
<td>moderate pain or discomfort</td>
<td></td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td>extremely anxious or depressed</td>
<td>not anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>no problems walking about</td>
<td>some problems walking about</td>
<td>confined to bed</td>
<td>Immediate death</td>
</tr>
<tr>
<td>unable to wash/dress self</td>
<td>some problems washing/dressing self</td>
<td>no problems with self-care</td>
<td></td>
</tr>
<tr>
<td>some problems performing usual activities</td>
<td>some problems performing usual activities</td>
<td>no problems performing usual activities</td>
<td></td>
</tr>
<tr>
<td>extreme pain or discomfort</td>
<td>moderate pain or discomfort</td>
<td>no pain or discomfort</td>
<td></td>
</tr>
<tr>
<td>not anxious or depressed</td>
<td>moderately anxious or depressed</td>
<td>extremely anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
WHICH DRUG THERAPY WOULD YOU CHOOSE?
Type the number in that box

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>confined to bed</td>
<td>no problems walking about</td>
<td>some problems walking about</td>
<td>Immediate death</td>
</tr>
<tr>
<td>no problems with self-care</td>
<td>unable to wash/dress self</td>
<td>some problems washing/dressing self</td>
<td></td>
</tr>
<tr>
<td>some problems performing usual activities</td>
<td>no problems performing usual activities</td>
<td>unable to perform usual activities</td>
<td></td>
</tr>
<tr>
<td>moderate pain or discomfort</td>
<td>not anxious or depressed</td>
<td>extreme pain or discomfort</td>
<td></td>
</tr>
<tr>
<td>moderately anxious or depressed</td>
<td></td>
<td>extremely anxious or depressed</td>
<td></td>
</tr>
</tbody>
</table>

Press a key between 1 and 4 to indicate your choice
Thank you very much for your cooperation.

You may complete this part of the interview by pressing any key.

Then please wait a moment while your answers are saved on the diskette before removing it from the drive. Thank you.

Press any key to finish this part of the interview.
APPENDIX G

RESPONDENT EVALUATION FORMS FOR EACH INDIVIDUAL PREFERENCE MEASUREMENT METHODOLOGY
MEASURING PREFERENCES FOR HEALTH OUTCOMES

Evaluation of Feeling Thermometer Tasks

Start Time

End Time
RESPONDENT NUMBER

MEASURING PREFERENCES FOR HEALTH OUTCOMES

Evaluation of Feeling Thermometer Tasks

Start Time

End Time
In this short questionnaire, we would like to obtain feedback from you regarding your evaluation of the feeling thermometer tasks you performed as a method for understanding your preferences for drug therapy outcomes.

INSTRUCTIONS: You are provided with eight descriptive scales pertaining to the nature of the feeling thermometer tasks you just completed as a part of this study. Each scale is based on two opposite concepts, such as “Difficult to do” and “Easy to do.” Please place a check (✓) within each scale to indicate how you feel the concepts relate to the feeling thermometer tasks under evaluation.

For example, if the concepts given are “Difficult to do” and “Easy to do” and you feel that the feeling thermometer tasks were extremely “Difficult to do,” please place your check (✓) as follows:

Feeling Thermometer Tasks

extremely moderately slightly neutral slightly moderately extremely

Difficult to do ✓ : ___ : ___ : ___ : ___ : ___ : ___ : ___ Easy to do

If you feel that the feeling thermometer tasks were moderately “Easy to do,” please place your check (✓) as follows:

Feeling Thermometer Tasks

extremely moderately slightly neutral slightly moderately extremely

Difficult to do ___ : ___ : ___ : ___ : ___ : ___ : ___ : ✓ Easy to do
If you feel that the feeling thermometer tasks were slightly "Difficult to do," please place your check (✓) as follows:

Feeling Thermometer Tasks

extremely moderately slightly neutral slightly moderately extremely

Difficult to do ___ : ___ : ✓ : ___ : ___ : ___ Easy to do

If you feel that both concepts are equally associated with the feeling thermometer tasks or if you consider the concepts to be neutral, then please place your check (✓) in the center, as follows:

Feeling Thermometer Tasks

extremely moderately slightly neutral slightly moderately extremely

Difficult to do ___ : ___ : ___ : ✓ : ___ : ___ Easy to do

If you feel that both concepts seem completely unrelated to the feeling thermometer tasks, please bring this to the attention of the interviewer.

Finally, please be sure to place a check (✓) on each scale. Also, please remember to mark in the middle of the space, as follows:

LIKE THIS

✓ : ___

NOT LIKE THIS

✓ : ___
Feeling Thermometer Tasks

exremely slightly neutral slightly moderately extremely

Boring _____:_____:_____:_____:_____:_____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:____:_____
RESPONDENT NUMBER

MEASURING PREFERENCES FOR HEALTH OUTCOMES

Evaluation of Chance Tasks

Start Time

End Time
In this short questionnaire, we would like to obtain feedback from you regarding your evaluation of the chance tasks you performed as a method for understanding your preferences for drug therapy outcomes.

INSTRUCTIONS: You are provided with eight descriptive scales pertaining to the nature of the chance tasks you just completed as a part of this study. Each scale is based on two opposite concepts, such as "Difficult to do" and "Easy to do." Please place a check (✓) within each scale to indicate how you feel the concepts relate to the chance tasks under evaluation.

For example, if the concepts given are "Difficult to do" and "Easy to do" and you feel that the chance tasks were extremely "Difficult to do," please place your check (✓) as follows:

Chance Tasks

extremely moderately slightly neutral slightly moderately extremely

Difficult to do ✓:   :   :   :   :   :    Easy to do

If you feel that the chance tasks were moderately "Easy to do," please place your check (✓) as follows:

Chance Tasks

extremely moderately slightly neutral slightly moderately extremely

Difficult to do ___:   :   :   :   :  :    Easy to do
If you feel that the chance tasks were slightly “Difficult to do,” please place your check (√) as follows:

<table>
<thead>
<tr>
<th>Chance Tasks</th>
<th>extremely</th>
<th>moderately</th>
<th>slightly</th>
<th>neutral</th>
<th>slightly</th>
<th>moderately</th>
<th>extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult to do</td>
<td>___ : ___ : √ : ___ : ___ : ___ : ___</td>
<td>Easy to do</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If you feel that both concepts are equally associated with the chance tasks or if you consider the concepts to be neutral, then please place your check (√) in the center, as follows:

<table>
<thead>
<tr>
<th>Chance Tasks</th>
<th>extremely</th>
<th>moderately</th>
<th>slightly</th>
<th>neutral</th>
<th>slightly</th>
<th>moderately</th>
<th>extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult to do</td>
<td>___ : ___ : ___ : ___ : ___ : ___ : ___</td>
<td>Easy to do</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If you feel that both concepts seem completely unrelated to the chance tasks, please bring this to the attention of the interviewer.

Finally, please be sure to place a check (√) on each scale. Also, please remember to mark in the middle of the space, as follows:

LIKE THIS

: ___ : √ : ___

NOT LIKE THIS

: ___ : √ : ___
Chance Tasks

Boring: Extremely moderately slightly neutral slightly moderately extremely

Sensible: Extremely moderately slightly neutral slightly moderately extremely

Difficult to do: Extremely moderately slightly neutral slightly moderately extremely

Useful: Extremely moderately slightly neutral slightly moderately extremely

Long: Extremely moderately slightly neutral slightly moderately extremely

Successful: Extremely moderately slightly neutral slightly moderately extremely

Clear: Extremely moderately slightly neutral slightly moderately extremely

Difficult to understand: Extremely moderately slightly neutral slightly moderately extremely

Interesting: Extremely moderately slightly neutral slightly moderately extremely

Senseless: Extremely moderately slightly neutral slightly moderately extremely

Easy to do: Extremely moderately slightly neutral slightly moderately extremely

Useless: Extremely moderately slightly neutral slightly moderately extremely

Short: Extremely moderately slightly neutral slightly moderately extremely

Unsuccessful: Extremely moderately slightly neutral slightly moderately extremely

Vague: Extremely moderately slightly neutral slightly moderately extremely

Easy to understand: Extremely moderately slightly neutral slightly moderately extremely

How confident are you that the chance tasks did a good job of measuring your preferences for drug therapy outcomes? Please circle only one option.

1. VERY UNCONFIDENT
2. SOMewhat UNCONFIDENT
3. NEITHER CONFIDENT NOR UNCONFIDENT
4. SOMewhat CONFIDENT
5. VERY CONFIDENT

Thank you very much for your cooperation!
RESPONDENT NUMBER

MEASURING PREFERENCES FOR HEALTH OUTCOMES

Evaluation of Choice Tasks

Start Time
End Time
In this short questionnaire, we would like to obtain feedback from you regarding your evaluation of the choice tasks you performed as a method for understanding your preferences for drug therapy outcomes.

INSTRUCTIONS: You are provided with eight descriptive scales pertaining to the nature of the choice tasks you just completed as a part of this study. Each scale is based on two opposite concepts, such as "Difficult to do" and "Easy to do." Please place a check (✓) within each scale to indicate how you feel the concepts relate to the choice tasks under evaluation.

For example, if the concepts given are "Difficult to do" and "Easy to do" and you feel that the choice tasks were extremely "Difficult to do," please place your check (✓) as follows:

Choice Tasks

extremely moderately slightly neutral slightly moderately extremely

Difficult to do  ✓: ___: ___: ___: ___: ___: ___: ___ Easy to do

If you feel that the choice tasks were moderately "Easy to do," please place your check (✓) as follows:

Choice Tasks

extremely moderately slightly neutral slightly moderately extremely

Difficult to do  ___: ___: ___: ___: ___: ✓: ___ Easy to do
If you feel that the choice tasks were slightly "Difficult to do," please place your check (✓) as follows:

**Choice Tasks**

<table>
<thead>
<tr>
<th>extremely</th>
<th>moderately</th>
<th>slightly</th>
<th>neutral</th>
<th>slightly</th>
<th>moderately</th>
<th>extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult to do</td>
<td>___</td>
<td>___</td>
<td>✓</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
</tbody>
</table>

If you feel that both concepts are equally associated with the choice tasks or if you consider the concepts to be neutral, then please place your check (✓) in the center, as follows:

**Choice Tasks**

<table>
<thead>
<tr>
<th>extremely</th>
<th>moderately</th>
<th>slightly</th>
<th>neutral</th>
<th>slightly</th>
<th>moderately</th>
<th>extremely</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult to do</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>✓</td>
<td>___</td>
<td>___</td>
</tr>
</tbody>
</table>

If you feel that both concepts seem completely unrelated to the choice tasks, please bring this to the attention of the interviewer.

Finally, please be sure to place a check (✓) on each scale. Also, please remember to mark in the middle of the space, as follows:

LIKE THIS: ___ ✓ ___

NOT LIKE THIS: ___ ✓ ___
Choice Tasks

How confident are you that the choice tasks did a good job of measuring your preferences for drug therapy outcomes? Please circle only one option.

1. VERY UNCONFIDENT
2. SOMewhat UNCONFIDENT
3. NEITHER CONFIDENT NOR UNCONFIDENT
4. SOMewhat CONFIDENT
5. VERY CONFIDENT

Thank you very much for your cooperation!
APPENDIX H

RESPONDENT EVALUATION FORMS FOR OVERALL COMPARATIVE EVALUATION AND OTHER FEEDBACK
RESPONDENT NUMBER

MEASURING PREFERENCES FOR HEALTH OUTCOMES

Overall Evaluation

Start Time

End Time
Thank you very much for your time. Finally, I'd like to get some feedback from you about some aspects of this study. Please note that your feedback is very important to us and will be used in future research. Please answer all questions as truthfully as possible and do not feel hesitant about making any negative comments about any aspect of the study.

PART I: Comparison of the Three Preference Measurement Tasks

In this interview, your preferences for the health outcomes of drug therapy were measured in three different ways -- using feeling thermometer tasks, choice tasks, and chance tasks. In this section of the feedback questionnaire, we would like you to compare these three preference measurement tasks.

INSTRUCTIONS: I'm going to be calling out eight descriptive phrases which may be used to compare the three preference measurement tasks. For each phrase, please let me know which task you think is most, middle, and least described by the phrase, as shown in the example below:

<table>
<thead>
<tr>
<th>Thoughtful</th>
<th>Most</th>
<th>Middle</th>
<th>Least</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>P</td>
<td>F</td>
<td></td>
</tr>
</tbody>
</table>

This individual has indicated that the choice task (C) is the most thoughtful, followed by the chance task (P). According to this individual, the feeling thermometer task (F) is the least thoughtful.

In order to make the comparison faster, you may use the following short-forms:

- F = feeling thermometer task
- C = choice task
- P = chance task

Please turn the page if you are ready to begin.
Please rank the three tasks in terms of how confident you are that they did a good job of measuring your preferences for drug therapy outcomes:

<table>
<thead>
<tr>
<th></th>
<th>Most</th>
<th>Middle</th>
<th>Least</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interesting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senseless</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult to do</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Useful</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Successful</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difficult to understand</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Considering all of the above factors, what is your overall ranking of the three tasks?

OVERALL BEST:  

MIDDLE:  

WORST:  
PART II: Other Feedback

Finally, we would like to get your feedback about some other general aspects of the interview.

1. The descriptions of the outcomes of drug therapy were presented in terms of five different aspects of health, i.e., mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Do you feel that there are any other aspects of health that we need to consider?

