AN ANALYSIS OF MULTI-ATTRIBUTE UTILITY THEORY AS A MODEL OF INTERNAL CONTROL EVALUATIONS BY EXTERNAL AUDITORS

DISSERTATION

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

By

Timothy Alan Farmer, B.S., M.A.

* * * * *

The Ohio State University
1983

Reading Committee:
Dr. Jesse F. Dillard
Dr. Daniel L. Jensen
Dr. Richard J. Jagacinski
Dr. George L. Smith, Jr.

Approved By

[Signature]
Adviser

Faculty of Accounting
Copyright by Timothy A. Farmer
ACKNOWLEDGEMENTS

This document represents the final phase of the long process of earning a doctorate. Throughout the process there were people whose help was instrumental in making the completion of this work a reality. I was the beneficiary of support from so many people that it isn't feasible to recognize all those who made contributions to my graduate education, but several people played such significant roles in my education that I must individually recognize them.

To the members of the auditing profession who acted as subjects for the dissertation experiment I extend most sincere thanks. The importance of these people to my study is self-evident. Special thanks go to Mark Osterheld, CPA of Peat, Marwick, Mitchell & Co. whose encouragement, expertise, and support made interface between the academic and professional environments much easier than I anticipated.

To my colleagues in the Accounting Doctoral Program at C.S.U. I extend my warmest appreciation for providing help, support, and some genuinely good times. Particular thanks go to Anthony Pico, CPA for his help in designing and testing the experimental cases.

- ii -
To my classmates, Professors Joseph H. Anthony and Richard A. Young, I extend deep appreciation. My association with them and their wives has made my life fuller and my tenure in graduate school easier than they otherwise would have been.

The guidance, help, and talents of my dissertation committee, Professors Richard J. Jagacinski, Daniel I. Jensen, and George L. Smith, Jr. were vital to the completion of this work. For their help I am truly grateful.

The generosity of some people is beyond measure. Such is the case of my committee chairman, Professor Jesse F. Dillard. Besides providing the guidance needed to complete this dissertation, Professor Dillard provided friendship of the truest kind. No words can adequately express my appreciation.

To my parents, Floyd H. Farmer, Sr. and Doris E. Farmer, I extend my deepest gratitude. Their encouragement and moral support through my many years in school has value which cannot be measured. Had I not felt their love and their concern for education throughout my life, I would have never even attempted to pursue an advanced degree. For reasons too numerous to mention, I dedicate this work to my Mom and Dad.
VITA


1979 . . . . . . . . B.S. in Accounting,
Lake Superior State College,
Sault Ste. Marie, Michigan

1979-1983. . . . . Graduate Teaching Associate
Academic Faculty of Accounting
The Ohio State University,
Columbus, Ohio

1982 . . . . . . . . M.A., The Ohio State University,
Columbus, Ohio

FIELDS OF STUDY

Major Field: Accounting

Financial Accounting. Professor John K. Shank

Managerial Accounting. Professor Daniel L. Jensen

Interdisciplinary Studies. Professor Thomas J. Burns
and Professor James C. Kimard

Minor Fields:

Industrial and Systems Engineering.
Professor George L. Smith, Jr.

Statistics. Professor Jean Powers
# TABLE OF CONTENTS

**ACKNOWLEDGEMENTS** ........................................... ii

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

**LIST OF TABLES** ................................................ viii

**LIST OF FIGURES** ................................................ ix

**Chapter**      | **Page**
---|---
I. **INTRODUCTION** ........................................... 1

Motivation for the Study ........................................ 1

Utility Theory as a Model of Auditor Decision-Making ............ 7

Research Hypotheses ............................................. 15

II. **LITERATURE REVIEW** ....................................... 17

MAUT and Its Application to Non-Accounting Problems ............ 17

Studies on the Elements of MAUT ................................ 18

Non-Accounting Applications of MAUT ........................... 24

Related Studies in Accounting Contexts .......................... 29

Experimental Studies in Auditor Judgment ........................ 29

Accounting Studies of Multiple Criteria Decision-Making ........ 42

III. **RESEARCH METHODOLOGY** ................................... 48

Overview of the Study ............................................. 48

Subjects .................................................................. 49

Elicitation of Multi-Attribute Utility Functions ............... 51

Definition of Attributes .......................................... 53

Tests for Preferential Independence .............................. 56

Tests for Utility Independence ................................... 60

Elicitation of Unidimensional Utility Functions ............... 64

Determination of Individual Scaling Constants ................. 67

Checks for Consistency ........................................... 70
<table>
<thead>
<tr>
<th>Appendix</th>
<th>page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. INTERNAL CONTROL EVALUATION POLICY OF A NATIONAL C.P.A. FIRM</td>
<td>135</td>
</tr>
<tr>
<td>B. FIRM MODEL OF INTERNAL CONTROL EVALUATION</td>
<td>139</td>
</tr>
<tr>
<td>C. INSTRUMENT TO TEST PREFERENTIAL INDEPENDENCE</td>
<td>144</td>
</tr>
<tr>
<td>D. CASE MATERIALS FOR EXPERIMENTAL TASK - PHASE II</td>
<td>151</td>
</tr>
<tr>
<td>E. SUBJECT DEMOGRAPHICS QUESTIONNAIRE</td>
<td>154</td>
</tr>
<tr>
<td>F. DOUBLE CROSS-VALIDATION AND ANALYSIS</td>
<td>155</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>158</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Random Sample of 42 Compliance Test Results</td>
<td>79</td>
</tr>
<tr>
<td>2.</td>
<td>Correlations Among Factors in Case Sample</td>
<td>80</td>
</tr>
<tr>
<td>3.</td>
<td>Conditional Utilities and Scaling Constants</td>
<td>92</td>
</tr>
<tr>
<td>4.</td>
<td>Kendall's tau Between Subjective and MAUT Ratings</td>
<td>94</td>
</tr>
<tr>
<td>5.</td>
<td>Correlations of Subject Ratings With Three Utility Models</td>
<td>96</td>
</tr>
<tr>
<td>6.</td>
<td>Sum of Square Deviations-Cross Validation Regressions</td>
<td>156</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Attribute-Independence and MAUF Form Relationships</td>
<td>12</td>
</tr>
<tr>
<td>2.</td>
<td>Allais Paradox</td>
<td>38</td>
</tr>
<tr>
<td>3.</td>
<td>Flowchart of MAUF Elicitation Procedures</td>
<td>52</td>
</tr>
</tbody>
</table>
Chapter I

INTRODUCTION

1.1 MOTIVATION FOR THE STUDY

Research in cognitive psychology has focused on many factors of human thought. Learning, memory, reasoning, problem-solving, and information processing have been studied in attempts to better understand the human thinking process: its properties, abilities, and limitations. Knowledge gained through research in these areas can be applied to memory, learning, and problem-solving tasks.

The research reported here examines the use of multi-attribute utility theory (MAUT) as a model of the decision-making behavior of external auditors in making internal control evaluations. Acquiring insight into the internalized methods of processing and aggregating internal control data used by auditors will lead to a clearer focus on problems encountered by these professionals in making their evaluations, the causes for the difficulties, and potential solutions to the problems. Research in this area will, hopefully, aid in overcoming quality control problems of
concern to accounting firms, their clients, and professional and regulatory bodies.

The American Institute of Certified Public Accountants (AICPA) (1982) defines accounting control as:

"...the plan of organization and the procedures and records that are concerned with the safeguarding of assets and the reliability of financial records and consequently are designed to provide reasonable assurance that:

a. Transactions are executed in accordance with management's general or specific authorization.

b. Transactions are recorded as necessary (1) to permit preparation of financial statements in conformity with generally accepted accounting principles or any other criteria applicable to such statements and (2) to maintain accountability of assets.

c. Access to assets is permitted only in accordance with management's authorization.

d. The recorded accountability for assets is compared with the existing assets at reasonable intervals and appropriate action is taken with respect to any differences." (Section 320.28)

Auditors use the study and evaluation of the client's system of internal control to perform the following three functions:

1) Determine whether an audit is possible. "If the system is inadequate or non-existent, it is virtually impossible for the auditor to evaluate whether the financial statements are fairly presented...If the internal control is so inadequate as to prohibit basic reliance, the auditor must either refuse to conduct an audit or disclaim an
opinion on the financial statements." (Arens and Loebbecke, 1980, p. 214)

2) Determine the audit evidence that must be collected. This second function addresses the standard of audit fieldwork prescribed by the AICPA: "There is to be a proper study and evaluation of the existing internal control as a basis for reliance thereon and for the determination of the resultant extent of the tests to which auditing procedures are to be restricted." (AICPA, Section 150.02, 1982)

3) Inform the senior management and the Board of Directors of serious weaknesses of the system (Arens and Loebbecke, 1980).

The evaluation of the internal control system presently requires the auditor's subjective judgment on the reliability of the system's components to detect material errors. Review procedures include preliminary evaluation of strengths and weaknesses, compliance tests for each critical control and re-evaluation of the system after testing has been performed. The extent of the substantive tests needed to verify the client's financial statements is determined by the compliance test results.

The 1977 Foreign Corrupt Practices Act called for greater attention to internal controls and their evaluation on the part of the auditors and management. This act required corporations under its jurisdiction to implement
systems of internal control and external auditors became involved in helping clients design and implement these systems. In response to the 1977 law, firms of Certified Public Accountants (C.P.A.) issued discussion papers stating their suggestions for management compliance with the law and the auditing firms' practices in evaluating and monitoring the internal control systems instituted by corporate management (Ernst and Ernst, 1978; Peat, Marwick, Mitchell & Co., 1978; Price Waterhouse, 1979; Deloitte, Haskins, and Sells, 1979). The descriptions of the internal control evaluation processes of these firms indicated considerable diversity within the "Big Eight" accounting firms. This finding was important to the research design due to the desire to create a generic experimental task for subjects. Research findings have indicated that auditor judgments are often widely varied across auditors, and within auditors over time (Ashton, 1974; Ashton and Brown, 1980). Variability in judgments and differences in procedures may have audit quality control implications. While the differences in procedures employed by various firms may have quality control implications, such an impact is not addressed in this study.

A concern over quality control in the auditing profession has been expressed by members of the profession as well as regulatory bodies. The AICPA supports voluntary
quality control and practice review programs and the Securities and Exchange Commission (SEC) in Accounting Series Release (ASR) #153 has established legal requirements of the professional for quality control of audits.

Subjective judgments by auditors have been found to be deficient in terms of judgments expected from professionals as they lack consensus among auditors, often suffer from instability over time, and may be based on heuristics which may bias decisions (Holstrum, 1981). Holstrum stated these characteristics of auditor judgment have importance relative to setting quality control standards for audit effectiveness and efficiency. Lack of consensus across auditors indicates receipt of different signals about the reasonableness of the financial statements (or reliability of the internal control system) across subjects or different interpretation of the signals received.

Loehbecke (1981) cited an unpublished paper by Cushing which dealt with the question of whether performance of audits by audit teams served to control conflicts between the utility functions of individual auditors and that of the audit team. The potential difference in utility functions of auditors may be a reasonable explanation for the lack of consensus among auditors since risk, which is a major consideration in audit judgments, would be reflected in the utility functions. As stated by Holstrum (1981), audit
judgment entails evaluating the joint risk of not detecting errors which have not been exposed by the internal control system, the auditor's test of detail, and the analytical review procedures.

The concept of risk for an individual auditor extends beyond the legal ramifications on the auditing firm from failure to detect material error, to personal risks associated with non-detection of errors. Such personal consequences may include loss of job or loss of status in the firm or the profession. The auditor's perception of the risks involved in his evaluation of internal control, whatever the source, is the "definition" of risk followed in the research. Individual utility functions reflect a person's affinity or aversity to perceived risk. Consequently, auditors with differing levels of risk tolerance or perceived risk may make different decisions given the same attributes of a system are observed by all the auditors.

The quality control concerns cited might eventually be addressed by the use of MAUT as a decision aid in evaluating internal control systems by providing a rigorous model for this task. Internal control judgments made by auditors are the result of combining the assessments of various attributes of the system under review. The application of MAUT incorporates these multiple criteria with the auditor's subjective evaluation of risk. The focus of this
study, however, is not on the construction of decision aids, but rather testing the validity of the MAUT model in an auditing context.

The audit setting to which MAUT is applied by this dissertation is the preliminary internal control evaluation of the Sales/Collections cycle. Within this context the MAUT model's ability to predict and describe decision-making behavior is evaluated. The statistical significance of correlations between MAUT-derived ratings of internal control systems and auditors' subjective evaluations are examined to assess the predictive ability of the MAUT model. To determine the relative merit of the MAUT model, the subject-prescribed scores of reliability are correlated with predictions of these ratings made with simpler models. The correlations from the various models are used as data in a nonparametric statistical analysis of the relative predictive performance of the models.

1.2 Utility Theory as a Model of Auditor Decision-Making

In the economics of consumer behavior, preferences are taken to characterize the choices made by a "rational man" faced with deciding how much he ought to spend on various commodities (Fishburn, 1968). From preference statements or behavior observed, a "utility function" may be described. Utility functions of this nature have been used a general
models of preference behavior in psychological studies as well as studies in other behavioral and social sciences.

The concept of utility used in consumer economics is not based on probabilistic notions. "In the classical economist's view, cardinal utility refers only to the intrinsic pleasure of the attribute under conditions of certainty." (Schoemaker, 1980) Underlying MAUT is unidimensional utility theory in the von Neumann-Morgenstern (1953) sense. This theory combines the concepts of expected value with a decision maker's affinity or aversion to risk and follows a strict set of axioms which are not incorporated into "utility functions" in the sense this term is used in consumer economics. Schoemaker (1980) describes these axioms informally as:

1. The complete ordering axiom: For any two lotteries, L1 and L2, the decision maker prefers either L1 to L2 or L2 to L1, or else is indifferent. Furthermore, if L1 is preferred to L2 and L2 to a lottery L3, then L1 must also be preferred to L3 (called transitivity).

2. The continuity axiom: If $x$ is preferred to $y$ and $y$ to $z$, then there must exist some probability $p$ (between 0 and 1) so that the decision maker is indifferent between a sure amount $y$ and a lottery offering $x$ and $z$ with probabilities $p$ and $(1-p)$ respectively.

3. The independence axiom: If the decision maker is indifferent between alternatives $x$ and $y$, then he should also be indifferent between two lotteries offering $x$ and $z$ in the first lottery and $y$ and $z$ in the second, with probabilities $p$ and $(1-p)$ in each lottery for any $y$ and $p$ values.
4. "The unequal probability axiom: If \( x \) is preferred to \( y \), then lottery \( L_1 \) should be preferred to \( L_2 \) when both lotteries contain only the outcomes \( x \) and \( y \) and when the probability of winning \( x \) is greater in \( L_1 \) than in \( L_2 \).

5. "The axiom of complexity: If two lotteries, \( L_1 \) and \( L_2 \), offer outcomes \( x \) and \( y \) for \( L_1 \) and produce two new lotteries, \( L_3 \) and \( L_4 \), as the outcomes for lottery \( L_2 \), with \( L_3 \) and \( L_4 \) offering only \( x \) and \( y \), then the decision maker should be indifferent between \( L_1 \) and \( L_2 \) if, and only if, the expected values of \( L_1 \) and \( L_2 \) are identical. (This axiom guarantees that the probabilities are calculated in accordance with traditional probability calculus.)" (p. 13-14)

According to the von Neumann-Morgenstern model, a decision maker who conforms to the foregoing axioms should select the alternative which maximizes his/her expected utility. (This is not the same as maximization of expected values since utility incorporates risk tolerances.) A risk neutral decision maker is one for whom \( U(E(x))=E(U(x)) \). A risk averse decision maker is one for whom \( U(E(x))>E(U(x)) \), and risk-seeking decision maker is characterized by \( U(E(x))<E(U(x)) \). That is, a decision maker is risk averse if he/she prefers the expected consequence of a lottery to that lottery. A risk preferent decision maker prefers the lottery over the expected consequence.

Unidimensional utility theory models decision-making with respect to alternatives with a single relevant attribute. For example, if selection of an investment opportunity in production of a new product can be made solely on the basis of cash flows that each alternative will
generate, unidimensional utility theory can be applied to the decision problem.

If individuals consider multiple factors to be relevant to the choice of one alternative over another, they must employ some method for combining the implications of each attribute for overall preference. If the decision of which prospective product to choose requires a decision maker to consider not only cash flows but also manpower requirements and lead-time to implement the decision, there exists more than one relevant attribute associate with each alternative and these attributes may be in conflict with one another. One project may have the most promising expected cash flows, but the time required to tool-up for its production may be longer than the time required to put an alternative on-line. This may be a very important consideration if market conditions required quick implementation to avoid being precluded from gaining a necessary market share. Trade-offs based on these conflicting objectives must be made.

The decision problems of the auditor are also characterized by multiple criteria which may be in conflict. A weakness in one aspect of internal control may be offset by strength in another in terms of making an overall judgment on internal control reliability. The procedures of auditing firms revealed in firm publications provide evidence of
the compensatory nature of the internal control evaluation process.

MAUT provides a formal framework in which to consider conflicting objectives. Whereas unidimensional utility theory focuses on the utility of an alternative \( u(X) \), where alternative \( X \) has one attribute, multi-attribute utility theory focuses on \( u(A) \), where \( A \) denotes an alternative with a vector of attributes. Since a decision maker may have difficulty specifying preferences for alternatives with several attributes, MAUT decompose the alternatives into their individual components, allowing the decision maker to specify preferences in a simplified setting. Multi-attribute utility functions (MAUF) are functions of the decision maker's utility functions for individual attributes. These aggregate functions may be additive, quasi-additive, or multilinear depending on the independence (or lack of independence) among individual attributes. Fischer (1979) provides a flowchart depicting the relationship between attribute-independence and form of the MAUF. This flowchart is reproduced in Figure 1. The starting point for the MAUT flowchart is based on the von Neumann-Morgenstern utility theory axioms stated earlier. Individual utility functions and scaling constants for the aggregation of component utility functions of the MAUF must be elicited by interview with the decision maker.
Tests of independence require the researcher to present the decision maker with combinations of attributes to determine whether a change in one of the attributes affects the implications of the other attributes on the choice problem. To determine the functional form of the MAUF, a test must be made of the decision maker's preferences to determine which of the independence conditions are valid. Three types of independence are defined:
**Preferential Independence:** Attribute X is preferentially independent of its complement, X−, if the preference order of consequences (where no uncertainty is involved) involving only changes in the levels of X does not depend on the levels at which attributes in the complement of X are held fixed (Keeney and Raiffa, 1976).

**Utility Independence:** Attribute X is utility independent of its complement, X−, if the conditional preference order for lotteries involving only changes in the levels of attribute X does not depend on the levels at which the attributes in X− are held fixed (Keeney and Raiffa, 1976).

**Additive Independence:** Attributes X1, X2,...,Xn are additive independent if preferences over lotteries on X1, X2,..., Xn depend only on their marginal probability distributions and not on their joint probability distribution (Keeney and Raiffa, 1976).

The notion of additive independence may be made clearer with an illustration:

Let y' and z' be arbitrarily chosen consequences related to attributes Y and Z respectively. The 50-50 lottery between (y,z) and (y',z') and the 50-50 lottery between (y,z') and (y',z) must be equally preferable in order for y and z to be additive independent since there is a probability equal to .5 associated with obtaining each of y, y', z, and z' in both lotteries.
If the multiple attributes of the decision problem are additive independent, the form of the MAUF is additive:

\[ u(x_1, x_2, \ldots, x_n) = \sum k_i u_i(x_i) \]

where

- \( u_i(x_i) \) is the unidimensional utility function for attribute \( i \) and
- \( k_i \) is a scaling constant elicited from the decision maker.

If \( x \) is utility independent of \( X- \), and if the \( x_i \) (\( \forall i \in \{1\} \)) are preferentially independent of the complement of \( x_i \), denoted \( X-i \), then

\[ u(x_1, x_2, \ldots, x_n) = \left( \frac{TT (k_i u_i(x_i) + 1)}{TT (k_i u_i(x_i) + 1)} - 1 \right) / k \]

usually written in the following form:

\[ ku(x_1, x_2, \ldots, x_n) + 1 = TT (k_i u_i(x_i) + 1) \]

where \( k_i \) and \( u_i(x_i) \) are as defined previously and \( k \) is an overall scaling constant to scale the function on the closed interval between 0 and 1. This is the quasi-additive form of utility function. The additive and quasi-additive functions are special forms of the multiplicative utility function. The general, multilinear form, of which the multiplicative MAUF is a case is as follows:

\[ u(x_1, x_2, \ldots, x_n) = \sum k_i u_i(x_i) - \sum k_i j u_i(x_i) u_j(x_j) + \ldots k_{ijl} \ldots u_i(x_i) u_j(x_j) u_l(x_l) \ldots u_n(x_n). \]

The procedures for eliciting unidimensional (conditional) utility functions and determining scaling constants are discussed in Chapter III, "Research Methodology."
1.3 RESEARCH HYPOTHESES

The emphasis of this study is two-fold. One objective is to test the empirical validity of MAUT in an auditing context. This is accomplished by determining Kendall's coefficient of rank correlation between the output of the MAUT model and the decisions made by practicing auditors. Another focus of the study concerns the ability of Multi-Attribute Utility Theory to model auditor decisions relative to other models. This objective is approached by examining the predictive ability of MAUT as compared to the predictive abilities of two simpler utility models in a nonparametric two-way layout.

The hypothesis of the first objective may be stated:

Ho(1): The ratings of internal control strength given by subjects to experimental cases will be significantly correlated with the ratings derived from the MAUT model.

A significant correlation between model and subjective ratings indicates that a preference ordering derived from the MAUT model closely resembles the ordering established by a subject's ratings. Such a finding provides support for the hypothesis that MAUT is a good model of auditor decision-making since utility theory models decision maker preferences.

The hypothesis of the second objective which concerns the relative predictive ability of MAUT may be stated:
Ho(2): The ratings derived from the MAUT model will be more highly correlated with subject ratings of internal control strength than will the output of simpler utility models which do not incorporate all the elements of MAUT.

If MAUT is superior to other models in terms of predictive ability, there should be a higher degree of rank correlation between MAUT and the subjects' ratings than between the subjects' ratings and other models.
Chapter II
LITERATURE REVIEW

2.1 MAUT AND ITS APPLICATION TO NON-ACCOUNTING PROBLEMS

Utility theory has roots in the literature of classical economic theory, but the axioms of utility theory established by von Neumann and Morgenstern (1953) provide the foundation for multi-attribute utility theory as it is used in this study. The basis for the methodological development of this research is Keeney and Raiffa's (1976) text which gives a detailed presentation of MAUT as it extends from von Neumann and Morgenstern's unidimensional theory. The axioms of utility theory espoused by von Neumann and Morgenstern were presented in Chapter I, "Introduction", and the concepts of MAUT advocated by Keeney and Raiffa which are central to this dissertation will be described in Chapter III, "Research Methodology."
2.1.1 Studies on the Elements of MAUT

The first issue to be examined in this study is the appropriateness of MAUT as a model of subject decision-making behavior. Using statistical and conjoint measurement procedures, Fischer (1976) modelled the decision-making of senior and graduate students in business and engineering to determine whether simple additive and multiplicative utility models were capable of explaining preferences for multi-dimensional outcomes (job choices) for both risky and riskless decisions. Fischer found high within-subject correlation between the rank ordering of alternatives as determined by the risky and riskless assessment models. (Kendall's tau ranged between .69 and .98, with a median of .90.) Similar procedures were followed in this study of auditor preferences in their risky task of evaluating internal control systems.

The procedures for constructing a multi-attribute utility function require subjects to attach subjective probabilities to events which indicate indifference points between the uncertain alternatives presented to them. Probability assessments are made for multiple attribute alternatives as well as for single attribute alternatives. When subjects are not accustomed to thinking consciously in terms of probabilities, they may find this task difficult—which leads to misspecification of the model. Einhorn and
McCoach (1977) and Barron (1980) present modifications of MAUT which attempt to simplify the utility elicitation task and compare the predictions of these models to more complex models. Einhorn (1976) provides an explanation for the results of comparisons between simple and complex models.

Einhorn and McCoach (1977) developed a simple multi-attribute utility procedure (SMAUP) and applied it to evaluation of performance of players in the National Basketball Association. They used Keeney and Raiffa's basic format, but assumed linear utility functions for individual attributes and assumed attribute independence to avoid some of the elicitation techniques required in Keeney and Raiffa's procedure. Attributes were selected by the researchers from statistics kept by the NBA for each player (field goal percentage, free throw percentage, rebounds, etc.). Scaling constants were derived from an aggregation and normalization process performed on the responses to a questionnaire completed by sports journalists concerning the relative importance of each attribute. A function which assumed equal weighting was also tested.

The models were tested by comparing the ratings they produced to the rosters of the NBA all-star teams for the '73-'74 and '74-'75 seasons. Player rankings were predicted quite accurately by the models. The authors found the equally weighted model and the more elaborate weighting
schemes performed similarly. Although SMAUP is useful for evaluations, the authors caution users to examine each particular situation in light of the simplifying assumptions made. If these assumptions are not reasonable, the SMAUP should not be used as a decision aid.

Einhorn (1976) uses the basketball player example of Einhorn and McCoach (1977) to test the sensitivity of ranking procedures to different weighting schemes, focusing particularly on equal weighting. He provides a theoretical justification for the empirical findings: equal weighting can be quite effective in predicting rank orderings. Since rank order of composites is only affected by the relative weights given to attributes and equal weighting cannot reverse the relative weightings of attributes, it avoids errors which might be made in estimating relative weights. Whenever additive models are used for predictive purposes, Einhorn concludes equal weighting schemes are likely to be useful. Einhorn's analysis dealt only with decisions in which ranking is of importance, not the level of the composites, and should be interpreted cautiously.

Barron (1980) developed and evaluated the Holistic Orthogonal Parameter Estimation (HOPE) elicitation procedure which assesses multi-attribute utility functions via sample holistic judgments. The HOPE procedure provides estimates of unidimensional utility functions and scaling
constants for the family of multiplicative utility functions in the risky case and for the family of multiplicative value functions in the riskless case. The procedure requires holistic assessment by the decision maker of a few multi-attributed outcomes. "Holistic responses may be either direct ratings, appropriate for riskless utilities, or standard gamble indifference probabilities appropriate for risky utilities." (Barron, 1980)

For data sets involving a small number of attributes, Barron found the HOPE procedure performed as well as the more familiar decomposition methods. When attention is restricted to non-dominated sets of alternatives the HOPE method was superior to an equal-weights method.

The second phase of the dissertation research focuses on the issue raised by Einhorn (1976), Einhorn and McCoach (1977) and Barron (1980), that of the sensitivity of MAUT to simplifications of its procedures for model-construction. Based on the findings of these studies one would expect that MAUT in its most complete form would model subjects' preferences better than simpler utility models yet all utility models would yield good predictions of observed behavior. This, of course, assumes that subjects actually conform to the utility theory axioms.

Differences in utility functions across decision makers may cause variations in judgments across the subjects.
Studies by Fishburn and Kochenberger (1979) and Crosby, Moskowitz, and Mahesh (1982) provide insights into the nature of utility functions and intra-subject differences in utility functions which may contribute toward understanding differences in MAU functions.

Fishburn and Kochenberger (1979) provided empirical evidence of the shapes of utility functions used in previously published studies. The authors transformed the data of these studies so that the utilities would be defined in terms of changes in wealth or percent return on investment and focused on a "target" or "zero gain" point. The utility functions were fitted in two pieces, on data above and below the target point. Results showed that about two-thirds of the below-target functions were risk seeking and about three-fifths of the above-target functions were concave or risk averse. In general the authors found below-target utility functions were steeper than above-target functions. The authors do not provide an interpretation of these findings relating to differences in subject demographics.

Crosby, Moskowitz, and Mahesh (1982) elicited unidimensional utility functions from seven audit partners and seven audit seniors. They did not seek commonality in scenarios upon which to base utility functions because seniors and partners are confronted with different decision environments in their normal activities. Instead, each auditor
was presented with reference contracts (i.e., lotteries) with which he would be comfortable. Later, utility functions were normalized to create comparability.

The utility functions of the auditors were constructed for monetary gains and losses using lotteries with 50-50 chances. Summed exponential functions were fitted to data elicited from subjects and results were examined to discover differences in risk attitudes of auditors. "Of the 14 subjects tested in this study, two exhibited constant risk proneness, nine exhibited constant or decreasing risk aversion, and three exhibited increasing risk aversion within the gains region. No significant relationships were found between these risk characteristics and the auditors' years of experience, years in a particular position, or his status as a senior or partner." (Crosby, Moskowitz, and Masch, 1982)

While the authors found good fits of the utility functions for data elicited on one dimension, they recognize that auditors often consider multiple factors in making decisions. Crosby, et al. (1982) stated that the nature of multi-attribute utility functions should be investigated in an auditing context.

The research conducted for this dissertation examines multi-attribute utility theory in an auditing context as suggested by these authors. The findings cited here and
the findings of the dissertation research provide insights into the characteristics of MAUT and form the basis for further study into the theory and its applications in specific contexts.

2.1.2 Non-Accounting Applications of MAUT

Multi-attribute utility theory has been used mainly as a tool for analyzing sets of alternative courses of action which have multiple objectives related to them. Although the research undertaken for this dissertation tested MAUT's validity as a model of decision-making in an auditing context rather than using it as a decision aid, the applications of MAUT presented in this section are relevant to this research as they illustrate the variety and types of high-level decision contexts which other authors felt were amenable to the use of MAUT.

Keeney (1973) applied multi-attribute utility theory techniques developed by earlier work to the problem of selecting the most effective strategy for developing the airport facilities of the Mexico City metropolitan area to ensure quality air service for the remainder of this century. The airport effectiveness attributes decided upon by an official of the Mexican Ministry of Public Works included cost, safety, capacity of airport facilities, noise levels, social disruption, and access times as components upon
which to judge the available alternatives with respect to
the interests (objectives) of the government, users, and
the non-users of the facility. The alternatives included
combinations of two possible sites and four categories of
air traffic (domestic, general, military, and internation-
al) which could have been routed to either location. Inde-
pendence among the six attributes was assumed and a multi-
pllicative model of the form:

\[ u(x_1, x_2, \ldots, x_6) = \left( \frac{TI (K_i u_i(x_i) + 1) - 1}{K} \right) \]

where \( K_i = \) scaling factor for individual utility func-
tions

\( K = \) overall scaling factor to unitize the func-
tion

\( u_i(x_i) = \) utility function for attribute \( i \), was fitted
to the individual utility functions and scaling constants
elicited from the Public Works official. From the multi-at-
tribute utility function developed, utility measures for
each airport alternative were derived by evaluating the
function at the component attribute measures for each al-
ternative. These measures provided a means of ranking the
alternatives which allowed explicit consideration of the
relevant factors involved in the decision.

Saaty, Vargas, and Barzilay (1982) developed a deci-
sion aid for high-level decisions which incorporates sub-
jective factors not easily quantified and included in the
decision analysis. They used President Carter's Iran
Hostage Rescue Mission to demonstrate the workings of their procedure called the Analytic Hierarchy Process (AHP). "The novel aspect and major distinction of this approach is that it structures any complex, multiperson, multicriterion, and multiperiod problem hierarchically." (Saaty, et al, 1982)

Measurements used in the methodology are derived from priorities established in pairwise comparisons of relevant elements by the decision maker(s) involved in the task. Relative importance of attributes were determined by normalizing and aggregating the priorities of the decision maker(s). Predictions are made by multiplying a matrix of importance weights for the factors by a priority vector of the alternatives. The "go"/"no go" decision was "made" by observing which alternatives received the highest "composite priority."

The AHP of Saaty, et al (1982) was compared to multi-attribute utility theory by constructing a MAUT model using the same factors as those used in the AHP model. The MAUT model revealed the same decision but the authors cited difficulties with MAUT which could make it harder to implement than AHP. The extra requirements of MAUT cited are:

(1) quantification of all attributes prior to the construction of their utility functions.

(2) evaluation of scaling constants "to assign weights to the attributes according to their importance in the global utility function." (Saaty, et al, 1982, p. 202)
(3) deciding whether to use a multiplicative or additive utility function.

The authors stated that requirement (1) above may pose problems if several decision makers do not agree on the variable that should represent a given attribute.

The statements made by the authors should not be construed as criticisms of the MAUT model. Item (1) above does make MAUT more difficult to implement, but the reason given by the authors is not a valid criticism of MAUT since MAUT is an individual preferences model. The form implemented by Saaty, et al. (1982) is not intended as a model of group decisions.

Statement (2), regarding scaling constants, indicates the authors tried to extend the interpretation of the scaling constants beyond the meaning implied by the theory. (See Keeney and Raiffa, 1976) The scaling constants cannot be interpreted as weights of importance. The measurement of the attributes can render this interpretation completely invalid.

The authors' statement about the difficulty of "deciding" which functional form is appropriate for the decision task indicates that only parts of the theory are being instituted. If all steps in MAUT are followed, the form of the utility function (additive versus multiplicative) follows from the independence conditions of the attributes and the scaling constants elicited from the decision maker. The
analyst really has no decision to make in this regard. Again, this "criticism" of MAUT stems from extending its use as an individual preference model to its use as a group decision aid.

Saaty, Vargas, and Barzilay (1982) did not present AHP as a descriptive model of decision-making, but rather as a decision aid which does not force the decision maker to be consistent with the underlying axioms assumed in MAUT.

Hannan, Smith, and Gilbert (1983) used MAUT and AHP to construct a model aimed at improving the performance of icebreakers in the Great Lakes, western rivers, and polar regions. The decision maker was a Coast Guard officer in headquarters who had previously been an icebreaker operator. He provided the information needed to construct the MAUT and AHP models but was aided by 12 icebreaker operators who provided their own utilities. A method for developing a group utility model proposed by Narsanyi and described by Keeney and Raiffa (1976) was used to combine the 12 utility functions. The group utilities were obtained by assigning equal weights to the utilities of the 12 operators. The AHP was used by the Coast Guard officer to arrive at importance weights for attributes.