1. NO
2. YES (PLEASE SPECIFY)

2. Do you have any other comments about any aspect of this study? If so, please write them down in the space provided below.

Thank you very much for your cooperation!
APPENDIX I

APPLICATION FOR EXEMPTION FROM FULL REVIEW SUBMITTED TO THE BEHAVIORAL AND SOCIAL SCIENCES REVIEW COMMITTEE
APPLICATION FOR EXEMPTION FROM HUMAN SUBJECTS COMMITTEE REVIEW

RETURN TWO (2) COPIES OF THE TYPED APPLICATION (including original signatures) TO: Office of Research Risks, Room 300, Research Foundation Building, 1960 Kenny Road, Campus. (ATTACH A BRIEF ABSTRACT DESCRIBING THE RESEARCH ACTIVITY IN LAY TERMS. ALSO, ANY QUESTIONNAIRES OR SURVEY INSTRUMENTS.)

Principal Investigator: Dev S. Pathak, D.B.A.
(Must be OSU Faculty) (Typed Name) (Signature)
Academic Title: Merrell-Dow Professor
Department: Pharmacy Practice & Administration
Campus Address: 136A - Parks Hall
Room Number - Building
Co-Investigator(s): Zafar Habim
(Signed)
Suzan Kucukarslan
(Typed Name) (Signature)
Jon C. Schommer

Yes No (Please respond to each item - A through F)

A. The ONLY involvement of human subjects in the proposed research activity will be in one or more of the exemption categories as described in the appendix of "Human Subjects Program Guidelines."

* Category(ies) # 2 ___.

B. The proposed research activity will involve minors (under the age of 18.)

C. The proposed research activity will involve pregnant women, mentally retarded, mentally disabled, and/or prisoners.

D. The proposed research activity will involve human in vitro fertilization.

E. The proposed research activity will involve an element of deception.

F. The proposed research activity will expose subjects to discomfort or harassment beyond levels encountered in daily life.

Source of Funding for Proposed Research: (Check A or B.)
A. OSURF: Sponsor __________________ RF Proposal/Project No. ____________
B. Other (Identify) Department ____________________________

Office Use: EXEMPTION STATUS: OCT 27 1994
Date ____________ Chairperson ____________

** Principal Investigator must submit a protocol to the appropriate Human Subjects Review Committee
APPENDIX J

SUMMARY STATEMENTS AND PROPOSAL FOR FULL REVIEW
SUBMITTED TO THE BEHAVIORAL AND SOCIAL SCIENCES
REVIEW COMMITTEE
Protocols received in the Office of Research Risks after the deadline date (NOON, Friday preceding meeting date) will be scheduled for the following meeting. If all time slots are filled and a protocol is received on or before the deadline date, the protocol will be scheduled for the following meeting. Only protocols that are complete will be scheduled for review. Incomplete protocols will be returned.

Principal Investigator(s): Dev S. Pathak, D.B.A.  
(Zafar Hakim, M.S.)

Typed Name  
Typed Name  
Typed Name

Academic Title: Merrell-Dow Professor

College: Pharmacy  
Department Pharmacy Practice & Administration

Campus Address: 136A Parks Hall, 500 West 12th Avenue, Columbus, OH 43210  
(Faculty Member's Campus Address)

PROTOCOL TITLE: (Include proposal title in parentheses if different from the protocol.)


SOURCE OF FUNDING FOR PROPOSED RESEARCH: (Check A or B):
A. OSURF: Sponsor __________________________ RF Proposal/Project No. ________
X. Other (Identify) Department Funds

Information about the funding/sponsorship of human subjects research activities is required for administrative purposes. Such information is generally not required as part of the human subjects review process.
The following summary must accompany your proposal. Be specific about exactly what subjects will experience when they participate in your research, and about the protections that have been included to safeguard them. Careful attention to the following may help facilitate the review process.

1. In a sentence or two, describe the background and purpose of the research.

Currently used methods to measure patients' preferences for being in particular states of health suffer from several drawbacks which may limit their usefulness. The purpose of this study is to use a new method - Choice-based Conjoint Analysis - to measure patients' preferences for health states and compare the results obtained with those of commonly used health status preference measurement methods, i.e., rating scale and standard gamble methods.

2. Briefly describe each condition or manipulation to be included within the study.

Patients' preferences for health states will be measured in three different ways:
(a) rating scale — rate imaginary drug therapies on a 0-100 scale
(b) choice-based conjoint — select one out of four imaginary drug therapy alternatives as the most desirable
(c) standard gamble — choosing between two imaginary drug therapies, one having known effects on health and the other having uncertain effects on health

3. What measures or observations will be taken in the study? If any questionnaires, check, or other instruments are used, provide a brief description and either include a copy or indicate when a copy will be submitted for review.
(a) background information on patients — demographics & disease-related information
(b) health status preferences — three different ways (as explained in 2 above)
(c) feedback — about preference measurement tasks and interview in general

STUDY QUESTIONNAIRES ARE ATTACHED

4. Will the subjects encounter the possibility of psychological, social, physical or legal risk? Yes ____ No X If so, please describe.

N/A

5. Will any stress to subjects be involved? Yes ____ No X If so, describe.

6. Will the subjects be deceived or misled in any way? Yes ____ No X If so, please describe and include an outline or script of the debriefing.

HS-008C (Rev. 6/92)
BEHAVIORAL & SOCIAL SCIENCES
SUMMARY SHEETS

7. Will there be a request for information which subjects might consider to be personal sensitive? Yes X No ____ If so, please describe.

As part of the background information, patients will be asked to state the highest education level they have attained and what other health problems they have had in the past five years. The reason for including these variables in the study is that the literature supports the hypothesis that patients' preferences for health states may be affected by these variables. Moreover, education level is a commonly asked question in survey research and has not been reported to be overly sensitive in past research. Finally, patients routinely answer questions about their other health problems at the clinic (see 10 below). If subjects do not wish to answer any question, they will have complete freedom to do so without any prejudice to them. (Check one)

8. Will the subjects be presented with materials which might consider to be offensive, threatening or degrading. Yes ____ No X If so, please describe.

9. Approximately how much time will be demanded of each subject?

One and a half hours.

10. Who will be the subjects in this study? How will they be solicited or contacted? Subjects will consist of hypertensive patients treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic in Columbus, Ohio. A letter from the clinic to this effect is attached. Sampled patients will be contacted over the telephone and requested to participate in the study.

TEXT FOR RECRUITING PATIENTS OVER THE TELEPHONE IS ATTACHED

11. What steps will be taken to insure that each subject's participation is voluntary? What, if any, inducements will be offered to the subjects for their participation?

Subjects will be informed that their decision to participate in the study will have no bearing on the quality of the care they receive at the VA clinic. If subjects decline to participate, no follow-up questions will be asked regarding the reason for refusal. Finally, no inducements will be offered to the subjects for their participation.

HS-008D (Rev. 6/92)
12. How will you insure that the subjects give their consent prior to participation? A written consent form will be used. Form is attached.

13. Will any aspect of the data be made a part of any permanent record that can be identified with the subject? Yes X No

14. Will the fact that a subject did or did not participate in a specific experiment or study be made a part of any permanent record available to a supervisor, teacher or employer? Yes X No

   A copy of the written consent form will be placed in the patient's medical record. This record is available only to clinic health-care professionals.

15. What steps will be taken to insure the confidentiality of the data?

   All data collection procedures will be conducted by a single investigator, including telephone calls and the personal interviews. Respondents will only be represented by respondent numbers on the data collection forms. No names will be recorded at any point during the interview.

16. If there are any risks involved in the study, are there any offsetting benefits that might accrue to either the subject or society?

   No risks are involved in the study.

17. Will any data from files or archival data be used? Yes X No

   The only data which will be used from the patients' medical records are their blood pressure readings. This is needed to test the hypothesis that patients differ in their health status preferences according to the severity of their condition. Further, patients' telephone numbers will be taken from their administrative records at the VA clinic. This is necessary to contact them and schedule appointments.
HEALTH STATUS PREFERENCE MEASUREMENT:
A COMPARISON OF THREE TECHNIQUES

A Proposal Submitted to the Office of Research Risks,
The Ohio State University, Columbus, Ohio

by

Zafar Hakim, M.S.
Ph.D. Candidate
(614)-292-1716

Dev S. Pathak, D.B.A.
Merrell-Dow Professor
(614)-292-6415

Suzan Kucukarslan, Ph.D.
Assistant Professor
(614)-292-1363

Jon C. Schommer, Ph.D.
Assistant Professor
(614)-292-3011

Division of Pharmacy Practice and Administration
College of Pharmacy
The Ohio State University
500 West 12th Avenue
Columbus, OH 43210.
Tel: (614)-292-2775
FAX: (614)-292-1335

and

William Arneson, M.S.
Chief, Pharmacy Service
Department of Veterans Affairs Outpatient Clinic
2090 Kenny Road
Columbus, OH 43221.
Tel: (614)-469-2547
FAX: (614)-469-5113

October 28, 1994
EXECUTIVE SUMMARY

Quality of life and health status are now considered important outcomes in evaluating drug therapy. Experts agree that health status is a multidimensional construct and should be measured with multiple indicators. However, in interpreting the results of research which have multiple indicators of the same construct, it is important to make decisions about which of the indicators are important than others. It is the premise of this study that the weighting of such indicators must be based on patient preferences for the effects measured by the indicators. In other words, what is needed is a method to quantify how much more desirable a given effect of a drug on health status (as represented by one indicator) is as compared to another. However, the most commonly used health status preference measurement methods suffer from various drawbacks, including the inability to consider all possible interactions between health status dimensions and using patients' judgments instead of choices as the response measure. Choice-based Conjoint Analysis is a promising new preference measurement method which is capable of handling interactions between health status dimensions and is based upon choice data.

The purpose of this study is to measure patients' health status preferences using Choice-based Conjoint Analysis and compare the results so obtained with those of the commonly used health status preference measurement methods, i.e., the rating scale and standard gamble methods. The study will also describe and differentiate between patients' preferences for health status on the basis of demographic and disease-related characteristics.

The conceptualization and operationalization of health status adopted for this study is that of the EuroQol. The study sample will be drawn from a sampling frame consisting of all hypertensive patients who are taking at least one antihypertensive drug and are currently being treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic in Columbus, Ohio. The estimated sample size is 130 patients. Data will be collected by personal interviews. The health status preferences of each patient will be measured by three different preference measurement methods. Data will be analyzed using multi-attribute preference theory, the CBC System for Choice-based Conjoint Analysis, and SAS.

By measuring and quantifying patients' preferences for health status, this study will provide clinicians with valuable information which they can use as inputs in therapeutic decision making and counseling for more
compliant patient behavior. Further, by introducing and validating a new health status preference measurement method and exploring the relationship between this new method and existing methods, the study will advance the field of health status preference measurement.

INTRODUCTION

Quality of life and health status are now considered important outcomes in evaluating drug therapy. The current interest in measuring health status reflects the coalescence of several related trends. Primary among these is the gradual change in the focus of medical care from decreasing mortality to limiting morbidity and the patient-reported impact of that morbidity. This change has been brought about mainly due to the nature of medical conditions observed in society, which have shifted from a predominance of acute and infectious diseases to a predominance of chronic conditions. Chronic conditions, for the most part, are incurable. The emphasis of therapy in such conditions, then, is usually a reduction in severity rather than cure or prolongation of life.

Further, there are behavioral consequences that come into play in such cases. Since chronic illnesses usually require lifelong therapy, the role of the patient in following the therapy becomes crucial in determining the effectiveness of therapy. If patients cannot see or feel a noticeable improvement in their functioning or well-being, they may be inclined to discontinue or be noncompliant with therapy. This in turn would have deleterious consequences for the health of the patient. In other words, patients need a measure of health which is centered around how they feel and function, since they may be unlikely to accept any therapy -- whatever its clinical or scientific merit -- if they see nothing in it for themselves.

Health-care providers are increasingly becoming convinced that one of the primary objectives of medical care is the enhancement of health status and quality of life of the patient. Indeed, the measurement of the patient's quality of life forms the centerpiece of the "outcomes management" movement, which is arguably a driving force behind medical care in the 1990s.

Policy makers are also driving the interest in health status and quality of life measurement from a regulatory perspective. This is because such information can be used in pharmacoeconomic evaluations, which compare
the costs and outcomes of alternative therapies. Regulatory authorities in Australia and Canada have begun requesting such information from pharmaceutical manufacturers along with safety and efficacy data for drug approval and registration decisions in these countries. In the US, the Food and Drug Administration Oncologic Drugs Advisory Committee has recommended that health-related quality of life data be considered along with survival data in the approval decision of anticancer drugs. Pharmaceutical manufacturers have now established entire departments whose sole function is the collection and analysis of quality of life data.

**NATURE OF THE PROBLEM**

In response to these demands, health-care researchers have developed several questionnaires which measure health status. Although there is no consensus on what specific dimensions of health should be included in a measure of health status, most commonly-used questionnaires conceptualize health status as comprising three major dimensions -- physical functioning, emotional functioning, and social functioning. Within each of these broad dimensions, there are several sub-categories, e.g., physical functioning may be viewed as consisting of mobility (i.e., the ability to walk about) and self-care (i.e., the ability to take care of oneself). However, although the specific dimensions to be included in a measure of health status may vary according to the purpose of the measure, there does seem to be a consensus among experts in the field that health status is a multidimensional construct and should be measured with multiple indicators.

In interpreting the results of research which have multiple indicators of the same construct, it is important to make decisions about which of the indicators are more important than others. When the evidence from one indicator conflicts with that from another, there must be some way of deciding which indicator carries the most weight. This is especially true for health status measurement questionnaires, which are usually multidimensional. For instance, a drug therapy may have beneficial effects on physical functioning but adverse effects on emotional functioning for the same patient. Most current health status measurement instruments can measure these beneficial and adverse effects on health status. However, they offer little guidance on how these conflicting effects have to be weighted in an overall drug therapy decision for the patient. It is the premise of this study that the weighting of such effects must be based on patient preferences for the effects. In other words, what is needed is a method to quantify how
much more desirable a given effect of a drug on health status is as compared to another.

The most commonly used health status preference measurement methods are the rating scale and standard gamble methods. The rating scale method requires patients to rate different hypothetical states of health on a 0-100 point scale, where 0 = least desirable and 100 = most desirable. The standard gamble method requires patients to evaluate different states of health by means of a gamble between a hypothetical health state that is certain versus a wager between two other hypothetical health states. Although more complex than the simple rating scale method, this latter method explicitly incorporates the notion of the probabilistic nature of the effects of drug therapy on health status.

However, the drawbacks of these methods include the inability to consider all possible interactions between health status dimensions and the fact that they use patients' judgments instead of choices as the response measure. The task of measuring patient preferences for health status is complicated by the fact that health status dimensions may interact with one another -- i.e., patients' preferences for physical functioning may be dependent on the particular level of emotional functioning they experience. Therefore, if any method does not consider these interactions, it may not truly measure patients' preferences for health status.

In order to understand the judgment versus choice distinction, it is necessary to think of the reason why patient preferences for health status need to be measured -- for the purpose of making a decision between several alternative therapies as to which one to prescribe for a patient. Making such a decision involves making a choice of one alternative over the others. Judgmental data (as obtained by the commonly used preference measurement methods) do not have a theoretical rationale linking them to choices -- choice data, on the other hand, have such a rationale. Choice-based Conjoint Analysis is a promising new preference measurement method which is capable of handling interactions between health status dimensions and is based upon choice data. Hence, this new method seems to offer advantages over the currently used health status preference measurement methods.
OBJECTIVES

The purpose of this study is to measure patients' health status preferences using Choice-based Conjoint Analysis and compare the results so obtained with those of the commonly used health status preference measurement methods, i.e., the rating scale and standard gamble methods. The study will also describe and differentiate between patients' preferences for health status on the basis of demographic and disease-related characteristics. The specific research questions and tasks required to answer these questions are listed in Appendix 1.

METHODOLOGY

The conceptualization and operationalization of health status adopted for this study is that of the EuroQol™ (see Appendix 2). The EuroQol™ is a health status measurement questionnaire developed by an interdisciplinary group of researchers from five countries in Europe.1 It conceptualizes health status as consisting of five different dimensions -- mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each of these five dimensions is further sub-divided into three levels of functioning. Therefore, a patient's state of health at a given point in time is viewed as a combination of one level from each of the five dimensions of the EuroQol™. Published applications of the EuroQol™ have not extended outside Europe, although this measure has universal applicability. Hence, this study will be the first to apply the EuroQol™ in the US. Correspondence regarding permission to use the EuroQol™ for this study is presented in Appendix 3.

The target population for this study is hypertensive patients taking one or more antihypertensive drugs. It was decided to use hypertensive patients because previous studies of antihypertensive drugs on health status have shown that these drugs have differential effects on health status dimensions17,18 and, given the chronic nature of hypertension, health status is considered an important outcome of antihypertensive drug therapy.19 The study sample will be drawn from a sampling frame consisting of all hypertensive patients who are taking at least one antihypertensive drug and are currently being treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic (or VAOPC) in Columbus, Ohio. The estimated sample size is 130 patients. Assuming a 85% participation rate (which has been reported in the literature for studies on health status preference measurement),20 153 patients will be approached to participate in
the study. Correspondence regarding permission to conduct the study at the VAOPC in Columbus, Ohio (pending approval of The Ohio State University Office of Research Risks) is presented in Appendix 4.

Data collection will proceed through the following stages (see Appendix 5). Sampled patients will be contacted over the telephone and requested to participate in the study. Patients who agree to participate will be sent a letter confirming their appointment time and date. Each participant will be interviewed by a researcher (i.e., Zafar Hakim) on the premises of the Clinic. The interview will last approximately one and a half hours. The interview will begin by obtaining informed consent from patients. Next, background demographic and disease-related information will be obtained, after which patients will be asked to evaluate their current health status using the EuroQol. This will be followed by the three preference measurement methods. The order of the three methods will be varied at random in order to avoid any learning effects. After each of the three methods is completed, feedback will be obtained about each specific method. Finally, at the end of the interview, an overall feedback questionnaire will be administered. One day after the interview, a personalized thank-you letter will be mailed to all study participants. All data collection letters and questionnaires are provided in Appendix 6.

The data will be analyzed at The Ohio State University College of Pharmacy. Preference measurement analysis will be conducted using multi-attribute preference theory (for rating scale and standard gamble methods) and the CBC System (for Choice-based Conjoint Analysis). Sub-group analysis will be conducted using SAS. After completion of the analysis, the Clinic will be provided with a final study report.

SIGNIFICANCE

Chronic conditions predominate in medical practice today. For such conditions, the role of the patient in following the therapeutic regimen becomes crucial in determining the effectiveness of therapy. However, as Goldsmith has noted, patients today are "increasingly vigilant and demanding ... increasingly suspicious of medical direction ... increasingly specific about their needs and how they want them met ...." In their study of hypertensive patients, Strull et al. found that patients desired information about their condition and wanted to discuss alternative therapies with their physicians, but did not desire to be very involved in actual clinical decision
Given these findings, it appears that although patients want to let physicians know their preferences for the outcomes of therapy and discuss therapeutic alternatives with their physicians, they would prefer to let physicians make the final therapeutic decision on their behalf after incorporating their preferences for the outcomes of therapy. This then would lead to greater acceptance of the drug therapy decisions by patients by what Mahler and Kulik have termed "sense of control," which in turn is associated with more compliant behavior by patients. As mentioned earlier, patient compliance is crucial in determining the effectiveness of therapy. Therefore, by measuring and quantifying patients' preferences for health status, this study will provide clinicians with valuable information which they can use as inputs in therapeutic decision making and counseling for more compliant patient behavior.

On a more methodological front, this study will advance the field of health status preference measurement by introducing and validating a new method for measuring health status preferences. By illustrating and empirically demonstrating the relationship between this new method and existing methods, greater understanding of the strengths and weaknesses of alternative methods may be gained. This in turn will prove valuable for use in future research on health status preference measurement methodology.
REFERENCES


APPENDIX K

RESPONSE FROM THE BEHAVIORAL AND SOCIAL SCIENCES REVIEW COMMITTEE
BEHAVIORAL AND SOCIAL SCIENCES
HUMAN SUBJECTS REVIEW COMMITTEE (HSRC)
THE OHIO STATE UNIVERSITY

Date NOVEMBER 4, 1994

RESEARCH PROTOCOL:

94B0328 HEALTH STATUS PREFERENCE MEASUREMENT: A COMPARISON OF THREE TECHNIQUES, Dev S. Pathak, Zafar Hakim, Pharmacy

presented for review by the Behavioral and Social Sciences Review Committee to ensure proper protection of the rights and welfare of the individuals involved with consideration of the methods used to obtain informed consent and the justification of risks in terms of potential benefits to be gained, The Committee action was:

___ APPROVED ___ DEFERRED*

X ___ APPROVED WITH CONDITIONS* ___ DISAPPROVED

___ NO REVIEW NECESSARY

*CONDITIONS/COMMENTS:

Subjects were deemed NOT AT RISK and the protocol was unanimously APPROVED WITH THE FOLLOWING CONDITIONS:

1. Revise the consent form as follows and forward a copy to the Committee:
   a. Include statement indicating that this study is being done in cooperation with The Ohio State University.
   b. Add statement indicating that the consent form will be placed in the patient's medical records.

2. Revise Solicitation letter to include statement indicating that consent form will be placed in the patients medical records.

3. Revise answer to Summary Sheet question #5 to indicate that subjects may experience stress due to the sensitive nature of some of the health questions.

Comment: The Committee makes the following suggestions:

- Omit statement "Thanks to modern science, we know what side effects . . . ." 
- To avoid sounding patronizing, refrain from using the response "Great" during interviews.
If you agree to the above conditions, please sign this form in the space provided below and return with any additional information requested to: BEHAVIORAL HUMAN SUBJECTS REVIEW DESK, 300 RESEARCH FOUNDATION, 1960 KENNY ROAD, CAMPUS, within one week. Upon such compliance, the approval form will be mailed to you. (In case of a deferred protocol, please submit the requested information at your earliest convenience.) The next meeting of the Committee will be two weeks from the meeting date indicated above.

DATE 11/15

Signatures(s)

HS-025A (Rev. 2/91)
(CONDITIONS/COMMENTS)

Principal Investigators
APPENDIX L

REVISED SUMMARY SHEETS SUBMITTED TO THE BEHAVIORAL AND SOCIAL SCIENCES REVIEW COMMITTEE
SUMMARY SHEETS

The following summary must accompany your proposal. Be specific about exactly what subjects will experience when they participate in your research, and about the protections that have been included to safeguard them. Careful attention to the following may help facilitate the review process.

1. In a sentence or two, describe the background and purpose of the research.

Currently used methods to measure patients' preferences for being in particular states of health suffer from several drawbacks which may limit their usefulness. The purpose of this study is to use a new method - Choice-based Conjoint Analysis - to measure patients' preferences for health states and compare the results obtained with those of commonly used health status preference measurement methods, i.e., rating scale and standard gamble methods.

2. Briefly describe each condition or manipulation to be included within the study.

Patients' preferences for health states will be measured in three different ways:
   (a) rating scale — rate imaginary drug therapies on a 0-100 scale
   (b) choice-based conjoint — select one out of four imaginary drug therapy alternatives as the most desirable
   (c) standard gamble — choosing between two imaginary drug therapies, one having known effects on health and the other having uncertain effects on health

3. What measures or observations will be taken in the study? If any background information on patients—demographics & disease-related information other instruments are used, provide a brief description and either include a copy or indicate when a copy will be submitted for review.
   (a) background information on patients — demographics & disease-related information
   (b) health status preferences — three different ways (as explained in 2 above)
   (c) feedback — about preference measurement tasks and interview in general

STUDY QUESTIONNAIRES ARE ATTACHED

4. Will the subjects encounter the possibility of psychological, social, physical or legal risk? Yes ☐ No ☐ If so, please describe.

N/A

5. Will any stress to subjects be involved? Yes ☒ No ☐ If so, describe.

It is possible that some of the descriptions of the imaginary drug therapies may be upsetting to some patients. It will be stressed at the outset as well as during the interview that the drug therapies described are only imaginary and that patients may stop the interview at any point if they feel uncomfortable, without any prejudice to them.

INTERVIEW SCHEDULE IS ATTACHED

6. Will the subjects be deceived or misled in any way? Yes ☐ No ☒ If so, please describe and include an outline or script of the debriefing.
7. Will there be a request for information which subjects might consider to be personal sensitive? Yes [X] No [ ] If so, please describe.

As part of the background information, patients will be asked to state the highest education level they have attained and what other health problems they have had in the past five years. The reason for including these variables in the study is that the literature supports the hypothesis that patients’ preferences for health states may be affected by these variables. Moreover, education level is a commonly asked question in survey research and has not been reported to be overly sensitive in prior research. Finally, patients routinely answer questions about their other health problems at the clinic (see 10 below). If subjects do not wish to answer any question, offensive, threatening or degrading, Yes [X] No [ ] If so, please describe.

8. Will the subjects be presented with activities which might consider they might consider offensive, threatening or degrading. Yes [X] No [ ] If so, please describe.

9. Approximately how much time will be demanded of each subject?

One and a half hours.

10. Who will be the subjects in this study? How will they be solicited or contacted? Subject: must be informed about the nature of what is involved as a participant, including particularly a description of anything they might consider to be unpleasant or a risk. Please provide an outline or script of the information which will be provided to subjects prior to their volunteering to participate. Include a copy of the written solicitation and an outline of the oral solicitation.

Subjects will consist of hypertensive patients treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic in Columbus, Ohio. A letter from the clinic to this effect is attached. A frame of such patients will be obtained from the clinic. Sampled patients will be contacted over the telephone and requested to participate in the study.

TEXT FOR RECRUITING PATIENTS OVER THE TELEPHONE IS ATTACHED

11. What steps will be taken to insure that each subject’s participation is voluntary? What, if any, inducements will be offered to the subjects for their participation?

Subjects will be informed that their decision to participate in the study will have no bearing on the quality of the care they receive at the VA clinic. If subjects decline to participate, no follow-up questions will be asked regarding the reason for refusal. Finally, no inducements will be offered to the subjects for their participation.

HS-008D (Rev. 6/92)
12. How will you insure that the subjects give their consent prior to participating? Will a written consent form be used? Yes _X__ No ____ If so, please include the form. A written consent form will be used. Form is attached.

13. Will any aspect of the data be made a part of any permanent record that can be identified with the subject? Yes ____ No _X__

14. Will the fact that a subject did or did not participate in a specific experiment or study be made a part of any permanent record available to a supervisor, teacher or employer? Yes _X__ No ____

A copy of the written consent form will be placed in the patient's medical record. This record is available only to clinic health-care professionals.