By using the "average utility" for 19 ice-breaking devices on 14 attributes given by MAUT models of the 12 icebreaker operators, a preference ranking of the devices
was achieved. The MAUT model was assumed to be appropriate for the decision to be made by the U.S. Coast Guard and thus was used as a decision aid by Hannan, et al (1983). The conclusions of the analysis have been instituted by the U.S. Coast Guard, which attests to the appropriateness of the MAUT model in that context. Empirical testing of the model's validity was not undertaken.

The studies referred to in this section illustrate the potential employment of MAUT (or similar analytical tools) as decision aids. A reasonable extension is to expand the use of the theory into accounting contexts which are characterized by many of the complexities of decision problems to which MAUT has already been applied. Accounting applications of MAUT are few in number and are reviewed in Section 2.2.2.

2.2 RELATED STUDIES IN ACCOUNTING CONTEXTS

2.2.1 Experimental Studies of Auditor Judgment

Ashton (1974) studied the internal control judgment of auditors. His experiment presented questions concerning internal control over payroll for a described company which were pre-answered (yes or no), upon which auditors were supposed to base their evaluations. A six-point scale, "extremely weak" to "adequate to strong" was used to measure the auditors' perceptions of the internal control
strength. The research checked the consistency of auditor judgment over time by having subjects repeat the evaluation task six to thirteen weeks after the initial phase of the experiment. Overall, Ashton found audit judgments to be quite consistent (both with respect to consensus among auditors and stability over time) but individual auditors were sometimes inconsistent in their judgments. Analysis of cue usage showed that all indicators in the cases were considered most important by at least one subject and all were considered least important by at least one subject. This finding provides evidence of variability in auditor judgments.

Joyce (1976) extended Ashton's work but examined the judgment process between the data collection phase and audit planning, whereas Ashton looked at the judgments between data collection and internal control evaluation. Joyce studied auditor decision-making in the context of internal control over accounts receivable rather than payroll as Ashton had done. Both Ashton and Joyce employed correlational statistics to measure consensus and analysis of variance (ANOVA) to model the decision behavior of individual subjects. Multivariate as well as univariate statistics were used to measure consensus and construct judgment models.
Inter-judge reliability was low (average Pearson product moment correlation coefficient was .373), contrary to Ashton's (1974) finding, but was consistent with findings of studies done prior to the work by Asthon. To explain why Ashton found consensus and studies by Joyce and others did not, Joyce stated it is conceivable that different auditors might generally agree on the quality of internal control in a given subsystem yet disagree on how much audit work should be devoted to that subsystder. This is not unreasonable considering the characteristics of sound internal control are well defined, while the basis for selection, timing, and extent of audit procedures to carry out in a specific circumstance may be defined less explicitly.

"The findings of prior auditing research and the great difficulty of acquiring professional expertise in a dynamic, probabilistic environment led to an expectation of high variance among judgments of the auditors participating in the study. The expectation was realized." (Joyce, 1976) Joyce hypothesized the level of consensus he found may be a reflection of the state of the art in auditing—not necessarily a reflection of professional competence. The existence of professional guidelines for auditing does not ensure that interpretation and application of the guidelines are consistent across auditors.
Ashton and Brown (1980) found average values of judgment insight, stability, and consensus were high. Auditors from Big Eight C.P.A. firms who had varying experience levels were asked to make 160 judgments based on eight internal control cues. Prior studies in this area analyzed 32 judgments on five to six cues. The large number of cases was administered to allow for a study of two-way and three-way interaction effects in the ANOVA. Tests of this data showed that these interaction effects were unimportant in explaining the variance of auditor judgments.

Two cues on separation of duties in payroll explained 51.4% of the variance, considerably more than any of the other cues. This study and prior studies suggest both judgment insight and consensus increase with increased levels of audit experience. Combined results also suggest the consistency of judgment policies is greater for more experienced auditors than for less experienced auditors in the study.

Mock and Turner (1979) found a great deal of variability in decisions about the extent of testing to be performed given changes in internal controls from one year to the next. The treatment effect of an analysis performed on the auditors' judgments explained only 7-15% of the variance.
Results showed less variability in auditor judgments when "strong" controls were present than when "fair" controls were in place. Although not stated by the authors, such variability in judgments might be explained in terms of the auditors' utility functions. The level of risk is lower when strong controls are in place than when controls are weak. Uncertainty about the reliability of the financial statements arises due to weaknesses in the internal control system. The auditors' differing degrees of risk aversion may result in varying sample sizes to compensate for the risk posed by weak controls. Since strong controls reduce the amount of risk related to those controls, the effect of differences in degree of risk aversion may be moderated. The extreme case of total elimination of risk would remove the significance of risk tolerance completely.

Mock and Turner (1981) examined the effect of guidance treatments and changes in internal control on sample size recommendations. They found large ranges for these recommendations made by the subjects in the study. Mock and Turner found that variability in subjects' sample size decisions was not consistently reduced by either guidance or internal accounting control treatments. The authors suggested several explanations for the variability of subject decisions. Among the reasons given were variability of the nature of test interpretation, judgments of parameters
which underlie statistical samples (beta risk, alpha risk, materiality, and acceptable amount of overstatement), and demographics (level of experience and specialized training). McCk and Turner cite as the major implication of the variability in sample size recommendations the possible risk of unwarranted reliance on small sample sizes and the risk of excessive audit costs when large sample sizes are specified. From the lack of explained variance of sample sizes in terms of statistically evaluated variables, the study implies many decision variables and decision-making procedures may have been used by the subjects.

Solomon (1982) attempted to assess how well audit teams can specify prior probability distributions (PPDs) for attribute and variable sampling relative to individuals. He also examined the "empirical validity" of account balance PPDs. Extremeness and calibration of account balance PPDs as well as consensus were investigated.

Auditors from Big Eight C.P.A. firms and one other large national firm were involved in a laboratory experiment which was performed a few weeks after the subjects had been trained in probability assessment. Participants were assigned randomly to groups representing three decision-making formats: (1) individual, (2) nominal group followed by an interactive group, and (3) an interactive group followed by a nominal group.
Group 1 read and responded to cases individually without interaction with others. Group 2 first read and responded to cases without group interaction and then subjects were asked to discuss the cases and reach a consensus on the appropriate PPV for each case. Group 3 reversed the steps followed by Group 2.

Solomon (1982) showed that the mean probability assigned by Group 3 was significantly higher than that assigned by individuals (Group 1). Group 2 also assigned a higher probability than that of Group 1, but the difference was not statistically significant.

"Calibration" as defined by Solomon is concerned with the appropriateness of the assessors' confidence in their subjective judgments. "For example, over the long run, an assessor is well calibrated if for all judgments assigned a probability of .75, 75 percent actually occur." (Solomon, 1982, p. 696) Chi-squared goodness of fit tests showed that all three groups were somehow miscalibrated as there were significant differences between observed and expected occurrences.

The author found a conservative bias which he attributed to training. Extremeness, which the author states is the condition of assigning a high probability to events which do not occur, was found to be greater for audit teams than for individuals.
Solomon concludes that there was generally a benefit from interaction but states:

"...the audit team PPDs were more extreme and exhibited greater consensus, but also exhibited greater miscalibration than the PPDs assessed by individuals. Hence, neither approach to assessing PPDs dominated the other." (Solomon, 1982, p. 703)

The findings of Ashton (1974), Ashton and Brown (1980), Joyce (1976), Mock and Turner (1979, 1981), and Solomon (1982) all indicate variability in auditor judgments based on evaluation of hypothetical audit environments. Most of these show a serious lack of consensus among the subjects. The authors generally attempted to determine which factors in the environment caused these differences through analysis of variance. Such testing focuses on statistical explanation of variance by cue usage and cue weightings, but the results do not provide clear reasons why these differences across auditors exist.

This study examines the judgments of auditors in comparison to predictions of those judgments made by utility models tailored to each individual's preference structure. If the models are able to predict accurately, a utility model is a good representation of the auditor's decision-making behavior. Differences in auditors' attitudes toward risk result in differences in utility functions, thereby providing an explanation for differences in judgments found by these authors.
Ashton (1982) provided some initial evidence of auditors' conformance to utility axioms. She studied responses of 414 Big Eight auditors to Allais' problem (used to test for violations of the expected utility axioms) in an auditing context. (See Figure 2) One variation was a problem involving bidding on two municipal audits and the other was a problem about whether to issue a qualified opinion concerning an unresolved matter. Sets of probabilities used in the study included those of the original Allais problem as well as combinations which were not as extreme as the original values.

Ashton found significantly more violations to the original Allais problem than to problems couched in auditing contexts. "...It should be noted that the violation rate in response to the original Allais problem (22 percent) is relatively low in comparison to violation rates in previous studies of the problem (15 to 59 percent). Thus, the results suggest that a relatively small number of auditors will make inconsistent decisions if confronted with situations similar to those suggested by Allais' problem." (Ashton, 1982)

The application of multi-attribute utility theory to internal controls evaluation may be viewed as a continuation of the investigations started by Ashton's examination of the Allais problem in an auditing context. The research
1. Do you prefer situation A1 or B1?

**Situation A1:**
Certainty of receiving $1,000,000.

**Situation B1:**
10 chances in 100 of gaining $5,000,000.
89 chances in 100 of gaining $1,000,000.
1 chance in 100 of gaining nothing.

II. Do you prefer situation A2 or B2?

**Situation A2:**
11 chances in 100 of gaining $1,000,000
89 chances in 100 of gaining nothing.

**Situation B2:**
10 chances in 100 of gaining $5,000,000
90 chances in 100 of gaining nothing.

Choosing A1 and B2 is a violation of utility theory:

A1 preferred to B1 implies:

\[
\begin{align*}
&u(1,000,000) > \\
&(.10)u(5,000,000) + (.89)u(1,000,000) + (.01)u(0) \\
\Rightarrow & (.11)u(1,000,000) > \\
&(.10)u(5,000,000) + (.01)u(0).
\end{align*}
\]

B2 preferred to A2 implies:

\[
\begin{align*}
&(.10)u(5,000,000) + (.90)u(0) > \\
&(.11)u(1,000,000) + (.89)u(0) \\
\Rightarrow & (.10)u(5,000,000) + (.01)u(0) > \\
&(.11)u(1,000,000).
\end{align*}
\]

*Figure 2: Allais Paradox*
performed for this dissertation may also be considered a variation on the prior studies aimed at a different level of understanding the decision-making behavior of auditors.

The hypothesis in this study differs from the results found in prior studies of auditor judgment in that it emphasizes differences in the perceptions of the environment of the audit in general, rather than physical factors of the firm that are evaluated in making a judgment. The study of differences in audit evidence-usage and weighting can provide information useful in understanding factors relied upon in the evaluation decision and will ultimately contribute toward improving audit effectiveness. Studies which reveal personal, internalized data filters, such as risk preferences, may account for differences in amount of attention paid to audit evidence across auditors. The results obtained will constitute a preliminary step in this direction.

Gaumnitz, et al (1982) attempt to tie together the studies previously cited which are dichotomized as research on consensus at the program planning phase or consensus on internal control strength. In the experimental task, the auditor-subjects first made an explicit evaluation of internal control strength and then stated the number of hours required to assess the propriety and collectibility of accounts receivable.
The authors found results of tests of inter-auditor consensus of internal control evaluations to be similar to those found by Ashton (1974) (mean correlation of .70) but their findings of consensus on audit hour requirements conflicted with Joyce's (1976) results. Gaumitz, et al (1982) found a correlation of .617, which is almost twice the correlation of .373 found by Joyce.

An ANOVA was performed to check for differences on the evaluation and hour specification tasks across C.P.A. offices as well as across levels of audit experience. Significant differences were found for other analyses.

The conflict with Joyce's (1976) results is explained by the authors as follows:

"...one possible explanation for our divergent findings is that an explicit evaluation of internal control is required to obtain a high correlation among subjects' hour estimates. That is, we suggest that internal control and audit hour judgments must somehow be 'linked' in order to obtain consensus in planned audit hours. This link is provided by the auditors' knowledge that an inverse relationship should exist between the strength of internal control and the number of audit hours planned." (Gaumitz, et al, 1982, p. 753)

Hamilton and Wright (1982) extend the work of others who studied auditor judgments at various stages in the audit process. These authors studied the relationship between years of experience and judgment consensus, stability of judgments, and the relative weighting of, and the degree of self-insight into the weighting of, internal control
indicators. The authors attempted to check for the generalizability of previous results. The experimental task was similar to that used in the Ashton (1974) study.

These researchers found a significant negative correlation between average consensus and years of experience. Highly significant differences were found for consensus average across firms. There was no significant difference in the stability measure (adjusted coefficients of determination) across levels of experience. By comparing subjectively specified cue-importance weights given by subjects to objectively determined cue weights from an ANOVA, the degree of self-insight was determined. "The results indicate considerable self-insight into the relative weightings of the five internal control indicators." (Hamilton and Wright, 1982, p. 764) More experienced auditors showed a slightly higher degree of self-insight than did the less experienced auditors in the study.

The studies presented in this section provide a motivation for studying MAUT in an auditing context. The authors cited here have contributed to the body of knowledge related to auditor decision-making by highlighting problems faced by auditors and offering some potential causes for judgment inconsistency and inter-auditor differences. In this study, MAUT is shown to be another viable medium for examining auditor decision-making in evaluating internal controls.
2.2.2 Accounting Studies of Multiple Criteria Decision-Making

The use of MAUT in accounting contexts has been very limited. Middaugh (1981) tests the applicability of MAUT as an aid in making capital budgeting decisions and Payne, Laughhun, and Crum (1982) test for violations of the independence axioms of MAUT in a capital budgeting context. A study by Arrington, Jensen, and Takutani (1982) is included with these works as an example of research in multiple criteria decision-making in an accounting context. While the Arrington, et al. article is not a true application of MAUT, it does address the issue of individual preferences as a determinant of decisions in accounting contexts.

Middaugh (1981) explored the potential use of MAUT in capital budgeting. MAUT was chosen as a normative model for capital budgeting because it yielded a functional model of a subject's preference behavior. He found eight out of ten correlations between subject rankings and rankings based on calculated utilities from the derived MAUFs were statistically significant ($p<.1$). The significant correlations ranged from .466 to .936. Middaugh concluded that MAUT does appear to be a viable framework in which to address the capital budgeting problem. From performing a sensitivity analysis of the functional forms of the MAUFs, the author found less time-consuming elicitation procedures to be robust and may be used with little "cost" in terms of
ability to predict. This result is consistent with the findings of Einhorn and McCoach (1977).

Payne, Laughhunn, and Crum (1982) studied the preferences of 128 industrial managers for hypothetical capital budgeting "gambles" in a check for violations of the independence axioms of Multi-Attribute Utility Theory. The results provided mixed support for utility independence, but there was generally strong support for preferential independence (called "event independence" in their study). Several factors limit the generalizability of the results of the Payne, Laughhunn, and Crum study including: (1) a non-random sample, (2) time constraints on subjects, (3) only two attributes upon which subjects were to base preference statements, and (4) the capital budgeting context. The last factor represents the most limiting factor in terms of implications for this dissertation research. A study of capital budgeting using two fairly objective attributes (such as the cash flows used in the Payne, Laughhunn, and Crum study) is much different from a study of auditing procedures based upon subjective attributes. Drawing implications of the Payne, et al. (1982) results for this study is tenuous, but the findings of certain violations and conditions for risk aversity and risk preference are, in general, relevant to work in the area of multi-attribute utility theory. The independence violations found by Payne, Laughhunn, and Crum
indicate that application of MAUT may not be appropriate in some circumstances but do not imply that the theory is generally faulty.

Arrington, Jensen, and Tokutani (1982) considered the use of corporate performance measures by managers and society in general and the means for aggregating multiple variables into an overall measure of corporate performance. The study attempts to scale the perceived importance of corporate multivariate disclosures and to rank companies according to their performance on a composite scale across the multivariate criteria. The method of composition used in the Arrington, et al. study is the AHP (Analytic Hierarchy Process) approach advocated by Saaty and his colleagues. (See Saaty, et al., 1982) AHP, along with utility theory, is part of a general focus of research on multivariate composition.

Arrington, et al. (1982) describe Saaty's AHP method in the context of their study as follows:

**AHP Procedure I: Eigenvector Scaling of Pairwise Elicitations of Perceived Variable Importance**

Pairwise comparisons of attributes are made based on perceived importance. A normalized eigenvector is derived therefrom which is a "scaled human perception of variable importance."

**AHP Procedure II: Scaling of Corporate Performance on Each Variate**

Measurement of variables may be direct (as in the case of variable which have objective measures) or indirect through subjective pairwise comparisons of attributes (e.g., perceived customer service reputation) of the set of companies.
AHP Procedure III: Composition of Multivariate Performance and Importance

An additive compensatory composition of variables is assumed by multiplying the vector of importance weights by the matrix of performance scores for the variables.