15. What steps will be taken to insure the confidentiality of the data?

All data collection procedures will be conducted by a single investigator, including telephone calls and the personal interviews. Respondents will only be represented by respondent numbers on the data collection forms. No names will be recorded at any point during the interview.

16. If there are any risks involved in the study, are there any offsetting benefits that might accrue to either the subject or society?

No risks are involved in the study.

17. Will any data from files or archival data be used? Yes _X__ No ____

The only data which will be used from the patients' medical records are their blood pressure readings. This is needed to test the hypothesis that patients differ in their health status preferences according to the severity of their condition. Further, patients' telephone numbers will be taken from their administrative records at the VA clinic. This is necessary to contact them and schedule appointments.
APPENDIX M

FINAL APPROVAL FORM OF THE
BEHAVIORAL AND SOCIAL SCIENCES REVIEW COMMITTEE
Research Involving Human Subjects

ACTIONS OF THE REVIEW COMMITTEE

With regard to the employment of human subjects in the proposed research protocol:

94B0328 HEALTH STATUS PREFERENCE MEASUREMENT: A COMPARISON OF THREE TECHNIQUES, Dev S. Pathak, Zafar Hakim, Pharmacy

THE BEHAVIORAL AND SOCIAL SCIENCES REVIEW COMMITTEE HAS TAKEN THE FOLLOWING ACTION:

- APPROVED
- DISAPPROVED
- X APPROVED WITH CONDITIONS*
- WAIVER OF WRITTEN CONSENT GRANTED

* Conditions stated by the Committee have been met by the Investigator and, therefore, the protocol is APPROVED.

It is the responsibility of the principal investigator to retain a copy of each signed consent form for at least three (3) years beyond the termination of the subject's participation in the proposed activity. Should the principal investigator leave the University, signed consent forms are to be transferred to the Human Subjects Review Committee for the required retention period. This application has been approved for the period of one year. You are reminded that you must promptly report any problems to the Review Committee, and that no procedural changes may be made without prior review and approval. You are also reminded that the identity of the research participants must be kept confidential.

Date: NOVEMBER 4, 1994

Signed: Patricia M. Schmier

(Chairperson)
APPENDIX N

PROPOSAL SUBMITTED TO THE VAOPC
RESEARCH AND DEVELOPMENT COMMITTEE
HEALTH STATUS PREFERENCE MEASUREMENT:
A COMPARISON OF THREE TECHNIQUES

A Proposal Submitted to the Research Committee of the
Department of Veterans Affairs Outpatient Clinic, Columbus, Ohio

by

Zafar Hakim, M.S.
Ph.D. Candidate
(614)-292-1716

Suzan Kucukarslan, Ph.D.
Assistant Professor
(614)-292-1363

Dev S. Pathak, D.B.A.
Merrell-Dow Professor
(614)-292-6415

Jon C. Schommer, Ph.D.
Assistant Professor
(614)-292-3011

Division of Pharmacy Practice and Administration
College of Pharmacy
The Ohio State University
500 West 12th Avenue
Columbus, OH 43210.
Tel: (614)-292-2775
FAX: (614)-292-1335

and

William Arneson, M.S.
Chief, Pharmacy Service
Department of Veterans Affairs Outpatient Clinic
2090 Kenny Road
Columbus, OH 43221.
Tel: (614)-469-2547
FAX: (614)-469-5113

October 13, 1994
EXECUTIVE SUMMARY

Quality of life and health status are now considered important outcomes in evaluating drug therapy. Experts agree that health status is a multidimensional construct and should be measured with multiple indicators. However, in interpreting the results of research which have multiple indicators of the same construct, it is important to make decisions about which of the indicators are important than others. It is the premise of this study that the weighting of such indicators must be based on patient preferences for the effects measured by the indicators. In other words, what is needed is a method to quantify how much more desirable a given effect of a drug on health status is as compared to another. However, the most commonly used health status preference measurement methods suffer from various drawbacks, including the inability to consider all possible interactions between health status dimensions and using patients' judgments instead of choices as the response measure. Choice-based Conjoint Analysis is a promising new preference measurement method which is capable of handling interactions between health status dimensions and is based upon choice data.

The purpose of this study is to measure patients' health status preferences using Choice-based Conjoint Analysis and compare the results so obtained with those of the commonly used health status preference measurement methods, i.e., the rating scale and standard gamble methods. The study will also describe and differentiate between patients' preferences for health status on the basis of demographic and disease-related characteristics.

The conceptualization and operationalization of health status adopted for this study is that of the EuroQol.® The study sample will be drawn from a sampling frame consisting of all hypertensive patients who are taking at least one antihypertensive drug and are currently being treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic in Columbus, Ohio. The estimated sample size is 130 patients. Data will be collected by personal interviews. The health status preferences of each patient will be measured by three different preference measurement methods. Data will be analyzed using multi-attribute preference theory, the CBC System for Choice-based Conjoint Analysis, and SAS.

By measuring and quantifying patients' preferences for health status, this study will provide clinicians with valuable information which they can use as inputs in therapeutic decision making and counseling for more
compliant patient behavior. Further, by introducing and validating a new health status preference measurement method and exploring the relationship between this new method and existing methods, the study will advance the field of health status preference measurement.

INTRODUCTION

Quality of life and health status are now considered important outcomes in evaluating drug therapy. The current interest in measuring health status reflects the coalescence of several related trends. Primary among these is the gradual change in the focus of medical care from decreasing mortality to limiting morbidity and the patient-reported impact of that morbidity. This change has been brought about mainly due to the nature of medical conditions observed in society, which have shifted from a predominance of acute and infectious diseases to a predominance of chronic conditions. Chronic conditions, for the most part, are incurable. The emphasis of therapy in such conditions, then, is usually a reduction in severity rather than cure or prolongation of life.

Further, there are behavioral consequences that come into play in such cases. Since chronic illnesses usually require lifelong therapy, the role of the patient in following the therapy becomes crucial in determining the effectiveness of therapy. If patients cannot see or feel a noticeable improvement in their functioning or well-being, they may be inclined to discontinue or be noncompliant with therapy. This in turn would have deleterious consequences for the health of the patient. In other words, patients need a measure of health which is centered around how they feel and function, since they may be unlikely to accept any therapy -- whatever its clinical or scientific merit -- if they see nothing in it for themselves.

Health-care providers are increasingly becoming convinced that one of the primary objectives of medical care is the enhancement of health status and quality of life of the patient. Indeed, the measurement of the patient's quality of life forms the centerpiece of the "outcomes management" movement, which is arguably a driving force behind medical care in the 1990s.

Policy makers are also driving the interest in health status and quality of life measurement from a regulatory perspective. This is because such information can be used in pharmacoeconomic evaluations, which compare
the costs and outcomes of alternative therapies. Regulatory authorities in Australia and Canada have begun requesting such information from pharmaceutical manufacturers along with safety and efficacy data for drug approval and registration decisions in these countries. In the US, the Food and Drug Administration Oncologic Drugs Advisory Committee has recommended that health-related quality of life data be considered along with survival data in the approval decision of anticancer drugs. Pharmaceutical manufacturers have now established entire departments whose sole function is the collection and analysis of quality of life data.

NATURE OF THE PROBLEM

In response to these demands, health-care researchers have developed several questionnaires which measure health status. Although there is no consensus on what specific dimensions of health should be included in a measure of health status, most commonly-used questionnaires conceptualize health status as comprising three major dimensions -- physical functioning, emotional functioning, and social functioning. Within each of these broad dimensions, there are several sub-categories, e.g., physical functioning may be viewed as consisting of mobility (i.e., the ability to walk about) and self-care (i.e., the ability to take care of oneself). However, although the specific dimensions to be included in a measure of health status may vary according to the purpose of the measure, there does seem to be a consensus among experts in the field that health status is a multidimensional construct and should be measured with multiple indicators.

In interpreting the results of research which have multiple indicators of the same construct, it is important to make decisions about which of the indicators are important and which can be ignored. When the evidence from one indicator conflicts with that from another, there must be some way of deciding which indicator carries the most weight. This is especially true for health status measurement questionnaires, which are usually multidimensional. For instance, a drug therapy may have beneficial effects on physical functioning but adverse effects on emotional functioning for the same patient. Most current health status measurement instruments can measure these beneficial and adverse effects on health status. However, they offer little guidance on how these conflicting effects have to be weighted in an overall drug therapy decision for the patient. It is the premise of this study that the weighting of such effects must be based on patient preferences for the effects. In other words, what is needed is a method to quantify how
much more desirable a given effect of a drug on health status is as compared to another.

The most commonly used health status preference measurement methods are the rating scale and standard gamble methods.\textsuperscript{11-14} The rating scale method requires patients to rate different states of health on a 0-100 point scale, where 0 = least desirable and 100 = most desirable. The standard gamble method requires patients to evaluate different states of health by means of a gamble between a health state that is certain versus a wager between two other health states. Although complex, this latter method explicitly incorporates the notion of the probabilistic nature of the effects of drug therapy on health status.

However, the drawbacks of these methods include the inability to consider all possible interactions between health status dimensions and the fact that they use patients' judgments instead of choices as the response measure. The task of measuring patient preferences for health status is complicated by the fact that health status dimensions may interact with one another -- i.e., patients' preferences for physical functioning may be dependent on the particular level of emotional functioning they experience. Therefore, if any method does not consider these interactions, it may not truly measure patients' preferences for health status.

In order to understand the judgment versus choice distinction, it is necessary to think of the reason why patient preferences for health status need to be measured -- for the purpose of making a decision between several alternative therapies as to which one to prescribe for a patient. Making such a decision involves making a choice of one alternative over the others. Judgmental data (as obtained by the commonly used preference measurement methods) do not have a theoretical rationale linking them to choices -- choice data, on the other hand, have such a rationale. \textsuperscript{15} Choice-based Conjoint Analysis is a promising new preference measurement method which is capable of handling interactions between health status dimensions and is based upon choice data. Hence, this new method seems to offer advantages over the currently used health status preference measurement methods.
OBJECTIVES

The purpose of this study is to measure patients' health status preferences using Choice-based Conjoint Analysis and compare the results so obtained with those of the commonly used health status preference measurement methods, i.e., the rating scale and standard gamble methods. The study will also describe and differentiate between patients' preferences for health status on the basis of demographic and disease-related characteristics. The specific tasks to be undertaken in order to achieve these objectives are listed in Appendix 1.

METHODOLOGY

The conceptualization and operationalization of health status adopted for this study is that of the EuroQoL® (see Appendix 2). The EuroQoL® is a health status measurement questionnaire developed by an interdisciplinary group of researchers from five countries in Europe. It conceptualizes health status as consisting of five different dimensions -- mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each of these five dimensions is further sub-divided into three levels of functioning. Therefore, a patient's state of health at a given point in time is viewed as a combination of one level from each of the five dimensions of the EuroQoL®. Published applications of the EuroQoL® have not extended outside Europe, although this measure has universal applicability. Hence, this study will be the first to apply the EuroQoL® in the US.

The target population for this study is hypertensive patients taking one or more antihypertensive drugs. It was decided to use hypertensive patients because previous studies of antihypertensive drugs on health status have shown that these drugs have differential effects on health status dimensions and, given the chronic nature of hypertension, health status is considered an important outcome of antihypertensive drug therapy. The study sample will be drawn from a sampling frame consisting of all hypertensive patients who are taking at least one antihypertensive drug and are currently being treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic in Columbus, Ohio. The estimated sample size is 130 patients. Assuming an 85% participation rate, 153 patients will be approached to participate in the study.
Data collection will proceed through the following stages. Sampled patients (see above) will be contacted over the telephone and requested to participate in the study. Patients who agree to participate will be sent a letter confirming their appointment time and date. Each participant will be interviewed by the principal investigator on the premises of the Clinic. The interview will last approximately one and a half hours. The interview will begin by obtaining background demographic and disease-related information from participants. This will be followed by the three preference measurement methods. The order of the three methods will be varied at random in order to avoid any learning effects. After each of the three methods is completed, feedback will be obtained about each specific method. Finally, at the end of the interview, an overall feedback questionnaire will be administered. One day after the interview, a personalized thank-you letter will be mailed to all study participants. All data collection letters and questionnaires are provided in Appendix 3.

The data will be analyzed at The Ohio State University College of Pharmacy. Preference measurement analysis will be conducted using multi-attribute preference theory (for rating scale and standard gamble methods) and the CBC System (for Choice-based Conjoint Analysis). Sub-group analysis will be conducted using SAS. After completion of the analysis, the Clinic will be provided with a final study report.

SIGNIFICANCE

Chronic conditions predominate in medical practice today. For such conditions, the role of the patient in following the therapeutic regimen becomes crucial in determining the effectiveness of therapy. However, as Goldsmith has noted, patients today are “increasingly vigilant and demanding ... increasingly suspicious of medical direction ... increasingly specific about their needs and how they want them met ....”20 In their study of hypertensive patients, Strull et al. found that patients desired information about their condition and wanted to discuss alternative therapies with their physicians, but did not desire to be very involved in actual clinical decision making.21 Given these findings, it appears that although patients want to let physicians know their preferences for the outcomes of therapy and discuss therapeutic alternatives with their physicians, they would prefer to let physicians make the final therapeutic decision on their behalf after incorporating their preferences for the outcomes of therapy. This then would lead to greater acceptance of the drug therapy decisions by patients by what
Mahler and Kulik have termed "sense of control," which in turn is associated with more compliant behavior by patients. As mentioned earlier, patient compliance is crucial in determining the effectiveness of therapy. Therefore, by measuring and quantifying patients' preferences for health status, this study will provide clinicians with valuable information which they can use as inputs in therapeutic decision making and counseling for more compliant patient behavior.