The product vector gives an implicit ordering of the utility companies used as examples in this study. The performance ranks are based upon the companies' scores on the following variables:

1. earnings margin rating
2. product price (electricity) to customers
3. research and development investment rating
4. deficiency rating on state-of-the-art pollution control

Arrington, et al. (1982) propose a modification of Saaty's AHP which allows subjects to be "inconsistent" and thereby "create" a model of subject perceptions which may be non-linear and non-compensatory, more in line with reality. The authors accomplish this by substituting pairwise comparisons of observed companies by subjects for the pairwise comparisons of attributes in AHP Procedure II. The comparisons of actual companies provides a subjective composition for scaling multivariate overall performance which may be compensatory or non-compensatory depending on the subject's responses.

Arrington, Jensen, and Tokutani (1982) state that utility theory is aimed at the task of developing a
structural model depicting a human respondent's utility or preference for complete combinations of variates, which is a more difficult task than the one reported by the authors which was aimed only at ordering or scaling alternatives. "It (AHP) does not...yield a utility or preference structural model, as in various approaches...that measure the main and/or interaction effect of the individual variates. Such added parameterization requires considerably more input from human respondents regarding thought processes." (Arrington, et al., 1982, p. 120) Because MAUT yields a model of subject preferences for any given set of measurements for variables which are represented in the model, comparisons between MAUT and AHP are limited. AHP claims only to order (or "scale") a set of alternatives and cannot be validly extended to analyze alternatives other than those for which pairwise comparisons were made in AHP Procedure II. The authors did not intend to uncover elements of the human thinking process and thus did not attempt to validate the model by comparing its output to observed phenomena. Arrington, et al. (1982) intended only to elicit judgments of multivariate pairwise comparisons of corporate social performances.

This study of auditor judgments examines MAUT as a general model of auditor behavior in an internal control setting. The examination of AHP in such a manner would not
be appropriate due to its limitations as a tool for a specific decision.
Chapter III
RESEARCH METHODOLOGY

3.1 OVERVIEW OF THE STUDY

The experiment consisted of two events. Over a period of approximately two weeks in August, 1983, interviews were conducted with each subject to elicit a multi-attribute utility function (MAUF) which models each auditor's preferences for internal control strength. Section 3.2 provides information about the subjects of the experiment and Section 3.3 describes the procedures for MAUF elicitation. A follow-up session with each subject was scheduled approximately two weeks after the initial discussions. During the second session, each participant was presented with a description of a hypothetical company and its internal control system. The text of the case was followed by the results of compliance tests of the transaction controls over the hypothetical company's sales. Forty-two different sets of results were given to subjects for their subjective evaluation of the reliability indicated by each case. It was suggested that such a rating may be accomplished by
first separating the cases into categories of relative strength by assigning scores of 90, 80, 70, and so on, to the results of compliance tests. By rank ordering the cases within each strength class, a complete ordering of the cases could be achieved. Demographic data were also collected during this phase of the experiment. Section 3.4 discusses the findings of a pilot study which tested Keeney and Raiffa's (1976) elicitation procedures and formed the basis for the task design. Section 3.5 describes the experimental task in more detail. Section 3.6 describes the statistical tests of the experimental data.

3.2 SUBJECTS

The firm whose procedures for evaluating internal controls are described in Appendix A provided subjects for the experiment. All participants are employed as auditors in the Columbus, Ohio office. Statistical tests which assume a random sample of subjects are used as tools of analysis even though the subjects of the study cannot be considered a random sample of practicing auditors. The violation of the assumption was deemed appropriate for this study in light of the problems a random sample would create and the fact that the study focuses on individual preferences. The problem of a random sample across firms lies in the existence of widely varied internal control evaluation
procedures among the C.P.A. firms. Hence, the selection of a single firm was made to avoid confounding effects of different policies.

The hypothesis tests conducted on data obtained are not meant to provide the basis for generalizable conclusions, but rather to give insights into the nature of the data from which inferences about the theory can be made. Generalizability of results was traded for tractability of the experimental task in limiting the sample to one public accounting firm. Since multi-attribute utility theory is an economic model of individual behavior, inferences about the validity of the MAUT model for individuals may still be made even though the study does not employ a random sample of auditors.

The fifteen auditors who participated in the study represented a range of experience levels. Experience ranged from a "Senior" with two years experience to a "Senior Manager" with 11 years experience. While the sample sizes at each level are not great enough to draw conclusions with overwhelming statistical support, some preliminary analysis of the effect of experience was accomplished with the results obtained from this limited sample. ¹

¹ The set of subjects consisted of 2 Seniors, 7 Supervising Seniors, 1 Manager and 5 Senior Managers.
3.3 ELICITATION OF MULTI-ATTRIBUTE UTILITY FUNCTIONS

The primary research task entails eliciting a multi-attribute utility function from each auditor-subject. The following subsections described the procedure used during the interview phase of the study. The interviews ranged in duration from 15 minutes to 30 minutes, averaging 24 minutes. An hour was allotted for each subject. Had more time been needed, it would have been available as subjects were not constrained by time. Figure 3 diagrams the steps followed in Phase 1 of the experiment.
Figure 3: Flowchart of MAUF Elicitation Procedures

Ask Subject to Specify Attributes & Ranges

Check for Preferential Independence

Are Attributes Mutually PI? NO

Check for Utility Independence

Are Attributes Mutually UI? NO

The MAUF is Multilinear

YES

The MAUF is Multiplicative

Elicit Conditional Utility Functions

Elicit Scaling Constants

Is $k = 1$? NO

The MAUF is Quasi-Additive

YES

The MAUF is Additive
3.3.1 Definition of Attributes

Keeney and Raiffa (1976) prescribe steps to be taken in eliciting multi-attribute utility functions. The first of these steps is to have subjects define the important objectives of the decision problem they face and the factors which indicate attainment of these objectives. These factors, called "attributes" of the decision problem, are incorporated into the multi-attribute utility function and are consequently problem-specific as well as subject-specific.

Adopting Keeney and Raiffa's procedures, each subject specifies the relevant factors considered in judging the strength of internal control for a specific company. The subject states the questions he/she addresses in making such a decision and tells how the factors are measured and the relevant range for each attribute. This is accomplished by asking: "What things do you look for in evaluating the internal controls over the Sales/Collections cycle or what procedures do you follow?" The subjects are also asked to specify how variables are (or may be) quantified.

Preliminary investigations of professional auditing literature indicated that the "Big Eight" C.P.A. firms have written policies and procedures for making the evaluation of internal controls (Deloitte, Haskins & Sells, 1979;
Ernst and Ernst, 1978; Peat, Marwick, Mitchell and Co., 1978; Price Waterhouse, 1979). The procedures of the firm which participated in this study are very explicit and are summarized in Appendix A. It was expected that subjects would conform to this written policy and respond to questions concerning relevant attributes of an internal control system by quoting firm policy. This expectation was borne out in pilot testing. When asked about the internal control system characteristics that are important in assessing its reliability, the subject of the pilot study's first phase gave a thorough explanation of the firm's procedures, including specification of the relevant attributes. The subject further stated that all subjects "should" give similar responses to questions concerning the relevant attributes of an internal control system.²

During the actual experiment, preparatory questions were asked to determine if subjects deviate from firm policies. The auditors were asked if, during their evaluation of the transaction controls over sales, they considered the characteristics of the system specified on the firm's worksheets. All subjects responded affirmatively and generally

² Based on preliminary discussions and the pilot study, the attributes and measures specified were expected to conform to the firm-model described in Appendix B, which is based on the firm's policies presented in Appendix A. This was in fact the case. When specifically asked if additional attributes were used, the participants did not add any additional factors to the four attributes hypothesized as being the relevant factors.
added a statement conceming the firm's requirement of proper documentation evidenced by completed workpapers. The subjects were asked whether or not any additional variables are important in their own evaluations of an internal control system as these would also be relevant to a specific subject's assessment of reliability and should consequently be incorporated into his/her utility function. A variety of answers were given, but subjects consistently stated that the firm-specified attributes are those which are consciously considered. Other things may be important under certain circumstances, but would not be easily quantified or even consciously considered. The overall control environment affects the evaluation in this manner but is taken into consideration in the sample size and degree of planned reliance on the system, so it really is not an issue by the time the internal controls are actually compliance tested.

The compliance test worksheet used by the participating firm specifies that the any more than two errors from a sample of 40 (a common sample size for the compliance test of controls over sales) will not allow the auditor to place a "moderate" degree of reliance on the controls. It further states that 0, 1, or 2 errors will provide the same degree of reliance. In attempting to determine the range of values for the attributes the subjects were asked if it is possible to find more than two errors (e.g., four errors in
a sample of 40) and still have a system that could be relied upon to the same degree as if only two errors were found. If the subject responds positively to such a question, the range of the attribute values must conform to this broadened interval. In response to this question subjects either stated that they would not want to rely on a system which produced more than two errors or they stated that a system with more than two errors could be reliable but the chance of it occurring would be close to zero.

Discussions with pilot study subjects and preparatory dialogues with subjects of the experiment confirmed the researcher's expectations about internal control attributes and the values with which to measure these cues. Subsequent steps in the utility function elicitation process were based on the policies for internal control evaluation prescribed by the participating firm.

3.3.2 Tests for Preferential Independence

Xi is preferentially independent of Xi- (its complement) if preferences for consequences which differ only in terms of the Xi attribute depend only on the xi value and not on the common value of xi-.

i.e. If (xi', xi-) is preferred to (xi", xi-) for one value of xi-, then (xi', xi-) must be preferred to (xi", xi-) for all values of xi-. (Keeney and Raiffa, 1976)

The subject is asked to evaluate conditional preferences for each attribute. That is, the decision maker is
asked for a preference ordering of the consequences in X given the X- is held at a fixed value. The subject is then asked if the preference ordering of the consequences changes if its complement is changed to a new value. If the subject maintains the same preference ranking on X over the changed value of the complement, then preferential independence is exhibited.

The interview with the auditor-subjects included questions to test for individual attribute preferential independence as follows:

(Researcher - R) Which would you rather see?

\[(x_P, x_A, x_C, x_M) = (2, 1, 1, 1)\]  \hspace{1cm} (1)

or

\[(x_P, x_A, x_C, x_M) = (0, 1, 1, 1)\]  \hspace{1cm} (2)

(Subject - S) I would say "(2)" of course.

(R) What about

\[(x_P, x_A, x_C, x_M) = (2, 1, 2, 1)\]  \hspace{1cm} (3)

or

\[(x_P, x_A, x_C, x_M) = (0, 1, 2, 1)\]  \hspace{1cm} (4)

(S) "(4)" of course.

Technically, this line of questioning should continue for all attributes and all changes in their complements, but subjects generally understand the concept that is being tested and respond with one statement that indicates that the levels of the other attributes would have no effect on their preferences in these situations.
Mutual preferential independence exists if all attributes are preferentially independent of their complements. With the finding of mutual preferential independence and verification that just one attribute is utility independent of its complement, mutual utility independence is assured. (See Keeney and Raiffa, 1976, p. 289) This result reduces the number of questions that must be asked to determine the functional form of the MAUF.

During the initial interviews with subjects the researcher tested single attribute preferential independence and assumed that preferential independence for pairwise comparisons of attributes would exist based on the results of the tests conducted. Tests for utility independence proceeded on the basis of that assumption. Later, pairwise tests were performed on nine of the fifteen subjects. (The six subjects not tested were unavailable for various reasons: two subjects had since left the firm, one subject was on vacation, and three were out of town on audit engagements.)

An example of the questions used to test subjects for pairwise preferential independence follows. The questions actually used are interpretations of the definition of preferential independence given by Keeney and Raiffa (1976) patterned after examples of the test performed during actual consulting projects. The complete instrument for this test is found in Appendix C.
Let:

\[ P = \text{number of errors in a compliance test of Population Completeness} \]

\[ A = \text{number of errors in a compliance test of Authorizations Review} \]

\[ C = \text{number of errors in a compliance test of Comparisons Accuracy} \]

\[ M = \text{number of errors in a compliance test of Mathematical Checks} \]

Assume: \( C = 2 \)

\[ M = 1 \]

For each pair of alternatives below, circle the more favorable compliance test report.

1A:

(a) \( P = 1 \) error, \( A = 2 \) errors
(b) \( P = 1 \) error, \( A = 1 \) error

2A:

(a) \( P = 1 \) error, \( A = 2 \) errors
(b) \( P = 2 \) errors, \( A = 0 \) errors

3A:

(a) \( P = 2 \) errors, \( A = 0 \) errors
(b) \( P = 1 \) error, \( A = 1 \) error

Now, assume: \( C = 1 \)

\[ M = 0 \]

Again, circle the more favorable report in each pair of alternatives below.

1E:

(a) \( P = 1, A = 2 \)
(b) \( P = 1, A = 1 \)

2E:

(a) \( P = 1, A = 2 \)
(b) \( P = 2, A = 0 \)

3E:

(a) \( P = 2, A = 0 \)
(b) \( P = 1, A = 1 \)

Now, Assume: \( C = 1 \)

\[ M = 2 \]
Again, circle the more favorable report in each pair of alternatives below.

1C: (a) \( P = 1, A = 2 \)
    (b) \( P = 1, A = 1 \)

2C: (a) \( P = 1, A = 2 \)
    (b) \( P = 2, A = 0 \)

3C: (a) \( P = 2, A = 0 \)
    (b) \( P = 1, A = 1 \)

This exercise tests for preferential independence of consequences on pairs of Population Completeness (P) and Authorization Reviews (A) from pairs of Comparisons (C) and Mathematical Checks (M). All combinations of pairs of the four attributes were tested.

3.3.3 Tests for Utility Independence

\( X_i \) is utility independent of the other attributes, \( X_i \), if preferences for lotteries over \( X_i \) with other attributes held constant, do not depend on what those fixed levels are. (Keeney, 1973)

Given a vector of attributes \( X = (X_1, X_2, ..., X_n) \) measured by the vector of values \( x = (x_1, x_2, ..., x_n) \) a check is made to determine if \( X_1 \) is utility independent of the complement of \( X_1 \), denoted \( X_1^c \). The decision maker is asked to specify the value of \( p \) which would make him/her indifferent between

\[(x_1', x_2', x_3', ..., x_n')\]

for certain or a \( p/(100-p) \) lottery composed of \( (x_1', x_2', x_3', ..., x_n') \) and \( (x_1^c, x_2', x_3', ..., x_n') \) where

\( x_i' = a \) particular value of attribute \( i \)
\( x_{i0} = \) the worst possible value of attribute \( i \)

\( x_{i*} = \) the best possible value of attribute \( i \).

This question is asked again with different values of \( x_{1-} \), and is repeated until the researcher is convinced that the same \( x_{1-} \) value will be given as an answer regardless of the \( x_{1-} \) value used. This finding indicates that \( x_{1-} \) is utility independent of \( x_{1-} \).

The definition of utility independence given by Keeney and Raiffa (1976) was set in an auditing scenario to test this condition in the dissertation experiment. Keeney and Raiffa provide several examples of the application of this definition in their text which served as models for the instrument used here.

An example of the protocol of this step is:

(Researcher - R) Suppose you know for sure that there is one error in the compliance test for Mathematical accuracy and there are no errors in the other tests.

i.e. \((x_{1'}, x_{2'}, x_{3'}, x_{4'}) = (1, 0, 0, 0)\).

Suppose also that on the other hand, you don't know how many errors for Mathematical checks there are "for sure," but you know that there is a \( p\% \) chance that there are zero errors for Mathematical checks (and no errors for the other tests as well) and a \( (100-p)\% \) chance that the Mathematical checks compliance test shows two errors (all other tests coming up clean, as in the previous case). Can you specify a \( p\% \) that would make you indifferent between having to live with the first and second situations I've specified?

This lottery is denoted: \(<(0,0,0,0), p, (2,0,0,0)>\)

(Subject - S) I'd say \( p=30 \) would make me feel the same about the two alternatives.
(E) Now suppose I ask you the same question but I change the Authorizations test results to one error in all cases.

\((1, 1, 0, 0, 0)\) for certain versus

\((0, 1, 0, 0), p, (2, 1, 0, 0)\)

What "p" would you specify?

If these attributes are utility independent, the subject should specify the same "p" value. This process is repeated with the same questions but changing the game to:

\((1, 2, 0, 0)\) for certain, versus

\((0, 2, 0, 0), p, (2, 2, 0, 0)\).

This set of questions tested whether or not Population Completeness was utility independent of Authorization Reviews. Subsequent questions tested the utility independence of Population Completeness from Comparisons Accuracy and Mathematical Checks by varying the values of these attributes in a manner identical to the process shown here for Population Completeness and Authorization Reviews. The researcher did not use a prepared instrument for this phase of the interview because the attributes were specified by subjects during the interview. (Ex post, it can be seen that the subjects all specified the same attributes. Had this been assured prior to the interviews, a test instrument could have been developed.) Instead of a prepared test instrument, the researcher wrote each gamble on a
sheet of blank paper as it was proposed to the subject. If the "p" specified is the same for all these questions, P is utility independent of P−.

The same set of questions could be asked for A, C, and M factors to verify mutual utility independence, but Keeney and Raiffa show that if mutual preferential independence is verified and just one attribute is found to be utility independent of its complement, then mutual utility independence is exhibited by the subject.

With the finding that X1 is utility independent of X1− and verification of mutual preferential independence, the form of the Multi-Attribute Utility Function can be established with the elicitation of scaling constants. (Section 3.3.5) If mutual utility independence is found to characterize the attributes specified by a subject, the quasi-additive or additive form of the MAUF is appropriate to model that subject's decision-making behavior. If the sum of the scaling constants (ki) equals one, then the additive form is the most appropriate functional form. If the sum of the scaling constants is not equal to one, then the quasi-additive form is most appropriate.
3.3.4 Elicitation of Unidimensional Utility Functions

Keeney and Raiffa's procedure for eliciting conditional utility functions entails proposing a series of simple lotteries to the subject. The process they describe is repeated for each individual attribute. Keeney and Raiffa provide a generic scenario for the techniques as well as giving examples of the questions that were asked by consultants in specific applications of the procedure. The technique begins with defining \( u_i(x_i^*) = 1 \) and \( u_i(x_{io}) = 0 \). The decision maker is asked to specify the value "\( p \)" such that he/she is indifferent between taking \( x_i^* \) for sure or a "\( p \)" chance of getting \( x_i^* \) versus a \((1-p)\) chance of getting \( x_{io} \). (The risky alternative is denoted \(<x_i^*, p, x_{io} >\).) The value of "\( p \)" that is specified is the utility of \( x_i^* \) since

\[
u_i(x_i^*) = p(u_i(x_i^*) + (1-p)(u_i(x_{io}))) = p\frac{1}{(1-p)} + (1-p)0 = p.
\]

This general outline was adapted for the auditing application undertaken in the dissertation research. Based on the internal control system attributes specified by the firm and confirmed by the pilot study subjects the following general form of questions was developed to elicit the conditional utility functions for the auditor-subjects. The general form of questions asked during elicitation interviews is:

I would like you to specify a probability that will make you indifferent between the value \( x_i \) and a lottery composed of the best and worst values of this attribute; a "\( p \)" chance to have the
best compliance test result and a \((1-p)\) chance of the worst result.

That is, you have the opportunity to take a chance of observing either \(x_{io}\) or \(x_{i*}\) in the internal control system or the value \(x_{i*}\) (with certainty). Can you tell me what sort of chance on the lottery would make you indifferent between it and the "sure thing?"

In general, this procedure is repeated for all points in a discrete-valued functions. The subject of the first phase of the pilot study had no trouble specifying the probabilities required by this procedure. Since pilot testing did not indicate any obvious problems with the procedure, the technique was not changed for the actual experiment. In this study, the values assumed by each attribute are limited to 0, 1, or 2. (See Appendix B) Therefore, only the utility for "1 error" needs to be elicited from the subjects. The utility of "zero errors" \((x_{i*})\) and the utility of "2 errors" \((x_{i0})\) are established by definition. That is, \(u_i("zero errors") = 1\) and \(u_i("2 errors") = 0.\)

The participants were actually asked the following questions during the elicitation of conditional utility functions.

(1) First, consider the Completeness controls. In a sample of size 40, finding 0, 1, or 2 errors is considered acceptable in compliance testing.

What probability, \(p\), would make you indifferent between auditing a system which you were certain would have one error in the compliance test of Completeness controls and auditing a system that had a \(p\) chance of having zero errors and a \((1-p)\) chance of having two errors in its compliance test for completeness of the sales data?
Consider the Authorization Review compliance test. You have a system which you are sure will have just one error in the compliance test of this control.

What probability, p, would make you just as comfortable to audit a system that had a p chance of yielding zero errors and a 1−p chance of showing two errors in the compliance test of Authorization Reviews?

Going now to Comparisons.

What probability, p, make you indifferent between auditing a system with a 100% chance of yielding only one error in a compliance test of Comparisons and auditing a system with a p chance of zero errors and a 1−p chance of two errors?

Finally, think about Mathematical Checks and the related compliance test thereof. Compare a certain ("for sure") outcome of the compliance test of one error to a p chance of having no errors in the sample and a 1−p chance of two errors in the sample.

What value of p would make you just as comfortable auditing the system with the uncertain outcome as you would feel about evaluating the system with the certain outcome?

The product of this exercise is a discrete utility function for each of the four transaction control attributes (Population completeness, Authorization review, Comparisons accuracy, and Mathematical checks).

Again, this phase of the interview was conducted without a prepared test instrument because subjects might have specified attributes for which to elicit utility functions that the researcher did not know about prior to undertaking the interviews.
3.3.5 Determination of Individual Scaling Constants

Scaling constants are derived by eliciting probabilities from subjects which make him/her indifferent between a lottery composed of all best versus all worst attribute values and a certain outcome of an event with one attribute at its best value and all others at their worst values. The elicited probability is the scaling constant corresponding to the attribute in question. In the case of an additive model, the sum of the scaling constants is one, but for a quasi-additive model (with interaction terms) an overall scaling constant is determined to make the values obtained from the MAUF in the closed interval between 0 and 1 (Keeney and Raiffa, 1976).

A generic example of the procedure followed to obtain the required probabilities follows.

Let $x_1$, $x_2$, and $x_3$ be three objective relevant to some hypothetical decision maker. The attribute measures are correspondingly $x_1$, $x_2$, and $x_3$. Define

$$u_i(x_{1o}) = 0 \text{ and } u_i(x_{i*}) = 1 \text{ for all } i.$$  

That is, the unidimensional utility of the worst possible value for each attribute is set at zero and the utility of the best possible value is set at one. Consequently,

$$u(x_{1o}, x_{2o}, x_{3o}) = 0 \text{ and }$$

$$u(x_{1*}, x_{2*}, x_{3*}) = 1.$$  

In words, the MAUF evaluated at the worst values for all attributes is zero and is one when evaluated at the best values for each attribute.
To find k1, propose the game:

A. \((x1^*, x2o, x3o)\) for certain

or the gamble

B. \(<(xi* for all i), p, (xio for all i)\>\)

and ask the decision maker to specify a value for \(p\) which would make him/her indifferent between alternatives A and B. For illustration, suppose the decision maker says \(p\) should be \(.4\). Then since

\[u(x1^*, x2o, x3o) = .4(u(x1*, x2*, x3*) + .6(u(x1o, x2o, x3o)),\]

the utility of option A is:

\[u(x1^*, x2o, x3o) = .4(1) + .6(0) = .4.\]

Therefore, \(k1 = .4\) since only the term of the MAUF involving \(u1(x1)\) alone will be non-zero.

This procedure is repeated for all attributes. If the sum of the \(k_i\) equals one, no overall scaling constant is needed as the MAUF is consequently additive. If the sum of the \(k_i\) is not equal to one, the overall scaling constant, \(K\), must be determined such that the utility function values are on the desired interval. This may be done using a computer algorithm or other means to derive the appropriate constant.

\(K\) is the solution to

\[1 + K = (1+Kk1)(1+Kk2)(1+Kk3) \ldots (1+Kkn).\]

In the actual elicitation interviews, the scaling constants related to the four transaction control attributes were determined using the following script.

You are considering the compliance testing of the transaction controls over Sale for a medium size manufacturing company. It has been decided that a
sample size of 40 will be used for the compliance tests of details. We will denote the four types of transaction control attributes as follows:

Population Completeness. . . . . . . P
Authorization Reviews. . . . . . . . . A
Comparisons. . . . . . . . . . . . . . . C
Mathematical Checks. . . . . . . . . M

Suppose that you knew for certain that the internal control system was going to yield compliance test results as follows: Population Completeness (P) - 0 errors; Authorization Reviews (A) - 2 errors; Comparisons Accuracy (C) - 2 errors; Mathematical Checks (M) - 2 errors.

Suppose that one the other hand that you weren't sure of the outcome of the compliance tests, but that there was a p\% chance that the results would be:

P: 0 errors
A: 0 errors
C: 0 errors
M: 0 errors

(The best possible situation)

and a (100-p)\% chance that the results would be:

P: 2 errors
A: 2 errors
C: 2 errors
M: 2 errors

(The worst possible situation)

What value of p would make you just as comfortable about auditing the firm with the certain outcome as you would about the firm with the uncertain (but potentially better or potentially worse) results?
The probability given as an answer to this scenario is the scaling constant related to the Population Completeness attribute. The same set of questions are asked in reference to each of the remaining attributes. The "certain" outcome is changed each time such that the attribute in question is that which showed zero errors (the other attributes showing two errors).

3.3.6 Checks for Consistency

Keeney and Raiffa (1976) indicate that checks be made periodically to ensure that the MAUF is representative of the subject's stated preferences. This involves asking the questions over again, inserting the answers given by the subject where appropriate. For example, "If you felt there was a 40% chance of \((x1^*, x2^*, x3^*)\) occurring and a 60% chance that \((x1o, x2o, x3o)\) would happen, you would be indifferent between that uncertain situation and knowing that \((x1', x2o, x3o)\) would happen for sure?", would be asked in the elicitation of scaling constants if a subject gave ".4" as an answer to the question "What value of \(p\) would make you indifferent between the options presented?".

Checks for consistency can also be made by changing the value of \(p\) slightly and asking for preference between the lottery and the certain event. The option with the higher expected utility must be selected by the subject to prove his/her consistency. Subjects were found to be very consistent. The auditors often stated they felt that consistency was important their decision making.
3.4 PILOT STUDY

3.4.1 Pilot Study Design

Before designing the experiment, interviews were conducted with representatives of five public accounting firms to gain an understanding of industry and firm procedures for making evaluations of internal control systems and to verify that the context of auditor decision-making was amenable to experimentation. Once the study's feasibility was determined, further inquiry was undertaken to determine the mode of implementing the experimental task.

The pilot study was conducted in two phases which parallel the design of the main experiment. One subject was used in the first phase and another was used in the second phase. Preliminary testing of the elicitation procedures was performed and, in a later session, the researcher conducted a trial run of the experimental task.

The subject of the first part of the pilot study was a Manager from the same firm which supplied subjects for the main study. An interview was conducted to test utility function elicitation procedures. This session took place approximately one month prior to the first utility function elicitation interview of the main study.
3.4.2 Pilot Study Findings - Part I

3.4.2.1 Attributes

The attributes specified by the subject were:

(1) Population Completeness (P)

(2) Authorizations Review (A)

(3) Comparisons Accuracy (C)

(4) Mathematical Checks (M)

In order to obtain "Moderate" reliance on the transaction controls, a sample size of 40 was specified for compliance testing. The sample size is determined by guidelines used in the participating C.P.A. firm, based on the type of internal control and the degree of reliance to be placed on the control.

This sample size allows for up to two errors with the system still being reliable to the desired degree. The same sample size is used for all four transaction control attributes of internal control over Sales.

3.4.2.2 Individual Utility Functions

The conditional utility functions for the subject of the pilot study are as follows:

<table>
<thead>
<tr>
<th>Possible Number of Errors</th>
<th>Utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1*</td>
</tr>
<tr>
<td>1</td>
<td>.6</td>
</tr>
<tr>
<td>2</td>
<td>0*</td>
</tr>
</tbody>
</table>

*Endpoints are defined to establish the range of the function.
Similarly, the conditional utilities for "1 error" of the other attributes are:

Authorization Reviews (A): \( u_A(1 \text{ error}) = .5 \)
Comparisons Accuracy (C): \( u_C(1 \text{ error}) = .6 \)
Mathematical Checks (M): \( u_M(1 \text{ error}) = .6 \)

3.4.2.3 Scaling Constants

For Population Completeness, the scaling constant, \( k_P \), is determined as follows:

\[
\begin{align*}
u(0,2,2,2) &= .8(u(0,0,0,0)) + .2(u(2,2,2,2)) = (.8)(1) + (.2)(0) = .8.
\end{align*}
\]

(See Keeney and Raiffa, 1976, pp. 302-303)

For Authorization Reviews, \( k_A = .7 \)
For Comparisons Accuracy, \( k_C = .7 \)
For Mathematical Checks, \( k_M = .7 \).

\( k \), the overall scaling constant, is the solution to the following equation:

\[
1 + k = (1 + .6k)(1 + .7k)(1 + .7k)(1 + .7k).
\]

For this subject, \( k = -.9942508 \).

The MAUF is therefore:

\[
1-.9942508u(x) = (1-.5965505u_P(x_P))(1-.6959756u_A(x_A))(1-.6959756u_C(x_C))(1-.6959756u_M(x_M)).
\]

Early discussions with colleagues indicated that some subjects may have no personal preferences about the impact of compliance tests on the "auditability" of the client. That is, there are subjects for whom personal risks and
preferences will not enter their analyses of the results of compliance testing. The single subject in Part I of the Pilot Study concurred with this contention and added that the experiment may show differences in the degree of internalization across years of service with the firm (Rank) because of recent changes in firm procedures for internal control evaluation.

Since 1980 the audit approach of the participating firm has been based on statistical sampling. The sample sizes of the compliance tests depend on the degree of "planned reliance." The number of irregularities that are deemed acceptable within the samples are specified by the firm's workpapers. Such quantification and specificity may have reduced the auditor's internalization of the audit data. For auditors trained in judgmental (as opposed to statistical) sampling (prior to 1980), there may be more internalization of audit data.

3.4.3 **Pilot Study Findings - Part II**

The second phase of the pilot study was conducted after the MAUPS were elicited from all subjects involved in the actual experiment. The audit Manager who participated in the first phase of the pilot study was unavailable to participate in the second phase, so a Senior Manager from the actual experiment subject pool was selected to test the
experimental task. The task presented to the subject was accomplished by him with no apparent difficulty. In the discussion which followed the completion of the task, the subject of the pilot study stated the task was an exercise that other subjects should not find difficult to perform and did involve the use of professional judgment required of an auditor evaluating an internal control system.

No changes were made in the experimental task as a result of conducting the pilot study since the subject responded to the task as the researcher had expected. Consequently, the data obtained from the subject of the second phase of pilot study were included with the data collected during the main phase of experimentation. (Subject 9 was the participant in this second phase of pilot testing.)

3.5 EXPERIMENTAL TASK

A potential order effect exists in the procedure followed for the experiment. All subjects were first interviewed to elicit multi-attribute utility functions and then were asked to evaluate the hypothetical cases during a subsequent meeting. Due to the nature of the elicitation process, this problem could not be avoided. Since the same cases were to be used for all subjects, the cases had to contain adequate information for all subjects to evaluate. Since the attributes considered important to each subject
were not known until interviews were conducted, case evaluations were by necessity made at a date following the initial interviews.

In the two weeks following the completion of the interview phase of the study, the researcher met individually with the subjects to administer the test instrument. The instrument consisted of a two-page case and a deck of 42 different cards bearing the results of four compliance tests (one for each of the attributes of the transaction controls: P, A, C, and M). Page one of the case packet detailed what the subject was being asked to do and page two described the control environment of the "client" and established the frame of references for the task (See Appendix D). Preliminary discussions with colleagues and practicing auditors indicated that the case materials should be quite simple in content (i.e., contain a very limited amount of detail) to help subjects confine their focus to the internal control being investigated. The subject of the first phase of pilot testing, who had taught professional development seminars on internal controls evaluation, also recommended that the cases be free of extraneous facts. This auditor also stated that subjects needed only the results of compliance tests and a brief description of the client to make the evaluations of system reliability requested in this experiment. The subject of part two of the
pilot study showed this statement to be true. The subject appeared to have no difficulty assigning ratings to the cases presented for evaluation.

The results of the hypothetical compliance tests were written on 3"x5" index cards to allow simple randomization of the cases. The deck was shuffled for each subject (including the subject of the second phase of the pilot study) to eliminate a potential order effect in the subjects' evaluations. Having the cases on cards allowed subjects to compare the compliance test results easily when assigning ratings, thereby allowing subjects to achieve consistency in their judgments.

Eighty-one possible compliance test reports could have been presented to subjects for evaluation (3 possible outcomes on 4 factors: \(3 \times 3 \times 3 \times 3 = 81\)). Based on procedures followed in prior studies (e.g., Dawes and Corrigan, 1972) it was decided that approximately ten cases per factor (cue), or approximately 40 cases, should comprise the subset of all possible cases. This was expected to provide enough data for analysis yet would be less taxing to the subjects than evaluation of all 81 cases.

Using a random number table, 42 cases were selected from the 81 combinations available. Independence among the factors was verified by determining the correlations among the four factors. Random selection of the cases did result
in a sample with statistically uncorrelated factors. Table 1 shows the 42 cases that were selected. Table 2 shows the correlations between each of the six pairs of factors.

The MAUT models elicited from the auditors produce a preference ordering of the cases for each subject by assigning an overall utility to each hypothetical system. A measure of the auditors' perceptions of internal control reliability was needed as data for analysis of the predictive ability of the MAUT model. Middaugh (1981) asked his subjects to assign a 0-100 rating to the investment projects which he compared to ratings produced by multi-attribute utility functions for capital budgeting decisions of managers. This type of rating system was used by the pilot study subject with no apparent trouble and was subsequently used during the main experiment.