On a more methodological front, this study will advance the field of health status preference measurement by introducing and validating a new method for measuring health status preferences. By illustrating and empirically demonstrating the relationship between this new method and existing methods, greater understanding of the strengths and weaknesses of alternative methods may be gained. This in turn will prove valuable for use in future research on health status preference measurement methodology.
REFERENCES


APPENDIX O

MINUTES AND OFFICIAL RESPONSE FROM THE VAOPC RESEARCH AND DEVELOPMENT COMMITTEE
1. The meeting was called to order by the Chair at 2:00 p.m.

2. The minutes of the July 21, 1994, meeting were not reviewed by the Committee and hence, not approved.

3. BUSINESS

   a. Dr. Henry Nasrallah has been approved for membership on the Research and Development Committee.

   b. **TOPIC: Proposal Titled “Health Status Preference Measurement, a Comparison of Three Techniques”**. This is a proposal submitted by Zafar Hakim.

**Discussion:** Veterans should be informed when they are called that the information they give will not influence their care. That is to be added to the information provided to the patient. This must be on the standard VA research consent form. It must also be noted that a copy of the consent form should go in to each veteran's record.
Action: Dr. Carlson moved to approve the protocol. This was seconded by Dr. Bauman pending approval by Dr. Kirkpatrick of the consent form and approval by The Ohio State University College of Pharmacy Human Subjects Committee.

Resolution: Dr. Kirkpatrick will write a letter to Dr. Pat Hack, who is the principle investigator from OSU and in this letter he will request: the Human Subjects Committee approval and a rewrite of the informed consent using VA Form 10-1086 with all its listed components. Once these submissions have been reviewed and approved by the Chairperson, Mr. Hakim will be able to proceed with the protocol.

4. There being no further business, the meeting was adjourned.

Submitted by:

WILLIAM ARNESON, R.Ph.
Recorder

RECOMMEND APPROVAL/DISAPPROVAL

ROBERT B. KIRKPATRICK, M.D.
Chairperson

APPROVED/DISAPPROVED

LILIANT T. THOME, M.D.
Director
October 26, 1994

Dear Dr. Pathak,

This is to inform you that your proposal, "Health Status Preference Measurement: A Comparison of Three Techniques" has been approved by the Columbus VAOPC Research and Development (R&D) Committee pending submission of: (a) the OSU Human Subjects Committee Approval and (b) a rewrite of the informed consent form using VA Form 10-1086 with all the listed components. Once these submissions have been reviewed and approved, you will be notified.

Please call Ruth Huisman, Quality Assurance Coordinator, if you have any questions regarding this issue. Her telephone number is (614) 469-2006.

Sincerely,

Robert Kirkpatrick, M.D.
Chairperson, Research and Development Committee

cc:

119
APPENDIX P

RESEARCH CONSENT FORMS
Subject Name: _________________________________________ Date ______________

Title of Study: Health Status Preference Measurement: A Comparison of Three Techniques.

Principal Investigator: Dev S. Pathak VAMC: Columbus, Ohio

DESCRIPTION OF RESEARCH BY INVESTIGATOR

1. Purpose of study and how long it will last:
2. Description of the study including procedures to be used:
3. Description of any procedures that may result in discomfort or inconvenience:
4. Expected risks of study:
5. Expected benefits of study:
6. Other treatment available:
7. Use of research results:
8. Special circumstances:
The purpose of this study is to understand what exactly patients being treated by the Columbus VA Outpatient Clinic like and dislike about drug therapy and the effects of drug therapy on health.

Participation in this study is entirely voluntary. If you decide to participate in the study, you will be asked some questions by the study interviewer. At first, you will be asked to provide some background information about yourself and your health in general. Next, your likes and dislikes about drug therapy and the effects of drug therapy on health will be measured in three different ways. One way involves rating different hypothetical or imaginary drug therapies on a 0-100 scale, in which 0 is the least desirable drug therapy and 100 is the most desirable drug therapy. A second way involves choosing what in your opinion is the most desirable of four alternative hypothetical or imaginary drug therapies, which will be described on the basis of their effects on your health. The third way involves choosing between two imaginary drug therapies, one of which has known effects on your health and the other has a certain chance of producing a good effect and a certain chance of producing a bad effect on your health. Please note that in all three ways, the drug therapies you will be evaluating are imaginary and have nothing to do with the drug therapy you are currently taking. After these three tasks, your feedback will be obtained on each task. We are also interested in finding out which task you feel is the best.

The entire interview is expected to last for about an hour and a half. We have been careful in selecting the questions to ask you in order to avoid putting you under undue stress. However, it is possible that some of the descriptions of the imaginary drug therapies may be upsetting to you. Therefore, if at any point during the interview you feel uncomfortable and wish to stop the interview, please feel free to do so. Such a decision will not affect the quality of care you receive at the VA and will not be held against you in any way.

Moreover, you decide not to participate in the study, your decision will not affect the quality of care you receive at the VA in any way. We would like to repeat that participation in this study is entirely voluntary, i.e., it is up to you.

This study will help health-care professionals such as doctors, nurses, and pharmacists understand what exactly patients like yourself are looking to gain from their drug therapy. These professionals can then include this information in their decisions so as to improve the quality of life of the patients under their care.
<table>
<thead>
<tr>
<th>Subject Name:</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title of Study: Health Status Preference Measurement: A Comparison of Three Techniques</td>
<td></td>
</tr>
<tr>
<td>Principal Investigator: Dev S. Pathak</td>
<td>VAMC: Columbus, OH</td>
</tr>
</tbody>
</table>

RESEARCH SUBJECTS' RIGHTS: I have read or have had read to me all of the above. Dr. ______________________ has explained the study to me and answered all of my questions. I have been told of the risks or discomforts and possible benefits of the study. I have been told of other choices of treatment available to me.

I understand that I do not have to take part in this study, and my refusal to participate will involve no penalty or loss of rights to which I am entitled. I may withdraw from this study at any time without penalty or loss of VA or other benefits to which I am entitled.

The results of this study may be published, but my records will not be revealed unless required by law.

In case there are medical problems or questions, I have been told I can call Dr. _______________________ at __________________ during the day and Dr. ______________________ at __________________ after hours. If any medical problems occur in connection with this study the VA will provide emergency care.

I understand my rights as a research subject, and I voluntarily consent to participate in this study. I understand what the study is about and how and why it is being done. I will receive a signed copy of this consent form.

Subject's Signature | Date |
---------------------|------|
Signature of Subject's Representative* | Subject's Representatives |
Signature of Witness | Witness (print) |
Signature of Investigator |

*Only required if subject not competent.
CONSENT FOR PARTICIPATION IN RESEARCH

I consent to participate in research entitled Health Status Preference Measurement: A Comparison of Three Techniques, which is jointly being conducted by the Department of Veterans Affairs Outpatient Clinic in Columbus, Ohio and The Ohio State University.

____________________, the principal investigator, or his/her authorized representative has explained the purpose of the study, the procedures to be followed, and the expected duration of my participation.

I acknowledge that I have had the opportunity to obtain additional information regarding the study and that any questions I have raised have been answered to my full satisfaction. Further, I understand that I am free to withdraw consent at any time and to discontinue participation in the study without prejudice to me.

Finally, I acknowledge that I have read and fully understand this consent form. I sign it freely and voluntarily. A copy has been given to me. I also understand that a copy of this consent form will be placed in my medical record.

Date: __________________

Signed: ____________________ ____________________
(principal investigator or his/her authorized representative) (participant)

Witness: ____________________
APPENDIX Q

VA RESEARCH AND DEVELOPMENT PROJECT DATA SHEET
**Health Status Preference Measurement: A Comparison of Three Techniques**

**Funding and Administration**

<table>
<thead>
<tr>
<th>Funding Code</th>
<th>Name if &quot;Other&quot;</th>
<th>Admin Code</th>
<th>Name if &quot;Other&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 0 0 0</td>
<td></td>
<td>0 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Research Focus**

- Health Status
- Health Status Indicators
- Quality of Life
- Abstract (cover an envelope)

**Signature of Principal Investigator**

(Date)
Abstract: Organize under the following headings: (a) objective; (b) research plan; (c) methodology, and (d) findings. Limit to 500 words.

ABSTRACT

Introduction: Quality of life and health status are now considered important outcomes in evaluating drug therapy. Experts agree that health status is a multidimensional construct and should be measured with multiple indicators. However, in interpreting the results of research which have multiple indicators of the same construct, it is important to make decisions about which of the indicators are important than others. It is the premise of this study that the weighting of such indicators must be based on patient preferences for the effects measured by the indicators. However, the commonly used health status preference measurement methods suffer from various drawbacks, including the inability to consider all possible interactions between health status dimensions and using patients' judgments instead of choices as the response measure. Choice-based Conjoint Analysis is a promising new method which is capable of overcoming these drawbacks.

Objectives: The purpose of this study is to measure patients' health status preferences using Choice-based Conjoint Analysis and compare the results so obtained with those of the commonly used health status preference measurement methods, i.e., the rating scale and standard gamble methods. The study will also describe and differentiate between patients' preferences for health status on the basis of demographic and disease-related characteristics.

Research Plan & Methodology: The conceptualization and operationalization of health status adopted for this study is that of the EuroQol®. The target population for this study is hypertensive patients taking one or more antihypertensive drugs. The study sample will be drawn from a sampling frame consisting of all hypertensive patients who are taking at least one antihypertensive drug and are currently being treated on an outpatient basis at the Department of Veterans Affairs Outpatient Clinic in Columbus, Ohio. The estimated sample size is 150 patients. Assuming an 85% participation rate, 153 patients will be approached to participate in the study. Sampled patients will be contacted over the telephone and requested to participate in the study. Each participant will be interviewed by a study investigator on the Clinic premises. The interview will begin by obtaining background demographic and disease-related information from participants. This will be followed by the three preference measurement methods. The order of the three methods will be varied at random in order to avoid any learning effects. After each of the three methods is completed, feedback will be obtained about each specific method. Finally, an overall feedback questionnaire will be administered. Preference measurement analysis will be conducted using multi-attribute preference theory (for rating scale and standard gamble methods) and the CBC System (for Choice-based Conjoint Analysis). Sub-group analysis will be conducted using SAS.

Significance of Findings: By measuring and quantifying patients' preferences for health status, this study will provide clinicians with valuable information which they can use as inputs in therapeutic decision making and counseling for more compliant patient behavior. Further, by introducing and validating a new health status preference measurement method and exploring the relationship between this new method and existing methods, the study will advance the field of health status preference measurement.
APPENDIX R

BACKGROUND INFORMATION QUESTIONNAIRE
MEASURING PREFERENCES FOR HEALTH OUTCOMES

Background Information

Start Time

End Time
Demographic Information

I would like to begin by asking you some background information about yourself. Since we will not be noting your name or social security number at any point in this interview, we will use this background information to describe patients similar to yourself in very general terms. In this way, we can be sure of representing your personal preferences without running the risk of specifically identifying you in any way.

Please note that you are not obliged in any way to answer these questions. If you feel uncomfortable for any reason whatsoever, please feel free to refuse to answer the question and ask me to move on to the next question. Shall we begin?

1. What is your age in years?

2. Which ONE of the following categories best describes the highest education level that you have achieved?
   1. GRADE SCHOOL OR LESS
   2. SOME HIGH SCHOOL
   3. COMPLETED HIGH SCHOOL
   4. TECHNICAL SCHOOL
   5. SOME COLLEGE
   6. COLLEGE DEGREE
   7. SOME GRADUATE WORK
   8. GRADUATE DEGREE
   9. OTHER (PLEASE SPECIFY)

3. Which ONE of the following categories best describes your main activity?
   1. IN EMPLOYMENT OR SELF-EMPLOYMENT
   2. RETIRED
   3. SEEKING WORK
   4. OTHER (PLEASE SPECIFY)
Severity Information

1. For how long have you been diagnosed as having high blood pressure?
   
   1. 1 YEAR OR LESS
   2. MORE THAN 1 YEAR BUT NO MORE THAN 5 YEARS
   3. MORE THAN 5 YEARS BUT NO MORE THAN 10 YEARS
   4. MORE THAN 10 YEARS BUT NO MORE THAN 15 YEARS
   5. MORE THAN 15 YEARS BUT NO MORE THAN 20 YEARS
   6. MORE THAN 20 YEARS BUT NO MORE THAN 25 YEARS
   7. MORE THAN 25 YEARS BUT NO MORE THAN 30 YEARS
   8. MORE THAN 30 YEARS

2. Patient's most recent blood pressure reading (get from chart)
   
   SYSTOLIC
   DIASTOLIC

3. Classification of patient (see JNC V)

   Systolic (mm Hg) | Diastolic (mm Hg)
   1. NORMAL        | < 130         | < 85
   2. HIGH NORMAL   | 130-139       | 85-89

   HYPERTENSION:
   3. MILD (STAGE 1) | 140-159       | 90-99
   4. MODERATE (STAGE 2) | 160-179      | 100-109
   5. SEVERE (STAGE 3) | 180-209       | 110-119
   6. VERY SEVERE (STAGE 4) | $\geq 210$   | $\geq 120$
Comorbidity Information

1. Besides high blood pressure, what other health problems have you had over the past five years? (call out each one in turn and wait for patient to respond)

1. ACQUIRED IMMUNODEFICIENCY SYNDROME (AIDS)
2. ALLERGIES
3. AMPUTATION OF ARM OR LEG
4. ARTHRITIS OR RHEUMATISM
5. BACK PAIN
6. BLOOD DISORDERS (e.g., ANEMIA, HEMOPHILIA)
7. BROKEN HIP OR OTHER BONE (FRACTURE)
8. CANCER OR LEUKEMIA
9. DEPRESSION
10. DIABETES
11. EPILEPSY
12. EYE, VISION, OR SIGHT PROBLEMS
13. HEARING PROBLEMS
14. HEART TROUBLE
15. HIGH CHOLESTEROL/HIGH BLOOD PRESSURE
16. KIDNEY DISEASE
17. LIVER DISEASE
18. LUNG DISEASE
19. MIGRAINE
20. OSTEOPOROSIS
21. PARKINSON'S DISEASE
22. PROSTATE TROUBLE
23. SLEEP DISORDERS
24. SPEECH PROBLEMS
25. THYROID PROBLEMS
26. ULCERS OF STOMACH AND/OR INTESTINE
27. URINARY TRACT DISORDERS
28. OTHER (PLEASE SPECIFY)
MEASURING PREFERENCES FOR HEALTH OUTCOMES

Summary Information

Position of Interview: 1. Before Appointment 2. After Appointment

Order of Tasks: 1. FSC 2. PCS 3. CFS

Overall Start Time
Overall End Time
APPENDIX T

MATERIALS FOR RECRUITING PATIENTS OVER THE TELEPHONE
Good ________ Mr./Ms. __________.

My name is ZAFAR HAKIM, and I am a graduate student in pharmacy at The Ohio State University. I am currently working on an important study with the pharmacy service at the VA Outpatient Clinic in Columbus, Ohio and I would like to invite you to participate in this study by answering a few questions on your next scheduled visit to the VA Outpatient Clinic.

The purpose of this study is to understand what exactly people such as yourself like and dislike about their drug therapy. If you decide to participate, you will answer some questions about the effects of drugs. These questions will take about one and a half hours and can easily be scheduled for a time before your next visit to the VA Outpatient Clinic. You will also sign an informed consent form, a copy of which will be placed in your medical records so that your doctor will know that you have taken part in this study. Please note that if you decide not to participate in this study, it will not affect the quality of care you receive at the VA in any way.

To make it more convenient for you, I will be present at the VA so that I can ask you the questions at the clinic itself. And yes, whatever you say, will be held in confidence. This study is jointly being conducted by The Ohio State University and the Department of Veterans Affairs, and the results of this study will help us get a clearer understanding of what exactly people such as yourself are looking for to gain from your drug therapy.

Would you, then, like to be a part of this study?

______________________________________________________________

IF YES, SCHEDULE A TIME FOR INTERVIEW

IF NO, THANK PATIENT AND HANG UP

EITHER WAY, FILL OUT DATASHEET FOR TELEPHONE RECRUITS
Fall-back Answers to Commonly-asked Questions

1. How did you get my telephone number?

YOUR TELEPHONE NUMBER WAS OBTAINED FROM THE ADMINISTRATIVE RECORDS AT THE VA OUTPATIENT CLINIC AT COLUMBUS, OHIO.

2. Who is conducting this study?

THIS STUDY IS JOINTLY BEING CONDUCTED BY THE OHIO STATE UNIVERSITY AND THE VA OUTPATIENT CLINIC AT COLUMBUS, OHIO.

3. Who is sponsoring this study?

THIS STUDY IS FUNDED ENTIRELY BY RESEARCH FUNDS FROM THE OHIO STATE UNIVERSITY COLLEGE OF PHARMACY.

4. What is the purpose of this study?

THE PURPOSE OF THIS STUDY IS TO UNDERSTAND THE PREFERENCES OF PATIENTS SUCH AS YOURSELF FOR THE OUTCOMES OF DRUG THERAPY.

5. What will I have to do if I agree to participate in this study?

AS A STUDY PARTICIPANT, ALL YOU HAVE TO DO IS MAKE AN APPOINTMENT WITH THE STUDY INVESTIGATOR. THIS APPOINTMENT WILL LAST ABOUT ONE AND A HALF HOURS AND CAN EASILY BE SCHEDULED FOR A TIME BEFORE YOUR NEXT PHYSICIAN VISIT AT THE VA. DURING THIS APPOINTMENT, YOU WILL BE ASKED YOUR OPINION ABOUT SOME HYPOTHETICAL OR IMAGINARY DRUG THERAPIES.

6. How will the study results be used?

THE STUDY WILL HELP HEALTH-CARE PROFESSIONALS UNDERSTAND WHAT EXACTLY PATIENTS SUCH AS YOURSELF ARE LOOKING TO GAIN FROM YOUR DRUG THERAPY. THESE PROFESSIONALS CAN THEN INCORPORATE SUCH PREFERENCES IN THEIR DECISIONS SO AS TO IMPROVE THE QUALITY OF LIFE OF THE PATIENTS UNDER THEIR CARE.
## Datasheet for Telephone Recruits

<table>
<thead>
<tr>
<th>NAME</th>
<th>SS #</th>
<th>TELEPHONE</th>
<th>RESPONSE</th>
<th>DATE</th>
<th>TIME</th>
<th>REMARKS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td>Maybe</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

1.
APPENDIX U

CODEBOOK
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line # 1</td>
<td>1 - 3 RN</td>
<td>Respondent Number (001 to 140)</td>
</tr>
<tr>
<td>4 - 9 DATEIN</td>
<td>Date of interview (MMDDYY)</td>
<td></td>
</tr>
<tr>
<td>10 ACCOMP</td>
<td>Accompanied by other: 1 = no 2 = yes</td>
<td></td>
</tr>
<tr>
<td>11 PI</td>
<td>Position of Interview: 1 = before appointment 2 = after appointment 3 = between appointments</td>
<td></td>
</tr>
</tbody>
</table>
| 12 ORDER | Order of Preference Measurement Tasks, where  
F = feeling thermometer  
S = standard gamble  
C = choice-based conjoint  
1 = FSC  
2 = FCS  
3 = CFS  
Note: other combinations (i.e., SFC, SCF, & CSF) were not possible since S had to be conducted after F |
<p>| 13 - 15 OTIME | Overall Time of Interview (in minutes) |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 - 17</td>
<td>AGE</td>
<td>Age (in years)</td>
</tr>
</tbody>
</table>
| 18      | EDUC     | Highest Education Level:  
|         |          | 1 = grade school or less  
|         |          | 2 = some high school  
|         |          | 3 = completed high school  
|         |          | 4 = technical school  
|         |          | 5 = some college  
|         |          | 6 = college degree  
|         |          | 7 = some graduate work  
|         |          | 8 = graduate degree  
|         |          | 9 = other |
| 19      | MACT     | Main Activity:  
|         |          | 1 = self/other employed  
|         |          | 2 = retired  
|         |          | 3 = seeking work  
|         |          | 4 = disabled  
|         |          | 5 = other |
| 20      | LDIAG    | Length of Time since Diagnosed of Hypertension:  
|         |          | 1 = 1 year or less  
|         |          | 2 = 1+ but NMT 5 years  
|         |          | 3 = 5+ but NMT 10 years  
|         |          | 4 = 10+ but NMT 15 years  
|         |          | 5 = 15+ but NMT 20 years  
|         |          | 6 = 20+ but NMT 25 years  
|         |          | 7 = 25+ but NMT 30 years  
|         |          | 8 = more than 30 years |
| 21 - 23 | SYSBP    | Systolic Blood Pressure  
<p>|         |          | (in mm Hg) |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>24 - 26</td>
<td>DIASBP</td>
<td>Diastolic Blood Pressure (in mm Hg)</td>
</tr>
</tbody>
</table>
| 27      | SEV      | Severity of Hypertension (according to JNC V):  
|         |          | 1 = normal  
|         |          | 2 = high normal  
|         |          | 3 = mild hypertension  
|         |          | 4 = moderate hypertension  
|         |          | 5 = severe hypertension  
|         |          | 6 = very severe hypertension  
|         |          | Note: when systolic and diastolic blood pressures fall into different categories, the higher category was selected for classification of severity (as per JNC V) |
| 28 - 33 | DATEBP   | Date of Blood Pressure Reading (MMDDYY) |
| 34      | OHP1     | Other Health Problems # 1 -- AIDS:  
|         |          | 1 = no  
|         |          | 2 = yes |
| 35      | OHP2     | Other Health Problems # 2 -- Allergies:  
|         |          | 1 = no  
<p>|         |          | 2 = yes |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>OHP3</td>
<td>Other Health Problems # 3 -- Amputation of Arm or Leg: 1 = no 2 = yes</td>
</tr>
<tr>
<td>37</td>
<td>OHP4</td>
<td>Other Health Problems # 4 -- Arthritis/Rheumatism: 1 = no 2 = yes</td>
</tr>
<tr>
<td>38</td>
<td>OHP5</td>
<td>Other Health Problems # 5 -- Back Pain: 1 = no 2 = yes</td>
</tr>
<tr>
<td>39</td>
<td>OHP6</td>
<td>Other Health Problems # 6 -- Blood Disorders: 1 = no 2 = yes</td>
</tr>
<tr>
<td>40</td>
<td>OHP7</td>
<td>Other Health Problems # 7 -- Fractures: 1 = no 2 = yes</td>
</tr>
<tr>
<td>41</td>
<td>OHP8</td>
<td>Other Health Problems # 8 -- Cancer: 1 = no 2 = yes</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>42</td>
<td>OHP9</td>
<td>Other Health Problems # 9 --</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Depression:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>43</td>
<td>OHP10</td>
<td>Other Health Problems # 10 --</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diabetes:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>44</td>
<td>OHP11</td>
<td>Other Health Problems # 11 --</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Epilepsy:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>45</td>
<td>OHP12</td>
<td>Other Health Problems # 12 --</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sight Problems:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>46</td>
<td>OHP13</td>
<td>Other Health Problems # 13 --</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hearing Problems:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>47</td>
<td>OHP14</td>
<td>Other Health Problems # 14 --</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Heart Trouble:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>48</td>
<td>OHP15</td>
<td>Other Health Problems #15 -- Hyperlipidemia:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>49</td>
<td>OHP16</td>
<td>Other Health Problems #16 -- Kidney Disease:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>50</td>
<td>OHP17</td>
<td>Other Health Problems #17 -- Liver Disease:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>51</td>
<td>OHP18</td>
<td>Other Health Problems #18 -- Lung Disease:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>52</td>
<td>OHP19</td>
<td>Other Health Problems #19 -- Migraine:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>53</td>
<td>OHP20</td>
<td>Other Health Problems #20 -- Osteoporosis:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>54</td>
<td>OHP21</td>
<td>Other Health Problems # 21 -- Parkinson’s Disease:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>55</td>
<td>OHP22</td>
<td>Other Health Problems # 22 -- Prostate Trouble:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>56</td>
<td>OHP23</td>
<td>Other Health Problems # 23 -- Sleep Disorders:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>57</td>
<td>OHP24</td>
<td>Other Health Problems # 24 -- Speech Problems:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>58</td>
<td>OHP25</td>
<td>Other Health Problems # 25 -- Thyroid Problems:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>59</td>
<td>OHP26</td>
<td>Other Health Problems # 26 -- Ulcers of Stomach/Intestine:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = no</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = yes</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| 60      | OHP27    | Other Health Problems # 27 -- Urinary Tract Disorders:  
|         |          | 1 = no  
|         |          | 2 = yes |
| 61      | OHP28    | Other Health Problems # 28 -- Other Problems:  
|         |          | 1 = no  
|         |          | 2 = yes |
| 62 - 63 | BTIME    | Time to complete Background Information (in minutes) |
| 64      | TMOB     | Mobility Today:  
|         |          | 1 = no problems walking about  
|         |          | 2 = some problems walking about  
|         |          | 3 = confined to bed |
| 65      | TSFC     | Self-care Today:  
|         |          | 1 = no problems with self-care  
|         |          | 2 = some problems washing/dressing self  
<p>|         |          | 3 = unable to wash/dress self |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| 66      | TACT     | Usual Activities Today:  
1 = no problems performing usual activities  
2 = some problems performing usual activities  
3 = unable to perform usual activities |
| 67      | TPDF     | Pain/Discomfort Today:  
1 = no pain or discomfort  
2 = moderate pain or discomfort  
3 = extreme pain or discomfort |
| 68      | TANX     | Anxiety/Depression Today:  
1 = not anxious or depressed  
2 = moderately anxious or depressed  
3 = extremely anxious or depressed |
| 69      | TGEN     | General Health Today versus 1 Year Ago:  
1 = better  
2 = much the same  
3 = worse |
<p>| 70-71   | EQTIME   | Time to complete EuroQol (in minutes) |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 3</td>
<td>FTM1</td>
<td>Feeling Thermometer Score for Card M1</td>
</tr>
<tr>
<td>4 - 6</td>
<td>FTM2</td>
<td>Feeling Thermometer Score for Card M2</td>
</tr>
<tr>
<td>7 - 9</td>
<td>FTM3</td>
<td>Feeling Thermometer Score for Card M3</td>
</tr>
<tr>
<td>10 - 12</td>
<td>FTS1</td>
<td>Feeling Thermometer Score for Card S1</td>
</tr>
<tr>
<td>13 - 15</td>
<td>FTS2</td>
<td>Feeling Thermometer Score for Card S2</td>
</tr>
<tr>
<td>16 - 18</td>
<td>FTS3</td>
<td>Feeling Thermometer Score for Card S3</td>
</tr>
<tr>
<td>19 - 21</td>
<td>FTU1</td>
<td>Feeling Thermometer Score for Card U1</td>
</tr>
<tr>
<td>22 - 24</td>
<td>FTU2</td>
<td>Feeling Thermometer Score for Card U2</td>
</tr>
<tr>
<td>25 - 27</td>
<td>FTU3</td>
<td>Feeling Thermometer Score for Card U3</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>28 - 30</td>
<td>FTP1</td>
<td>Feeling Thermometer Score for Card P1</td>
</tr>
<tr>
<td>31 - 33</td>
<td>FTP2</td>
<td>Feeling Thermometer Score for Card P2</td>
</tr>
<tr>
<td>34 - 36</td>
<td>FTP3</td>
<td>Feeling Thermometer Score for Card P3</td>
</tr>
<tr>
<td>37 - 39</td>
<td>FTA1</td>
<td>Feeling Thermometer Score for Card A1</td>
</tr>
<tr>
<td>40 - 42</td>
<td>FTA2</td>
<td>Feeling Thermometer Score for Card A2</td>
</tr>
<tr>
<td>43 - 45</td>
<td>FTA3</td>
<td>Feeling Thermometer Score for Card A3</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>Line #3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 3</td>
<td>FTAB</td>
<td>Feeling Thermometer Score for Card AB</td>
</tr>
<tr>
<td>4 - 6</td>
<td>FTAW</td>
<td>Feeling Thermometer Score for Card AW</td>
</tr>
<tr>
<td>7 - 9</td>
<td>FTCM</td>
<td>Feeling Thermometer Score for Card CM</td>
</tr>
<tr>
<td>10 - 12</td>
<td>FTCS</td>
<td>Feeling Thermometer Score for Card CS</td>
</tr>
<tr>
<td>13 - 15</td>
<td>FTCU</td>
<td>Feeling Thermometer Score for Card CU</td>
</tr>
<tr>
<td>16 - 18</td>
<td>FTCP</td>
<td>Feeling Thermometer Score for Card CP</td>
</tr>
<tr>
<td>19 - 21</td>
<td>FTCA</td>
<td>Feeling Thermometer Score for Card CA</td>
</tr>
<tr>
<td>22 - 24</td>
<td>FTBO</td>
<td>Feeling Thermometer Score for Card BO</td>
</tr>
<tr>
<td>25 - 27</td>
<td>FTE3</td>
<td>Feeling Thermometer Score for Card E3</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>28 - 30</td>
<td>FTE8</td>
<td>Feeling Thermometer Score for Card E8</td>
</tr>
<tr>
<td>31 - 33</td>
<td>FTHT</td>
<td>Feeling Thermometer Score for Card HT</td>
</tr>
<tr>
<td>34 - 36</td>
<td>FTID</td>
<td>Feeling Thermometer Score for Card ID</td>
</tr>
<tr>
<td>37 - 38</td>
<td>FTTTIME</td>
<td>Time to complete Feeling Thermometer Tasks (in minutes)</td>
</tr>
</tbody>
</table>
| 39      | FTINT    | Feeling Thermometer Evaluation for Interest:  
|         |          | 1 = extremely boring  
|         |          | 2 = moderately boring  
|         |          | 3 = slightly boring  
|         |          | 4 = neutral  
|         |          | 5 = slightly interesting  
|         |          | 6 = moderately interesting  
|         |          | 7 = extremely interesting |
| 40      | FTSEN    | Feeling Thermometer Evaluation for Sensibility:  
|         |          | 1 = extremely senseless  
|         |          | 2 = moderately senseless  
|         |          | 3 = slightly senseless  
|         |          | 4 = neutral  
|         |          | 5 = slightly sensible  
|         |          | 6 = moderately sensible  
|         |          | 7 = extremely sensible  