Auditor-subjects were asked to assign a 0-100 rating to each of the 42 cases to indicate their feeling of comfort concerning the reliability of the internal control systems that produced the given compliance test results. As subject assigned values to the cases by writing ratings on the cards, either the researcher or the subject put the cards in order according to the subject-assigned ratings. After completing an initial pass through the cards, the subject made a further evaluation of all the cases which were assigned the same rating, attempting to encourage the
<table>
<thead>
<tr>
<th>Population Completeness</th>
<th>Authorization Reviews</th>
<th>Comparisons Accuracy</th>
<th>Mathematical Checks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
subject to make the rating as fine as possible. At the end of some sessions, ties among cases remained as subjects stated that they felt some cases did indicate identical degrees of reliability of the transaction controls.

The worksheet for transaction controls evaluation calls for the evaluation of the internal control system to determine the controls that are in existence, then links this worksheet to the compliance test worksheet which documents the tests of system attributes performed and the results obtained therefrom. The completion and analysis of these data provide the basis for the subjective conclusions
which the auditor must make in the final sections of the worksheet for transaction controls evaluation.

The experimental task given to subjects to perform was similar to the activity required on an actual audit. The elicitation of MAUPs focuses on the attributes upon which subjects base their evaluations of the internal control systems. The specification of system ratings by subjects corresponds to drawing conclusions asked for on the worksheet for transaction controls evaluation. The ratings requested in the experiment, like the conclusions drawn in an actual audit setting, require the auditors to combine reports of compliance tests in light of the characteristics of the system and the constraints of the statistical tests. Therefore, although numerical ratings are not made in actual audit situations, the experimental task is reasonably close to the procedures followed while making an internal control evaluation in an actual audit engagement.

Upon completion of the case evaluation task the subjects responded to a six-item demographics questionnaire. (See Appendix E) Based on prior studies, the researcher expected education, rank, and experience to have an effect on subject decision-making behavior. The questionnaire provided information about rank used in statistical analysis as well as data that may be useful in subsequent studies to test the effect of demographic factors on subject behavior.
3.6 **ANALYSIS**

3.6.1 **Testing the Predictive Ability of MAUT**

As stated in Chapter I, the hypothesis of the first objective is:

The rating of internal control strength given by subjects to experimental cases will be significantly correlated with the ratings derived from the MAUT model.

The data presented to auditor-subjects for subjective evaluation were also evaluated using the MAUT elicited from each auditor. These ratings were paired with that auditor's subjective rating of the internal control systems and Kendall's nonparametric coefficient of rank correlation was determined. Statistical significance of these correlations show whether or not MAUT is a good predictor of auditor decisions with respect to internal control evaluation. The classical theory correlation coefficient is not used because the underlying distributions of the variables are not known. The predictions used to list the theory are "ratings" which determine a preference ordering of the internal control systems. The ratings provided by the models carry no inherent meaning.

Testing the ability of the multi-attribute utility theory model to predict auditor's decisions attempts to present evidence of the theory's aptness in this context. Further evidence, though, concerning its relative merit as a descriptive model requires a comparison of the MAUT
model's predictions to those made by "simpler" models. Here "simpler" is used to refer to models which are either based on relaxed conditions of MAUT or are not utility models at all. It is hypothesized that the ability of the MAUT model to predict auditor judgments will be greater than that of the simpler models due to:

(1) Incorporation of risk preferences into the MAUT model, and

(2) Direct elicitation of the MAUP from subjects being modelled.

The predictions from the MAUT model were compared to predictions from two simpler utility models using a nonparametric two-way layout for ordered alternatives. (Hollander and Wolfe, 1973) This analysis examines the superiority of the elaborate model-building procedures of MAUT. The method used for this analysis is described in Section 3.6.2.

3.6.2 Nonparametric Statistical Analysis

The second research hypothesis is:

The ratings derived from the MAUT model will be more highly correlated with subject ratings of internal control strength than will the output of simpler utility models which do not incorporate all the elements of MAUT.

To determine the relative merit of the model which incorporates subject-elicited scaling constants for individual attributes, a nonparametric two-way layout is performed.
The correlations between subjective ratings of cases describing internal control systems and the ratings of these cases derived from each subject's MAUT model constitute the data for the MAUT model. The correlations between subject-specified evaluations and predictions of two simpler utility models which are constructed under relaxed conditions of MAUT are the data related to these simpler models. The models may be defined:

A. Equally Weighted Additive Model (EW)

\[ V = u_1(x_1) + u_2(x_2) + \ldots + u_n(x_n) \]

where \( V \) is the rating using this model and \( u_i(x_i) \) is the unidimensional utility function for attribute \( i \).

B. Additive MAUT Model Weighted with Scaling Constants (SW)

\[ V' = \sum k_i u_i(x_i) \]

where \( V' \) is the rating using this model and \( k_i \) is the scaling constant for attribute \( i \).

C. Multi-Attribute Utility Theory Model (as defined earlier).

Models A and E are constructed by the researcher using portions of the data obtained during interviews aimed at developing the MAUF for each subject.

For a subject who exhibits additive independence among attributes, Models E and C will be identical.
Step 1: The ratings of the hypothetical internal control systems as given subjectively (call this "Model S") are compared with the ratings based on Model A (the equally weighted additive model). The correlation between Model A and Model S is determined using Kendall's rank correlation coefficient (tau). These statistics are calculated for each subject.

Step 2: Similarly, the correlation between ratings of Model B (scale weighted additive model) and Model S is determined.

Step 3: The correlation between the MAUT model (Model C) and Model S as determined previously is used in this phase of testing as well.

Step 4: The correlations are examined to determine whether one method of scaling the multi-attribute utility function improves the correspondence between the mathematical model and the auditors' subjective procedure for evaluating the internal control systems. The outcome of this analysis sheds light on the issue of whether the tests for preferential, utility, and additive independence (the results of which determine the functional form of the model) are important for modelling auditors' decision-making behavior.

The following relationship among the correlations is expected to be found:

This finding will indicate that the refinements of the MAUT procedure aid in modelling the subjects' behavior. If this is true, $\text{Corr}(C, S)$ should be greater than the other correlations.

To test this relationship, a distribution-free test for ordered alternatives based on Friedman rank sums (attributed to Page) is employed. (Hollander and Wolfe, 1973) A two-way layout design is used in which treatments are ordered, a priori, in an intended direction. The design is:

\[
\begin{array}{ccc}
\text{Model A} & \text{Model B} & \text{Model C} \\
\hline
S_1 & \text{Corr}(A, S) & \text{Corr}(B, S) & \text{Corr}(C, S) \\
1 & 1 & 1 & 1 \\
2 & 2 & 2 & 2 \\
3 & 3 & 3 & 3 \\
\end{array}
\]

Where $\text{Corr}(A, S)i$, $\text{Corr}(B, S)i$, and $\text{Corr}(C, S)i$ are the previously described correlations for each model/subject combination.

The hypothesized relation can be expressed in terms of a treatment effect $t$ as follows:

$$H_0: \ t(\text{MAUT}) = t(\text{SW}) = t(\text{ER})$$
H1: \( t(MAUT) \geq t(SW) \geq t(EW) \) (with at least one strict inequality)

where \( t(.) \) are the "treatment effects" corresponding to each of the utility models being examined.
Chapter IV

EXPERIMENTAL RESULTS

4.1 ELICITED UTILITIES AND SCALING CONSTANTS

An experiment was conducted to study the applicability of multi-attribute utility theory in an auditing context. MAUT is hypothesized to be an appropriate model of auditor decision-making behavior because it is an individual preferences model which incorporates multiple factors and the decision maker's propensities toward risk as it impacts his/her decision-making behavior. These characteristics are representative of the complexities of the auditor's decision-making task involved in internal control systems evaluation.

In Phase I of the experiment, interviews were conducted with 15 practicing auditors at the Columbus, Ohio office of an international public accounting firm. The purpose of these interviews was to elicit a multi-attribute utility function (MAUF) from each auditor which would model that subject's decision-making behavior with respect to making an evaluation of the internal control system, specifically the "transaction controls" over credit Sales.
The MAUPs were elicited using Keeney and Raiffa's (1976) technique. The attributes were specified by firm policy, so the researcher expected four attributes to be considered relevant to internal controls evaluation by the auditors acting as subjects. The firm specifies four control attributes which are:

1. Population Completeness (P)
2. Authorization Reviews (A)
3. Comparisons Accuracy (C)
4. Checks of Mathematical Accuracy (M).

The auditor-subjects answered questions about their perceptions of a hypothetical audit client and the compliance testing of the transaction controls over Sales for the hypothetical company.

The specification of the compliance test sample size and the degree of planned reliance to be placed on the controls established the range of values for the attributes. Audit procedures dictated by the firm state that given a sample size of 40 and moderate reliance on transaction controls, the acceptable error rates (number of irregularities in the compliance test) are 0, 1, and 2. Beyond two errors, the controls cannot be relied upon. The conditional utilities (ui("1 error")) and scaling constants (K and k_i) collected during Phase I are shown in Table 3. The conditional utilities and scaling constants in Table 3 and the
functional form which follows, constitute each subject's MAUT model. The quasi-additive MAUF is:

\[ 1 + Ku(x) = (Kk_1 u_1(x_1) + 1)(Kk_2 u_2(x_2) + 1)(Kk_3 u_3(x_3) + 1) \ldots (Kk_{n\text{un}} u_{n\text{un}}(x_{n\text{un}}) + 1) \]

For example, the MAUF for Subject #3 is:

\[ 1-.9901107u(x)= (1-.4950554uP(xP))(1-.7920886uA(xA))(1-.6930775uC(xC))(1-.6930775uM(xM)) \]

The conditional utilities for each attribute revealed the following information:

**Population Completeness**

1. 5 out of 15 (5/15) subjects are risk neutral (i.e., uP(1 error) = .5.)
2. 10 out of 15 (10/15) subjects are risk preferent (i.e., uP(1 error) > .5.)

**Authorization Reviews**

1. 7/15 subjects are risk neutral
2. 8/15 subjects are risk preferent

**Comparisons Accuracy**

1. 6/15 subjects are risk neutral
2. 9/15 subjects are risk preferent

**Mathematical Checks**

1. 7/15 subjects are risk neutral
2. 8/15 subjects are risk preferent.

Seven subjects assigned the same utility to ui(1 error) for all attributes.
Considerable diversity in subject specified probabilities for scaling constants is revealed in Table 3. Four subjects specified the same value for each attribute's scaling constant. (The values differed across subjects.) The sum of the scaling constants was greater than one for twelve subjects (and consequently the overall scaling constant was less than one for these subjects). The sum of the scaling constants was less than one for the remaining three subjects (and thus the overall scaling constant for each of these auditors was greater than one). There were no subjects for whom the scaling constants summed to one and thus none of the subjects' preferences were characterized by additive multi-attribute utility functions. (The sums of scaling constants ranged from a low of .4 to a high of 3.3.) Keeney and Raiffa (1976) state that no definite interpretation of the scaling constants can be made, but it is interesting to note that Population Completeness was assigned the highest value in six of fifteen cases and shared the top value in three other instances. (This does not include the cases in which the ki's were identical.)

The second phase of experimentation asked the auditor-subjects who participated in the first phase to evaluate a set of hypothetical reports of compliance test results. The researcher met with the subjects individually so that questions could be answered.
### TABLE 3
Conditional Utilities and Scaling Constants

<table>
<thead>
<tr>
<th>S</th>
<th>uP</th>
<th>uA</th>
<th>uC</th>
<th>uM</th>
<th>kF</th>
<th>kA</th>
<th>kC</th>
<th>kM</th>
<th>Overall Scaling Constant K*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>-.9126286</td>
</tr>
<tr>
<td>2</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.3</td>
<td>.25</td>
<td>.25</td>
<td>.25</td>
<td>-.1297867</td>
</tr>
<tr>
<td>3</td>
<td>.7</td>
<td>.85</td>
<td>.8</td>
<td>.8</td>
<td>.5</td>
<td>.8</td>
<td>.7</td>
<td>.7</td>
<td>-.9901107</td>
</tr>
<tr>
<td>4</td>
<td>.8</td>
<td>.8</td>
<td>.7</td>
<td>.7</td>
<td>.8</td>
<td>.7</td>
<td>.7</td>
<td>.8</td>
<td>-.9962253</td>
</tr>
<tr>
<td>5</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
<td>.2</td>
<td>.7545541</td>
</tr>
<tr>
<td>6</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.4</td>
<td>.5</td>
<td>.4</td>
<td>.4</td>
<td>-.8218063</td>
</tr>
<tr>
<td>7</td>
<td>.6</td>
<td>.5</td>
<td>.7</td>
<td>.7</td>
<td>.4</td>
<td>.4</td>
<td>.3</td>
<td>.3</td>
<td>-.6359550</td>
</tr>
<tr>
<td>8</td>
<td>.7</td>
<td>.75</td>
<td>.59</td>
<td>.59</td>
<td>.41</td>
<td>.41</td>
<td>.41</td>
<td>.41</td>
<td>-.7920866</td>
</tr>
<tr>
<td>9</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.4</td>
<td>.4</td>
<td>.4</td>
<td>-.8218063</td>
</tr>
<tr>
<td>10</td>
<td>.6</td>
<td>.6</td>
<td>.6</td>
<td>.5</td>
<td>.25</td>
<td>.2</td>
<td>.15</td>
<td>.15</td>
<td>1.0618128</td>
</tr>
<tr>
<td>11</td>
<td>.8</td>
<td>.75</td>
<td>.7</td>
<td>.6</td>
<td>.75</td>
<td>.75</td>
<td>.5</td>
<td>.5</td>
<td>-.9780643</td>
</tr>
<tr>
<td>12</td>
<td>.9</td>
<td>.85</td>
<td>.85</td>
<td>.85</td>
<td>.9</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>-.9991869</td>
</tr>
<tr>
<td>13</td>
<td>.7</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.75</td>
<td>.5</td>
<td>.6</td>
<td>.25</td>
<td>-.9506947</td>
</tr>
<tr>
<td>14</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
<td>.1</td>
<td>.1</td>
<td>.1</td>
<td>.1</td>
<td>6.6080239</td>
</tr>
<tr>
<td>15</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.8</td>
<td>.7</td>
<td>.7</td>
<td>.6</td>
<td>.6</td>
<td>-.9837215</td>
</tr>
</tbody>
</table>

The overall scaling constant, K, was determined by solving the following equation for K:

\[ 1 + K = (1+K_P)(1+K_A)(1+K_C)(1+K_M) \]

This was accomplished with the use of a computer algorithm written by Richard A. Young, a colleague at the Ohio State University.
4.2 KENDALL'S COEFFICIENT OF RANK CORRELATION

A measure of MAUT's ability to model the subjects' evaluations of compliance tests of transaction controls is Kendall's coefficient of rank correlation. Kendall's tau was computed for each subject by comparing the subject's ratings of each of the 42 cases of compliance test results to the rating predicted with that subject's MAUT. Table 4 shows the results of this computation, including the levels of significance for each correlation coefficient. It can be seen that subjects' ratings are generally highly correlated with the ratings predicted using the elicited MAUT. All correlations are significantly different from random (t=0).

The alpha level corresponds to the hypothesis test related to Kendall's tau. This is a test of independence between the subject-specified ratings and the model-derived ratings. That is:

$H_0: \tau = 0$

$H_1: \tau \neq 0$.

Alpha is the lowest level of significance at which the null hypothesis can be rejected. From Table 4 it can be seen that all of the tests can be rejected at any reasonable level of significance, indicating that the two sets of ratings are not statistically independent.

A test on the correlations between subject ratings and predictions of the MAUT model to check for a rank effect
### TABLE 4

Kendalls tau Between Subjective and MAUT Ratings

<table>
<thead>
<tr>
<th>Subject</th>
<th>tau</th>
<th>alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.94339</td>
<td>.0001</td>
</tr>
<tr>
<td>2</td>
<td>.72890</td>
<td>.0001</td>
</tr>
<tr>
<td>3</td>
<td>.42824</td>
<td>.0001</td>
</tr>
<tr>
<td>4</td>
<td>.86107</td>
<td>.0001</td>
</tr>
<tr>
<td>5</td>
<td>.99809</td>
<td>.0001</td>
</tr>
<tr>
<td>6</td>
<td>.84815</td>
<td>.0001</td>
</tr>
<tr>
<td>7</td>
<td>.72562</td>
<td>.0001</td>
</tr>
<tr>
<td>8</td>
<td>.59763</td>
<td>.0000</td>
</tr>
<tr>
<td>9</td>
<td>.71415</td>
<td>.0001</td>
</tr>
<tr>
<td>10</td>
<td>.77749</td>
<td>.0001</td>
</tr>
<tr>
<td>11</td>
<td>.72314</td>
<td>.0001</td>
</tr>
<tr>
<td>12</td>
<td>.61447</td>
<td>.0000</td>
</tr>
<tr>
<td>13</td>
<td>.68580</td>
<td>.0001</td>
</tr>
<tr>
<td>14</td>
<td>.83397</td>
<td>.0001</td>
</tr>
<tr>
<td>15</td>
<td>.71464</td>
<td>.0001</td>
</tr>
</tbody>
</table>

was made using Wilcoxon's rank sum procedure (Hollander and Wolfe, 1973). It was hypothesized that the degree of internalization of risk and the implications of multiple attributes would be greater for auditors with more experience. That is,
Ho: \text{corr(Managers)} = \text{corr(Seniors)}

Ha: \text{corr(Managers)} > \text{corr(Seniors)}

The statistic, \( W \), is the sum of the rank of the manager's correlations (from a ranking of all 15 subject's correlations). For the data, \( W=49 \). \alpha_{(n=6,m=9,W=49)} = .477. This indicates that the null hypothesis cannot be rejected at any reasonable level of significance. Therefore, it cannot be concluded that there is a rank effect of a difference in degree of predictibility of the MAUT model. The inability to find the hypothesized rank effect may be due to the low power of the test resulting from a small sample size.