Note: reverse coded already
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>FTDDO</td>
<td>Feeling Thermometer Evaluation for Difficulty in Doing:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = extremely difficult to do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = moderately difficult to do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = slightly difficult to do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = neutral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = slightly easy to do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = moderately easy to do</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 = extremely easy to do</td>
</tr>
<tr>
<td>42</td>
<td>FTUSE</td>
<td>Feeling Thermometer Evaluation for Usefulness:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = extremely useless</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = moderately useless</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = slightly useless</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = neutral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = slightly useful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = moderately useful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 = extremely useful</td>
</tr>
<tr>
<td>Note: reverse coded already</td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>FTLEN</td>
<td>Feeling Thermometer Evaluation for Length:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = extremely long</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = moderately long</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = slightly long</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = neutral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = slightly short</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = moderately short</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 = extremely short</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| 44      | FTSUC    | Feeling Thermometer Evaluation for Success:  
          |          |   1 = extremely unsuccessful  
          |          |   2 = moderately unsuccessful  
          |          |   3 = slightly unsuccessful  
          |          |   4 = neutral  
          |          |   5 = slightly successful  
          |          |   6 = moderately successful  
          |          |   7 = extremely successful  
          |          |  Note: reverse coded already |
| 45      | FTCLR    | Feeling Thermometer Evaluation for Clarity:  
          |          |   1 = extremely vague  
          |          |   2 = moderately vague  
          |          |   3 = slightly vague  
          |          |   4 = neutral  
          |          |   5 = slightly clear  
          |          |   6 = moderately clear  
          |          |   7 = extremely clear  
<pre><code>      |          |  Note: reverse coded already |
</code></pre>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| 46      | FTDUN    | Feeling Thermometer Evaluation for Difficulty in Understanding:  
|         |          | 1 = extremely difficult to understand  
|         |          | 2 = moderately difficult to understand  
|         |          | 3 = slightly difficult to understand  
|         |          | 4 = neutral  
|         |          | 5 = slightly easy to understand  
|         |          | 6 = moderately easy to understand  
|         |          | 7 = extremely easy to understand  
| 47      | FTCNF    | Feeling Thermometer Evaluation for Confidence:  
|         |          | 1 = very unconfident  
|         |          | 2 = somewhat unconfident  
|         |          | 3 = neither confident or unconfident  
|         |          | 4 = somewhat confident  
|         |          | 5 = very confident  
<p>| 48 - 49 | FTETIME  | Time to complete Feeling Thermometer Evaluation (in minutes) |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line # 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 3</td>
<td>SGM1</td>
<td>Standard Gamble Score for Card M1</td>
</tr>
<tr>
<td>4 - 6</td>
<td>SGM2</td>
<td>Standard Gamble Score for Card M2</td>
</tr>
<tr>
<td>7 - 9</td>
<td>SGM3</td>
<td>Standard Gamble Score for Card M3</td>
</tr>
<tr>
<td>10 - 12</td>
<td>SGS1</td>
<td>Standard Gamble Score for Card S1</td>
</tr>
<tr>
<td>13 - 15</td>
<td>SGS2</td>
<td>Standard Gamble Score for Card S2</td>
</tr>
<tr>
<td>16 - 18</td>
<td>SGS3</td>
<td>Standard Gamble Score for Card S3</td>
</tr>
<tr>
<td>19 - 21</td>
<td>SGU1</td>
<td>Standard Gamble Score for Card U1</td>
</tr>
<tr>
<td>22 - 24</td>
<td>SGU2</td>
<td>Standard Gamble Score for Card U2</td>
</tr>
<tr>
<td>25 - 27</td>
<td>SGU3</td>
<td>Standard Gamble Score for Card U3</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>--------------------------------------</td>
</tr>
<tr>
<td>28 - 30</td>
<td>SC.P1</td>
<td>Standard Gamble Score for Card P1</td>
</tr>
<tr>
<td>31 - 33</td>
<td>SGP2</td>
<td>Standard Gamble Score for Card P2</td>
</tr>
<tr>
<td>34 - 36</td>
<td>SGP3</td>
<td>Standard Gamble Score for Card P3</td>
</tr>
<tr>
<td>37 - 39</td>
<td>SGA1</td>
<td>Standard Gamble Score for Card A1</td>
</tr>
<tr>
<td>40 - 42</td>
<td>SGA2</td>
<td>Standard Gamble Score for Card A2</td>
</tr>
<tr>
<td>43 - 45</td>
<td>SGA3</td>
<td>Standard Gamble Score for Card A3</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>1 - 3</td>
<td>SGAB</td>
<td>Standard Gamble Score for Card AB</td>
</tr>
<tr>
<td>4 - 6</td>
<td>SGAW</td>
<td>Standard Gamble Score for Card AW</td>
</tr>
<tr>
<td>7 - 9</td>
<td>SGCM</td>
<td>Standard Gamble Score for Card CM</td>
</tr>
<tr>
<td>10 - 12</td>
<td>SGCS</td>
<td>Standard Gamble Score for Card CS</td>
</tr>
<tr>
<td>13 - 15</td>
<td>SGCU</td>
<td>Standard Gamble Score for Card CU</td>
</tr>
<tr>
<td>16 - 18</td>
<td>SGCP</td>
<td>Standard Gamble Score for Card CP</td>
</tr>
<tr>
<td>19 - 21</td>
<td>SGCA</td>
<td>Standard Gamble Score for Card CA</td>
</tr>
<tr>
<td>22 - 24</td>
<td>SGBO</td>
<td>Standard Gamble Score for Card BO</td>
</tr>
<tr>
<td>25 - 27</td>
<td>SGE3</td>
<td>Standard Gamble Score for Card E3</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>-------------</td>
</tr>
<tr>
<td>28 - 30</td>
<td>SGE8</td>
<td>Standard Gamble Score for Card E8</td>
</tr>
<tr>
<td>31 - 33</td>
<td>SGHT</td>
<td>Standard Gamble Score for Card HT</td>
</tr>
<tr>
<td>34 - 36</td>
<td>SGID</td>
<td>Standard Gamble Score for Card ID</td>
</tr>
<tr>
<td>37 - 38</td>
<td>SGTTIME</td>
<td>Time to complete Standard Gamble Tasks (in minutes)</td>
</tr>
</tbody>
</table>
| 39      | SCINT    | Standard Gamble Evaluation for Interest:  
|         |          | 1 = extremely boring  
|         |          | 2 = moderately boring  
|         |          | 3 = slightly boring  
|         |          | 4 = neutral  
|         |          | 5 = slightly interesting  
|         |          | 6 = moderately interesting  
|         |          | 7 = extremely interesting |
| 40      | SGSEN    | Standard Gamble Evaluation for Sensibility:  
|         |          | 1 = extremely senseless  
|         |          | 2 = moderately senseless  
|         |          | 3 = slightly senseless  
|         |          | 4 = neutral  
|         |          | 5 = slightly sensible  
|         |          | 6 = moderately sensible  
|         |          | 7 = extremely sensible |