4.3 \textsc{Nonparametric Test of Comparison}

Kendall's \( \tau \) (coefficient of rank correlation) was computed on subjects' ratings with values predicted using (1) an equally weighted linear additive utility model, (2) a linear additive utility model weighted with scaling constants elicited from subjects during Phase I interviews, and (3) the MAUF, incorporating conditional utilities, individual scaling constants, and the overall scaling constant, \( K \). The results of these computations are shown in Table 5.

The hypothesis to be tested using a nonparametric two-way layout is:
TABLE 5

Correlations of Subject Ratings With Three Utility Models

<table>
<thead>
<tr>
<th>Subject in Firm</th>
<th>Equally Weighted Model (EW)</th>
<th>Scale Weighted Model (SW)</th>
<th>MAUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Sr. Mgr.</td>
<td>.98817H</td>
<td>.98817H</td>
<td>.94339L</td>
</tr>
<tr>
<td>2 S. Sr.</td>
<td>.67299L</td>
<td>.76470H</td>
<td>.72890</td>
</tr>
<tr>
<td>3 Sr.</td>
<td>.59723H</td>
<td>.47959</td>
<td>.42824L</td>
</tr>
<tr>
<td>4 S. Sr.</td>
<td>.85836</td>
<td>.81622L</td>
<td>.86107L</td>
</tr>
<tr>
<td>5 S. Sr.</td>
<td>1.00000H</td>
<td>1.00000H</td>
<td>.99809L</td>
</tr>
<tr>
<td>6 Sr. Mgr.</td>
<td>.85472</td>
<td>.86759H</td>
<td>.84815I</td>
</tr>
<tr>
<td>7 S. Sr.</td>
<td>.67266L</td>
<td>.74708H</td>
<td>.72562</td>
</tr>
<tr>
<td>8 Sr. Mgr.</td>
<td>.58459L</td>
<td>.58459L</td>
<td>.59763H</td>
</tr>
<tr>
<td>9 Sr. Mgr.</td>
<td>.65382L</td>
<td>.74463H</td>
<td>.71415</td>
</tr>
<tr>
<td>10 Sr. Mgr.</td>
<td>.64198L</td>
<td>.76777</td>
<td>.77749H</td>
</tr>
<tr>
<td>11 S. Sr.</td>
<td>.50622L</td>
<td>.60851</td>
<td>.72314H</td>
</tr>
<tr>
<td>12 S. Sr.</td>
<td>.46044L</td>
<td>.53349</td>
<td>.61447H</td>
</tr>
<tr>
<td>13 Sr.</td>
<td>.44896L</td>
<td>.62390</td>
<td>.68580H</td>
</tr>
<tr>
<td>14 S. Sr.</td>
<td>.95759H</td>
<td>.95759H</td>
<td>.83397I</td>
</tr>
<tr>
<td>15 Mgr.</td>
<td>.70988L</td>
<td>.75090H</td>
<td>.71464</td>
</tr>
</tbody>
</table>

Frequency: 4 H, 9 L, 8 H, 2 L, 6 H, 5 I

H = model with highest correlation among the three models.
I = model with lowest correlation among the three models.
Hc: \( t(\text{MAUP}) = t(\text{SW}) = t(\text{EW}) \)

Ha: \( t(\text{MAUP}) \geq t(\text{SW}) \geq t(\text{EW}) \) (with at least one strict inequality).

The statistic for Page's test, \( L \), is computed by weighting the sum of the column ranks for each model by its position in the hypothesis.

\[
\]

At alpha = .05, the null hypothesis cannot be rejected.

Even separating the sample of all 15 subjects into "All Seniors" (Seniors and Supervising Seniors) and "All Managers" (Managers and Senior Managers) there still appears to be no significant statistical support for the hypothesis. For "All Managers", \( L = (1)(15) + (2)(20) + (3)(12) = 74 \). \( L^* (n=6, \alpha=.05) = 79 \). Since \( L < L^* \), the null hypothesis again cannot be rejected.

For "All Seniors", \( L = (1)(15) + (2)(20) + (3)(19) = 112 \). \( L^* (n=6, \alpha=.05) = 116 \). Since \( L < L^* \), the null hypothesis cannot be rejected for this subset of the subjects either.

Separating by individual ranks also showed no apparent support for the alternative hypothesis.

Page's test addresses a very specific question, that of a hypothesized trend in the data. Since the hypothesized
trend was not found, a general test of differences among the models was performed.

A two-way layout based on Friedman Rank Sums tested the following relationship:

\[ H_0: t(EW) = t(SW) = t(MAUT) \]
\[ H_a: \text{the t's are not all equal.} \]

The same ranking procedure was followed as that used in Page's test. In the two-way layout, average ranks across blocks are used and are denoted, R.i.

\[ R_{EW} = 1.6, \quad R_{SW} = 2.3333, \quad R_{MAUT} = 2.0666. \]

The tables included in Hollander and Wolfe (1973) did not extend to fifteen subjects, so a large sample approximation was used for the \( S \) statistic. Because ties existed in the ranks, a correction for ties was needed. The resulting modified \( S \) statistic (\( S' \)) is:

\[ S' = \frac{744}{168} = 4.429 \]

This statistic has an asymptotic Chi-squared distribution, so the large sample approximation employs Chi-squared on two degrees of freedom, which at alpha=.05 is the critical value of 6 (approximately). Since \( S'<6 \) (the critical value), the null hypothesis cannot be rejected.

The two overall tests (Page's test and the general two-way layout) showed no differences across models, but the numerous combinations of orderings might account for the inconclusive findings of the overall tests.
Pairwise tests of differences between models were performed to check for differences across models. Wilcoxon's signed rank, one-sample, test for differences in location was used.

**MAUT vs. EW**

Ho: \( t(\text{MAUT}) = t(\text{EW}) \)

Ha: \( t(\text{MAUT}) > t(\text{EW}) \)

Since Wilcoxon's test analyzes difference in the sample data, the statistical hypothesis is:

Ho: \( D = 0 \)

Ha: \( D > 0 \).

Wilcoxon's \( T^+ = 86 \). The critical value \( t(0.05, 15) = 90 \). Since \( T^+ \) is less than the critical value, the null hypothesis cannot be rejected. The lowest alpha level at which this test could be rejected is 0.076.

**MAUT vs. SW**

Ho: \( D = 0 \)

Ha: \( D > 0 \)

For the comparison of the data generated by these models, Wilcoxon's statistic, \( T^+ = 54 \). Again, the critical value is \( t(0.05, 15) = 90 \). Therefore, the null hypothesis cannot be rejected. Furthermore, the null cannot be rejected at any reasonable level of significance.

While the sample size of this study is very small, the data obtained in the study were tested for differences due
to experience. Of the six subjects (4, 8, 10, 11, 12, and 13) whose correlations between the subjective ratings and MAUF predictions were the highest of the three correlations computed, 1 was a "Senior", 3 were "Supervising Seniors", zero were "Managers" and 2 were "Senior Managers." Hence, from a subjective assessment, there does not appear to be a "rank effect" within the subset of subjects with this characteristic. There were two subjects who exhibited a complete reversal of the hypothesized ordering of computed correlations. Of these subjects, one was a "Senior" and one was a "Senior Manager." Since only two subjects exhibited this characteristic, it is difficult to draw any conclusions with respect to a rank effect on this data.

In terms of conditional probabilities, the chance that the hypothesized relationship is true given a subject has a particular rank is:

\[ p(\text{corr (Subj., MAUF)} > \text{Others} / \text{rank i}) \]

- Senior (1/2) \[ \cdot 5 \]
- Supervising Senior (3/7) \[ 0.4285714 \]
- Manager (0/1) \[ 0 \]
- Senior Manager (2/5) \[ 0.4 \]
- Overall (6/15) \[ 0.4 \]

To test these data for difference due to rank, it is hypothesized:

- \( H_0: p = 0.4 \)
- \( H_a: p \neq 0.4 \)
where $p = .4$ is proposed as the value against which to perform this test as it is the overall sample probability of having the MAUI correlation the highest among the three correlations. Using Wilcoxon's signed rank procedure,

$T = 3$, which does not allow for rejection of the null hypothesis at any reasonable significance level. The power of this test, given the very small sample size, is quite low (.172) indicating the even if the hypothesized relationship is true, the test may not be able to detect it.

The hypothesis of a rank effect was further tested by looking at the effect of rank from another perspective. The conditional probability that a subject holds a particular rank given his/her correlations exhibited the hypothesized relationships is:

- Senior (1/6) \( .16666667 \)
- Supervising Senior (3/6) \( .5 \)
- Manager (0/6) \( 0 \)
- Senior Manager (2/6) \( .33333333 \)
- Overall (6/15) \( .4 \)

Similarly testing $p = .4$ using Wilcoxon's signed rank procedure, $T=2$, which again does not allow rejection of the null hypothesis at any reasonable level of significance given the power of the test. Hence, an effect of rank on the ability of the model to predict auditor ratings of the
reliability of internal control systems cannot be concluded from the data obtained in the dissertation research.

An analysis of these data was not performed to check for the effect of age because there was a monotonic relationship between age and rank in the firm. Nonparametric analyses employing ranks of the data would be essentially identical tests.
Chapter V

CONCLUSIONS, DISCUSSION, AND LIMITATIONS

5.1 IMPLICATIONS OF HYPOTHESIS TESTS

5.1.1 Ability of MAUT to Model Auditors Evaluations

The ability of MAUT to model the evaluations of hypothetical internal control systems made by auditor-subjects was illustrated in Table 4 which shows the values of Kendall's coefficient of rank correlation between model-predictions and observed ratings for each subject. It was hypothesized that MAUT is an appropriate model of auditor decision-making because of the personal preferences, risk attitudes, and multiple criteria which characterize professional judgments. The internal control setting encompasses these elements yet provides a manageable context for empirical testing. The generally high correlations between predicted and actual ratings indicate that MAUT is not only conceptually appropriate as a theory of auditor decision-making but is also empirically valid in this context for the subjects of the study.
The intent of the study was to provide preliminary evidence concerning the validity of MAUT as a model of auditor decision-making and it successfully accomplishes this task even though the research hypotheses were not upheld by statistical testing. Further research should be undertaken to test the ability of MAUT to model the decisions auditors make during other phases of the audit process (e.g. evaluating other elements of internal control systems, program planning, and analytical review) to determine the full extent of the appropriateness of MAUT in auditing. Cross-firm studies of auditors must be made to provide the basis for conclusions about the generalizability of the theory in the auditing profession.

5.1.2 Performance of MAUT Relative to Other Models

The ability of MAUT to model auditor decision-making in the described context was compared to two models which incorporate only portions of MAUT. The equally weighted additive model is simply the sum of the conditional utilities for the internal control system attributes represented by compliance test results. It employs the elicited utilities, but does not consider the scaling constants nor the independence conditions of the complete theory and therefore ignores the impact of changes in one attribute's value on the other attributes and does not account for combinations
of attribute values as prescribed by the theory. The scale weighted additive model goes one step beyond the equally weighted model to incorporate the scaling constants elicited from subjects but assumes for all subject that attributes are additively independent. (That is, preferences for lotteries are based on marginal probabilities only.) The MAUT model incorporates the impact of changes in one attribute on the others through its functional form. Combining this element of the theory with the von Neumann and Morgenstern utilities for attributes and scaling constants, MAUT establishes a model which more closely reflects the complexities of decision-making under uncertainty faced by professionals. It was consequently hypothesized that the degree of predictive ability would increase with increases in model complexity, the MAUT model representing the highest level of complexity studied.

This hypothesis was tested by comparing the rank correlations between model predictions and subject ratings using a nonparametric two-way layout. (Table 5 shows these correlations.) The results showed no statistical support for the hypothesis when the data for all 15 subjects were analyzed. On a subject-by-subject basis, analysis of the data from six of the 15 auditors showed correlations in the hypothesized direction and for only one subject was there a complete reversal of the hypothesized relationships. The
correlations for nine subjects were in some other non-hypothesized order. MAUT can model decision-making behavior better than the simpler models, studied here, evidenced by data obtained from Subjects 4, 8, 10, 11, 12, and 13. While MAUT may not be appropriate as a model of internal control evaluations for all auditors, there is support for its validity in some cases.

5.2 **DISCUSSION REGARDING INDIVIDUAL SUBJECTS**

5.2.1 **Subject 1**

Subject #1 (S1), a Senior Manager with eight years experience with the firm is one of the four subjects who specified the same scaling constant for each attribute. As a result of this commonality, the equally weighted and scale weighted additive models ranked cases identically and thus had the same degree of rank correlation with subject ratings. These models had a higher correlation with subject ratings than did the MAUT model (.99817 versus .94339).

This subject responded very quickly to the questions asked by the researcher during elicitation of the MAUF. S1 gave "5" as a response to all questions which required a subjective probability. (Such a response during the elicitation of conditional utilities indicates risk neutrality with respect to that attribute.) The subject did not wait for the researcher to elicit utilities for each attribute.
In responding to the first question asked in eliciting up("1 error") the subject added "...and if you ask me that question for the other factors I'll say .5 as well." When the researcher posed the questions required to elicit the individual scaling constants, which ask subjects to make preference statements for lotteries with multiple attribute values specified, the subject again responded " .5 for all the factors we're looking at." S1 added that none of the attributes were more important than the others which was the reasoning for specifying the same scaling constants for all attributes (even though Keeney and Raiffa (1976) show that the scaling constants are not importance weights).

The statements by the subject that all the factors were the same, combined with a potential influence of the probability elicitation in the simple lotteries for the conditional utility functions, may indicate the reason the .5 was given for scaling constants. Had the subject not been so quick to respond to the researcher's questions, the answers may have been closer to .25, leading to the formulation of an additive MAUP. The subject dismissed questions asked during a consistency check by stating with virtually no consideration of the questions that the same responses should be maintained.

The responses given by subject #1 and the statements made during the elicitation process indicates that an
additive model may have been a more appropriate functional form of the MAUF. Had this been warranted by the subjective probabilities explicitly given by the subject, the three models used for comparison would have produced identical correlations with the subject ratings.

5.2.2 Subject 2

Subject #2 (S2), a Supervising Senior with four years experience in the firm, appeared to take the experimental task more seriously than did the first subject. S2 stated that the specification of "0.5" for the conditional utilities was a "gut reaction" and was "...supported by the idea of expected value." The subject specified "0.25" as the probability related to scaling constants for three of the four attributes but gave a probability of "0.3" related to Population Completeness. This factor was viewed as more important by the subject and thus needed a higher probability to satisfy his preference structure (Again, the subject interpreted the probability specified as an indication of relative importance, an interpretation Keeney and Raiffa contend is not validly attributed to scaling constants.)

The sum of the scaling constants elicited from Subject #2 (0.30, 0.25, 0.25, 0.25) is 1.05, very close to 1.00, the number needed to indicate that an additive utility function would be the most appropriate form of the MAUF. The scale
weighted additive model and MAUF constructed with the scaling constants elicited both predicted better than the equally weighted model, but the scale weighted additive model performed slightly better than the hypothesized MAUT model. (See Table 5) This finding indicates that the elicited probabilities should not be viewed as error-free (since they are subjectively-given) and that a range of values around 1.00 for the sum of the $k_i$ should be deemed appropriate for using an additive model.

5.2.3 Subject 3

Subject #3 (S3) is a senior with two years of auditing experience. S3 was very cooperative through both phases of experimentation and stated the reasons for the answers given during MAUF elicitation. That is, the specification of probabilities used to determine scaling constants were based on subject perceptions of the importance of each attribute. The results of correlating this subject's case ratings with predictions made using the utility functions were consistent with the contention that internalization of risk and attribute-implications on system reliability is greater with increased years of experience (rank). This subject's correlation coefficients were the lowest of all subjects for the scale weighted and MAUI models and this subject had the fewest years of audit experience.
The correlation between Subject #3's ratings and the equally weighted utility function was the highest among the three models compared, and the correlation with the quasi-additive MAUF was the lowest (though all were highly significant). This finding may also be attributed to the subject's lack of experience.