Note: reverse coded already
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| 41      | SGDDO    | Standard Gamble Evaluation for Difficulty in Doing:  
1 = extremely difficult to do  
2 = moderately difficult to do  
3 = slightly difficult to do  
4 = neutral  
5 = slightly easy to do  
6 = moderately easy to do  
7 = extremely easy to do |
| 42      | SGUSE    | Standard Gamble Evaluation for Usefulness:  
1 = extremely useless  
2 = moderately useless  
3 = slightly useless  
4 = neutral  
5 = slightly useful  
6 = moderately useful  
7 = extremely useful |
| 43      | SGLEN    | Standard Gamble Evaluation for Length:  
1 = extremely long  
2 = moderately long  
3 = slightly long  
4 = neutral  
5 = slightly short  
6 = moderately short  
7 = extremely short |
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>SGSUC</td>
<td>Standard Gamble Evaluation for Success:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = extremely unsuccessful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = moderately unsuccessful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = slightly unsuccessful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = neutral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = slightly successful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = moderately successful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 = extremely successful</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: reverse coded already</td>
</tr>
<tr>
<td>45</td>
<td>SGCLR</td>
<td>Standard Gamble Evaluation for Clarity:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = extremely vague</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = moderately vague</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = slightly vague</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 = neutral</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5 = slightly clear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 = moderately clear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7 = extremely clear</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: reverse coded already</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>----------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| 46      | SGDUN    | Standard Gamble Evaluation for Difficulty in Understanding:  
|         |          | 1 = extremely difficult to understand  
|         |          | 2 = moderately difficult to understand  
|         |          | 3 = slightly difficult to understand  
|         |          | 4 = neutral  
|         |          | 5 = slightly easy to understand  
|         |          | 6 = moderately easy to understand  
|         |          | 7 = extremely easy to understand  
| 47      | SGCONF   | Standard Gamble Evaluation for Confidence:  
|         |          | 1 = very unconfident  
|         |          | 2 = somewhat unconfident  
|         |          | 3 = neither confident or unconfident  
|         |          | 4 = somewhat confident  
|         |          | 5 = very confident  
| 48 - 49 | SGETIME  | Time to complete Standard Gamble Evaluation  
<p>|         |          | (in minutes)  |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line # 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 - 2</td>
<td>CBCTIME</td>
<td>Time to Complete CBC Task (in minutes)</td>
</tr>
</tbody>
</table>
| 3       | CBCINT   | CBC Evaluation for Interest:  
|         |          | 1 = extremely boring  
|         |          | 2 = moderately boring  
|         |          | 3 = slightly boring  
|         |          | 4 = neutral  
|         |          | 5 = slightly interesting  
|         |          | 6 = moderately interesting  
|         |          | 7 = extremely interesting |
| 4       | CBCSEN   | CBC Evaluation for Sensibility:  
|         |          | 1 = extremely senseless  
|         |          | 2 = moderately senseless  
|         |          | 3 = slightly senseless  
|         |          | 4 = neutral  
|         |          | 5 = slightly sensible  
|         |          | 6 = moderately sensible  
|         |          | 7 = extremely sensible |
| Note: reverse coded already | | |
| 5       | CBCDDO   | CBC Evaluation for Difficulty in Doing:  
|         |          | 1 = extremely difficult to do  
|         |          | 2 = moderately difficult to do  
|         |          | 3 = slightly difficult to do  
|         |          | 4 = neutral  
|         |          | 5 = slightly easy to do  
|         |          | 6 = moderately easy to do  
<p>|         |          | 7 = extremely easy to do |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| 6       | CBCUSE   | CBC Evaluation for Usefulness:  
|         |          | 1 = extremely useless  
|         |          | 2 = moderately useless  
|         |          | 3 = slightly useless  
|         |          | 4 = neutral  
|         |          | 5 = slightly useful  
|         |          | 6 = moderately useful  
|         |          | 7 = extremely useful  
|         |          | Note: reverse coded already |
| 7       | CBCLEN   | CBC Evaluation for Length:  
|         |          | 1 = extremely long  
|         |          | 2 = moderately long  
|         |          | 3 = slightly long  
|         |          | 4 = neutral  
|         |          | 5 = slightly short  
|         |          | 6 = moderately short  
|         |          | 7 = extremely short  
| 8       | CBCSUC   | CBC Evaluation for Success:  
|         |          | 1 = extremely unsuccessful  
|         |          | 2 = moderately unsuccessful  
|         |          | 3 = slightly unsuccessful  
|         |          | 4 = neutral  
|         |          | 5 = slightly successful  
|         |          | 6 = moderately successful  
|         |          | 7 = extremely successful  
<p>|         |          | Note: reverse coded already |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| 9       | CBCCLR   | CBC Evaluation for Clarity:  
 1 = extremely vague  
 2 = moderately vague  
 3 = slightly vague  
 4 = neutral  
 5 = slightly clear  
 6 = moderately clear  
 7 = extremely clear  
  
Note: reverse coded already |
| 10      | CBCDUN   | CBC Evaluation for Difficulty in Understanding:  
 1 = extremely difficult to understand  
 2 = moderately difficult to understand  
 3 = slightly difficult to understand  
 4 = neutral  
 5 = slightly easy to understand  
 6 = moderately easy to understand  
 7 = extremely easy to understand  
  
| 11      | CBCCONF  | CBC Evaluation for Confidence:  
 1 = very unconfident  
 2 = somewhat unconfident  
 3 = neither confident or unconfident  
 4 = somewhat confident  
 5 = very confident |
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 - 13</td>
<td>CBCETIME</td>
<td>Time to complete CBC Evaluation (in minutes)</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>COMPINT1</td>
<td>Comparison of Tasks on Interest for Most Interesting:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>2</td>
<td>COMPINT2</td>
<td>Comparison of Tasks on Interest for Middle Interesting:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>3</td>
<td>COMPINT3</td>
<td>Comparison of Tasks on Interest for Least Interesting:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>4</td>
<td>COMPSEN1</td>
<td>Comparison of Tasks on Sensibility for Most Sensible:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: reverse coded already</td>
</tr>
<tr>
<td>5</td>
<td>COMPSEN2</td>
<td>Comparison of Tasks on Sensibility for Middle Sensible:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| 6       | COMPSEN3   | Comparison of Tasks on Sensibility for Least Sensible:  
         |            | 1 = feeling thermometer  
         |            | 2 = standard gamble  
         |            | 3 = choice-based conjoint  
         |            | Note: reverse coded already |
| 7       | COMPDDO1   | Comparison of Tasks on Difficulty in Doing for Least Difficult to Do:  
         |            | 1 = feeling thermometer  
         |            | 2 = standard gamble  
         |            | 3 = choice-based conjoint  
         |            | Note: reverse coded already |
| 8       | COMPDDO2   | Comparison of Tasks on Difficulty in Doing for Middle Difficult to Do:  
         |            | 1 = feeling thermometer  
         |            | 2 = standard gamble  
         |            | 3 = choice-based conjoint  |
| 9       | COMPDDO3   | Comparison of Tasks on Difficulty in Doing for Most Difficult to Do:  
         |            | 1 = feeling thermometer  
         |            | 2 = standard gamble  
         |            | 3 = choice-based conjoint  
<pre><code>     |            | Note: reverse coded already |
</code></pre>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>COMPUSE1</td>
<td>Comparison of Tasks on Usefulness for Most Useful:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>11</td>
<td>COMPUSE2</td>
<td>Comparison of Tasks on Usefulness for Middle Useful:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>12</td>
<td>COMPUSE3</td>
<td>Comparison of Tasks on Usefulness for Least Useful:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>13</td>
<td>COMPLEN1</td>
<td>Comparison of Tasks on Length for Least Long:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: reverse coded already</td>
</tr>
<tr>
<td>14</td>
<td>COMPLEN2</td>
<td>Comparison of Tasks on Length for Middle Long:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>15</td>
<td>COMPLEN3</td>
<td>Comparison of Tasks on Length for Most Long:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: reverse coded already</td>
</tr>
<tr>
<td>16</td>
<td>COMPSUC1</td>
<td>Comparison of Tasks on Success for Most Successful:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>17</td>
<td>COMPSUC2</td>
<td>Comparison of Tasks on Success for Middle Successful:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>18</td>
<td>COMPSUC3</td>
<td>Comparison of Tasks on Success for Least Successful:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>19</td>
<td>COMPCLR1</td>
<td>Comparison of Tasks on Clarity for Most Clear:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>--------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>20</td>
<td>COMPCLR2</td>
<td>Comparison of Tasks on Clarity for Middle Clear:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>21</td>
<td>COMPCLR3</td>
<td>Comparison of Tasks on Clarity for Least Clear:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>22</td>
<td>COMPDUN1</td>
<td>Comparison of Tasks on Difficulty in Understanding for Least Difficult to Understand:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: reverse coded already</td>
</tr>
<tr>
<td>23</td>
<td>COMPDUN2</td>
<td>Comparison of Tasks on Difficulty in Understanding for Middle Difficult to Understand:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = feeling thermometer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = standard gamble</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 = choice-based conjoint</td>
</tr>
<tr>
<td>Columns</td>
<td>Variable</td>
<td>Explanation</td>
</tr>
<tr>
<td>---------</td>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| 24      | COMPDUN3     | Comparison of Tasks on Difficulty in Understanding for Most Difficult to Understand:  
|         |              | 1 = feeling thermometer  
|         |              | 2 = standard gamble  
|         |              | 3 = choice-based conjoint  
|         |              | Note: reverse coded already                                                  |
| 25      | COMPCON1     | Comparison of Tasks on Confidence for Most Confident:  
|         |              | 1 = feeling thermometer  
|         |              | 2 = standard gamble  
|         |              | 3 = choice-based conjoint                                                   |
| 26      | COMPCON2     | Comparison of Tasks on Confidence for Middle Confident:  
|         |              | 1 = feeling thermometer  
|         |              | 2 = standard gamble  
|         |              | 3 = choice-based conjoint                                                   |
| 27      | COMPCON3     | Comparison of Tasks on Confidence for Least Confident:  
|         |              | 1 = feeling thermometer  
|         |              | 2 = standard gamble  
|         |              | 3 = choice-based conjoint                                                   |
| 28      | COMPOVR1     | Overall Comparison of Tasks for Best Overall:  
|         |              | 1 = feeling thermometer  
|         |              | 2 = standard gamble  
<p>|         |              | 3 = choice-based conjoint                                                   |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| 29      | COMPOVR2 | Overall Comparison of Tasks for Middle Overall:  
|         |          | 1 = feeling thermometer  
|         |          | 2 = standard gamble  
|         |          | 3 = choice-based conjoint |
| 30      | COMPOVR3 | Overall Comparison of Tasks for Worst Overall:  
|         |          | 1 = feeling thermometer  
|         |          | 2 = standard gamble  
|         |          | 3 = choice-based conjoint |
| 31      | OTHER1   | Other Aspects of Health that Need to be Considered # 1:  
|         |          | 1 = cognitive functioning  
|         |          | 2 = sexual functioning  
|         |          | 3 = energy/vitality  
|         |          | 4 = dexterity  
|         |          | 5 = none |
| 32      | OTHER2   | Other Aspects of Health that Need to be Considered # 2:  
|         |          | 1 = cognitive functioning  
|         |          | 2 = sexual functioning  
|         |          | 3 = energy/vitality  
|         |          | 4 = dexterity  
|         |          | 5 = none |
| 33      | OTHER3   | Other Aspects of Health that Need to be Considered # 3:  
|         |          | 1 = cognitive functioning  
|         |          | 2 = sexual functioning  
|         |          | 3 = energy/vitality  
|         |          | 4 = dexterity  
<p>|         |          | 5 = none |</p>
<table>
<thead>
<tr>
<th>Columns</th>
<th>Variable</th>
<th>Explanation</th>
</tr>
</thead>
</table>
| 34      | OTHER4   | Other Aspects of Health that Need to be Considered # 4:  
1 = cognitive functioning  
2 = sexual functioning  
3 = energy/vitality  
4 = dexterity  
5 = none |
| 35      | OTHER5   | Other Aspects of Health that Need to be Considered # 5:  
1 = cognitive functioning  
2 = sexual functioning  
3 = energy/vitality  
4 = dexterity  
5 = none |
| 36      | OTHCOMM  | Other Comments:  
1 = very interesting  
2 = cards don't make sense  
3 = fruitless endeavor  
4 = many variables/confusing  
5 = none |
| 37 - 38 | OVRTIME  | Time to complete Overall Evaluation (in minutes) |
BIBLIOGRAPHY


100. Berg RL. (1973), Health Status Indexes. Hospital Research and Educational Trust, Chicago, IL.


185. Browne MW. (1992). *FITMOD: A Computer Program for Calculating Point and Interval Estimates of Fit Measures.* The Ohio State University, Columbus, OH.


290. Davies-Avery A and Ware JE. (1981), Measuring Health Perceptions in the Health Insurance Experiment. R-2711-HHS; The RAND Corporation, Santa Monica, CA.


318. Dohrenwend BS, Dohrenwend BP, and Cook D. (1973), "Ability and Disability in Role Functioning in Psychiatric Patient and Non-Patient


334. Dupuit J. (1934), *De l'utilite et De Sa Mesure.* La Riforma Sociale, Torino, France.


385. Epstein AM, Hall JA, Tognetti J, et al. (1989), "Using Proxies to Evaluate Quality of Life: Can They Provide Valid Information About Patients' Health Status and Satisfaction with Medical Care?" Medical Care 27 (3, Supplement): S91-S98.


1316


*Health Psychology* 8 (6): 723-735.

Experiment." *Medical Care* 32 (10): 975-988.

711. Kaplan RW and Saccuzzo DP. (1982), *Psychological Testing: 

712. Kaplan S. (1987), "Patient Reports of Health Status as Predictors of 
Physiologic Health Measures in Chronic Disease." *Journal of Chronic 
Diseases* 40 (Supplement 1): 27S-35S.

Observability of Preferences by Experimental Methods." *Econometrica* 

Decision Making Under Risk." In: Munier BR (ed). *Risk, Decision, and 
Rationality.* D. Reidel Publishing Company, Dordrecht, The 

715. Karnofsky DA and Burchenal JH. (1949), "The Clinical Evaluation of 
Chemotherapeutic Agents in Cancer." In: MacLeod CM (ed). *Evaluation of 

716. Kasper JF, Mulley AG, and Wennberg JE. (1992), "Developing Shared 
Decision-making Programs to Improve the Quality of Health Care." 

Responsiveness of Utility Measures." *Medical Care* 32 (2): 183-188.


750. Knight FH. (1921), Risk, Uncertainty, and Profit. Houghton Mifflin, Boston, MA.


of the American Pharmaceutical Association, Dallas, TX; March 20-24, 1993.


1072. Pihlblad CT and Rosencranz HA. (1967), The Health of Older People in the Small Town. Department of Sociology, University of Missouri, Columbia, MI.
1073. Pihlblad CT and Rosencranz HA. (1969), *Social Adjustment of Older People in the Small Town*. Department of Sociology, University of Missouri, Columbia, MI.


1092. Raiffa H. (1968b), Preferences for Multiattribute Alternatives. The RAND Corporation, Santa Monica, CA.


1131. Romano J. (1950), "Basic Orientation and Education of the Medical Student." JAMA 143 (June 3): 409-412.


1328. Tsevat J, Dawson NV, and Matchar DB. (1990), "Assessing Quality of Life and Preferences in the Seriously Ill Using Utility Theory." *Journal of Clinical Epidemiology* 43 (Supplement): 73S-77S.


1409. Wickens CD. (1984), *Engineering Psychology and Human Performance*. Merrill, Columbus, OH.