There is no objective evidence in the data collected to indicate why the additive models performed better than the quasi-additive MAUT model. Since this subject had only two years of audit experience, the lack of conformance to the second research hypothesis may be attributed to the experimental task and its strain on the subject's cognitive processes. In two years, the audit senior may not have evaluated 42 internal control systems and this fact could have contributed to producing the results. In evaluating the 42 cases presented to the auditors in Phase II of the experiment, Subject #3 used a heuristic to simplify the rating task by first separating the compliance test cases by number of errors without regard to attributes and focused on cases containing two or three "0 error" results. The cases were then ordered by perceived strength and ratings assigned accordingly. Such a decision rule may resemble more closely an additive model than a multiplicative model and thereby account for the results observed.
5.2.4 Subject 

Subject #4 (S4) is a Supervising Senior with three years experience as an auditor. This subject considered carefully the questions asked during Phase I of the experiment and based responses to requests for probabilities on perceptions of importance. Subject #4 made statements to that effect during both the elicitation of conditional utilities and scaling constants. S4 further stated the problem of combining results of several compliance tests was a very real part of the decision-making being done in the planning phase of the audit to which S4 was assigned at the time.

The correlation of Subject #4's ratings with the predictions made using the quasi-additive utility function (.86107) was the greatest among the three models compared. The MULT model is therefore concluded to be the most appropriate of the three models to model the evaluations of compliance tests by Subject #4. No evidence is available to determine the reason why the equally weighted model performed better than the scale weighted model (.85836 and .81622 respectively). The cause of the relative superiority of the equally weighted model may be an artifact of the mathematics.
5.2.5 **Subject 5**

Subject #5 (S5) is a Supervising Senior with three years experience as an auditor. Subject #5 is the second of the four auditors who specified scaling constants which were the same across attributes. This subject also displayed risk neutral characteristics by specifying "0.5" as the conditional utility for each "1 error" consequence. During the elicitation of the MAUF, Subject #5 stated more than once that the factors were all equally important and that each attribute should be "consistently" evaluated. (e.g. One error in a test for Population Completeness should be viewed the same as one error in a test for Mathematical Checks, according to the auditor.)

This subject evaluated the cases presented in Phase II of the experiment in accordance with the statements made in Phase I. The same ratings were given to all cases which had the same set of numerals, regardless of the compliance test to which they were assigned. (e.g. 1, 0, 2, 1 was rated the same as 2, 0, 1, 1 the same as 2, 1, 1, 0.) Such a rating scheme is consistent with and was predicted by the two additive models for Subject #5, both of which were equally weighted. Since this auditor was perfectly consistent in the subjective evaluations, a perfect 1.00000 rank correlation resulted between the subjective ratings and the predictions of the two (equally weighted) additive utility models.
The correlation between the subjective ratings given by Subject #5 and the MAUF model was also extremely high (.99809). The subject clearly used a decision rule characterized by an additive model which accounts for the slightly lower coefficient of rank correlation between the observed ratings and the predictions of the quasi-additive MAUF. If the subject would have given "25" as the probability used to derive the scaling constants common to all the attributes, the MAUF would have been additive and would yield predicted ranks identical to those given by the other two models. These ratings would have consequently correlated perfectly with the subject's ratings. The misspecification of the model should not be attributed to a deficiency of the underlying theory, but rather to the assumption that subjects can perfectly specify their subjective probabilities for hypothetical events. As indicated in the analysis of results obtained from subject #2 which showed the sum of the $k_i$ equal to 1.05, further research should be conducted to determine tolerance limits for subjective probabilities within which additive independence is warranted and additive MAUFs are appropriate characterizations of subject preferences.
5.2.6 Subject 6

Subject #6 (S6) is a Senior Manager with eleven years of experience as an auditor. With the participating C.P.A. firm. All three utility models predicted ratings of the hypothetical cases which correlated very highly with the observed ratings assigned by the subject. The scale weighted additive model out-performing the equally weighted additive model and the MAUT models (Kendall's tau related to these models are .86759, .85472, and .84815 respectively.) The reason for the relative predictability of the three models is not clear from the data collected. While it is intuitively appealing that the scaling constants lead to greater predictive ability between the additive models, analysis of the scaling constants (kP=.4, kA=.5, kC=.4, and kM=.4) and the comments made by the subject does not reveal the reason for the superiority of both additive models over the quasi-additive model.

Subject #6 appeared to be uncomfortable specifying probabilities for a hypothetical client and repeatedly stated that "it all depends on the individual client." S6 also oscillated during elicitation of conditional utilities what "it should be" and what the subject "felt it was." This hesitancy was also related to the generic nature of the hypothetical client. This subject commented that the researcher would probably find different results for the
younger auditors than for more experienced auditors because of the shift away from a highly judgmental audit style, a further indication that evaluations are dependent on an intimate knowledge of the client and cannot be completely captured in a model which is appropriate for all situations. It may thus be concluded that a single MAUT model is not an appropriate way to characterize this auditor's behavior in evaluating internal control systems even though fairly accurate predictions may be made.

5.2.7 Subject 7

Subject #7 (S7) is a Supervising Senior with three and one-half years of auditing experience. The rank correlation between the subject's ratings of the hypothetical cases and the predictions made using each of the three utility models indicates that the scale weighted additive model best predicts this subject's evaluations of compliance test results. (tau=.74708) The MAUF produced the second-best correlation (tau=.72562) and the equally weighted model yielded the worst predictions of the three (tau=.67266), although all correlation coefficients were highly significant.

These findings imply that a compensatory decision rule is being used by Subject #7 in the evaluation of transaction controls. Little more can be concluded with respect to Subject #7 as this auditor appeared to have no difficulty
specifying probabilities requested during MAUF elicitation nor did the subject express any problems encountered in evaluating the compliance test results. Due to the number of cases to evaluate and the relatively limited amount of experience with internal controls evaluation (3-1/2 years), the subject may have simplified the decision rule to accommodate the task given in Phase II. The subject's decision rule in the case evaluation phase may have resembled more closely the scale weighted additive model than the complex quasi-additive MAUF because of a "need" to simplify the task.

5.2.8 Subject 8

Subject #8 (S8) is a Senior Manager with eight years of auditing experience and is the third of four subjects who specified the same scaling constant for all attributes (ki = .41). Unlike the other three subjects so characterized, the subjective ratings given by this auditor are most highly correlated with the predictions of the quasi-additive MAUF. The MAUT model is therefore the most appropriate of the models examined to predict this subject's evaluations of compliance test results.

Several factors may have contributed to the finding that the MAUF was the best model of the evaluations made by Subject #8. This subject had far more formal education than did most of the other participants ("All But Dissertation"
- ABD, Carnegie-Mellon University) and was very familiar with probability theory and its applications. The probabilities specified by this subject may be expected to reflect less bias than probabilities given by subjects who are inexperienced in probability assessment. Consequently, the model constructed in Phase I of the experiment may reflect this subject's preferences more accurately than the models constructed for other subjects reflect their preferences.

The level of professional experience possessed by this subject combined with the subject's stated extensive experience as a subject in research projects could have helped the subject evaluate hypothetical cases as actual client systems would be evaluated. Whereas subjects with relatively low levels of experience may need to simplify the task to make it feasible in light of their stored experiences, S8 may have less need to use simplification as this subject has a large pool of experiences to draw upon for help in the evaluation task. It is likely that the educational experiences, especially at the doctoral level, helped S8 appreciate the need for serious consideration of the task on the part of subjects involved and accordingly gave the best answers that could be given.3 Finding the hypothesized relationships in the case of Subject #8 and not

---

3 While it is assumed that all subjects answered honestly and to the best of their ability, human error must be considered in analyzing the results obtained.
finding such a relationship for other subjects may be attributable to a more accurate model and "cleaner" data because of the subject's higher degree of familiarity with academic research and its objectives.

5.2.9 **Subject 9**

Subject #9 (S9) is a Senior Manager with ten years of auditing experience. The ratings assigned to the hypothetical compliance test results by S9 were most highly correlated with predictions made using the scale weighted additive model (tau=.74463). The predictions made with the quasi-additive MAUP were the second-best predictions (tau=.71415). These findings indicate that the level of complexity added by the scaling constants aids in modeling the evaluations made by Subject #9.

The fact that the MAUP was not the best predictor may be partially explained by the method and results of the subjective evaluation. Subject #9 assigned ratings to each set of compliance tests as they appeared in the shuffled deck of 3"x5" index cards. S9 did not compare the cards and the ratings thereto assigned to ensure consistency in the evaluations. Consequently at least one of the assigned ratings does not conform to the von Neumann and Morgenstern "Complete Ordering Axiom" (Schoemaker, 1980) (e.g. "0, 1, 1, 2" was assigned a rating of 76, whereas "0, 1, 1, 1" was
assigned a rating of 75.5. The latter should be preferred to the former.) This inconsistency and other errors in the data caused by cognitive limitations may account for the slight superiority of the additive model over the quasi-additive model.

Concluding that the additive model is most appropriate due to the closeness of the sum of the scaling constants to one is not warranted since this sum is equal to 1.7 (0.5+.4+.4+.4). Hence the data allow only the conclusion that MAUF is a good predictor and that errors in the observed values may prevent definitive conclusions about the relative predictive accuracy of the models studied.

5.2.10 Subject 10

Subject #10 (S10) is a Senior Manager with six years of experience as an auditor. During Phase I of the experiment, this subject initially stated that as long as all the compliance test results were acceptable according to firm policy, then no differences existed among results conforming to the prescribed criteria. After the researcher explained the need to have the subject express personal viewpoints, Subject #10 acted according to his/her own perceptions. The MAUF elicited from this auditor yielded predictions of case evaluations which are more highly correlated with subject-given ratings than are the predictions
from the simpler utility models, just as hypothesized. 
\( \tau(EW,S) = .64198, \quad \tau(SW,S) = .76777, \) and 
\( \tau(MAUT,S) = .77749. \) It may be concluded that for Subject 
\#10, the quasi-additive MAUF is the most appropriate model 
of the subject's evaluation of compliance test results from 
among the models studied.

5.2.11 Subject 11

Subject \#11 (S11) is a Supervising Senior with three 
years of auditing experience. The subject was very familiar 
with probability theory and asked whether or not the 
elicitation procedures were designed to determine how con-
servative the subject was. The familiarity and comfort 
with probabilities could have contributed toward specifica-
tion of a MAUT which was an accurate characterization of 
the subject's preferences. The correlation between the rat-
ings given by Subject \#11 and the predictions made using 
the MAUF of this subject \( \tau(MAUT,S) = .72314 \) was the high-
est of the three models. Furthermore, the hypothesized re-
lationship between model complexity and predictive ability 
was found in the data related to this subject. It can 
therefore be concluded that the evaluations of transaction 
control compliance tests made by Subject \#11 are more accu-
rately predicted by the MAUF than by the simpler models em-
ployed in this study.
5.2.12 **Subject 12**

Subject #12 (S12), a Supervising Senior with three years of auditing experience, specified conditional utilities and scaling constants to characterize a MAUF which yielded predictions that are more highly correlated with subject-assigned ratings than either of the simpler utility models to which it was compared. \( \text{tau}(\text{EW}, S) = .46044, \) \( \text{tau}(\text{SW}, S) = .53349, \) and \( \text{tau}(\text{MAUT}, S) = .61447 \) This finding is consistent with the hypothesis and hence it may be concluded that the MAUF is the most appropriate utility model to predict the evaluations of transaction control compliance tests made by Subject #12.

5.2.13 **Subject 13**

Subject #13 (S13) is a Senior with two and one-half years of auditing experience. He was very interested in the experimental task and exhibited quite an extensive knowledge of probability theory and expected value theory. During elicitation of conditional utilities and scaling constants, the subject spoke aloud and stated that he/she was "calculating expected error rates and evaluating their acceptability." This familiarity and use of probabilistic concepts may have contributed to the resulting correlation between the subject's ratings and the predictions of the MAUF (\( \text{tau} = .68580 \)) which was higher than the correlations
related to the other utility models (\(\tau(EW, S) = 0.44896\) and \(\tau(WS, S) = 0.62390\)). One may conclude that of the models examined, the MAUF best characterizes this subject's preferences in the context of the study.

5.2.14 **Subject 14**

Subject 14 (S14) is a Supervising Senior and has worked as an auditor for four years. S14 was the fourth subject to specify the same scaling constant for all attributes in the MAUF (\(k_i = 0.1\)). Because of this specification, the equally weighted and scale weighted additive utility functions yield identical rank predictions for case ratings and their Kendall's coefficient of rank correlation between these predictions and observed ratings is the same for both models (\(\tau = 0.95759\)). Furthermore, these models predicted subject ratings better than the quasi-additive MAUF (\(\tau = 0.83397\)).

Statements made by the subject during interviews with the researcher made definitive conclusions impossible. The subject's response to questions asked during the initial elicitation of scaling constants was:

"As long as there is any chance of getting something other than the worst acceptable situation, I would prefer the alternative giving me that chance."

Consequently "p=0" was given as a response for all questions related to scaling constants. This resulted in a de-
generate MAUT and predictions could not be made except to presume that the subject would perceive no differences among the 42 compliance test results presented to him for evaluation.

A Phase II interview was scheduled with Subject #14 to verify the behavior implied by the degenerate MAUT. The subject was asked if ratings could be assigned to the 42 cases to indicate the reliability of the transaction controls. S14 stated it could. This contradicted the subject’s earlier statements and prompted the researcher to repeat the elicitation procedure. Upon doing so, the subject indicated that the indifference probability was actually not zero, but “something close to zero, like .1.” This statement was used to predict the ratings given by Subject #14.

The contradictory statements made by Subject #14 indicate that model misspecification or an inconsistent decision rule (or both) accounts for the results obtained. That is, MAUT should not be ruled out as a theory of this auditor’s decision-making behavior because of the apparent superiority of the simple models in predicting subject ratings. The impact of any errors in the data obtained from Subject #14 on the additive models may have been less than the impact on the quasi-additive model due to the latter model’s interaction terms, thereby causing the correlation results found. Since the axioms underlying MAUT assume the
decision maker acts consistently, conclusions regarding the appropriateness of MAUT as a model of Subject #14's preferences are tenuous.

5.2.15 **Subject 15**

Subject #15 (S15) is a Manager with five years of auditing experience. The model which produced predictions most highly correlated with ratings given to cases in Phase II by Subject #15 is the scale weighted additive utility model (tau=.75090). The MAUT performed second-best (tau=.71464). From these results it may be concluded that the scaling constants aid in modelling the auditor's preferences since both models which incorporate scaling constants predicted better than the equally weighted model (tau=.70588).

The superiority of the additive model over the quasi-additive model cannot be attributed to any objective data collected from interviews with Subject #15. It may be presumed that this subject, as other may have done, simplified the decision rule due to the fairly large number of cases to evaluate. As a result, this simpler evaluation method may more closely resemble the scale weighted additive model than the more complex quasi-additive model.
5.3 **IMPLICATIONS OF THE RESEARCH FOR THE AUDITING PROFESSION**

The research conducted was designed to reveal insights about the validity of multi-attribute utility theory as a model of auditor evaluations, not as a study of the quality of auditor decision-making procedures. The models constructed for each subject were used as characterizations of the auditors' preference structures in the context of internal controls evaluation; they were not considered decision aids nor were they tested to determine the potential of MAUPS as audit decision aids. Conclusions relating to the specific intent of this study were made in previous sections of this chapter. While the main focus of the research was on theoretical issues rather than professional issues, the findings of the study do have implications which may be important to advancement of the profession.

5.3.1 **Implications for Auditor Training**

Auditor training and professional development courses appear to be very effective in the firm which participated in this study. As stated in Chapter IV - "Experimental Results", all fifteen subjects specified the same four characteristics of a transaction control as the attributes they examine in making an evaluation of these controls. The four attributes are those specified by the firm in its training manuals and pre-printed working papers. All subjects were
thoroughly familiar with the procedures and knew how to implement them. Thus, it can be concluded that the firm's training program was effective in teaching the factors the firm's policy-makers deemed important in evaluating an internal control system.

The effectiveness of the firm's training program in one area of policy implies that other auditing procedures may be formalized thus creating greater objectivity in the auditing process. The MAUPS for each subject shown in Table 3 indicate variations among auditors in the combination rules used to aggregate the results of compliance test of transaction control reliability. If the firm wished to ensure objectivity in situations which require aggregation of multiple factors, a combination rule could be prescribed and communicated to employees effectively through training programs and auditing documents. It may be concluded that auditors will conform to procedures prescribed by the firm's officers. This conclusion has implications for auditing firms which desire more objectivity in their auditing procedures than they have currently.

5.3.2 Implications for Quality Control

The varied multi-attribute utility functions of the fifteen subjects of this study indicate that there exists the potential for differing audit opinions across auditors
within the same firm. The possibility for differing opinions indicates that the quality of audits may be difficult to control.

The subjects of this study all specified the same attributes of the internal controls being evaluated, those specified by firm policy. If subsequent studies show auditors actually do conform to explicitly stated firm policies can help control the quality of audits. By implementing objective procedures and documenting the auditors' activities and conclusions, the quality of audits may be monitored and controlled.

Auditors are considered professionals due in part to their expert knowledge of accounting systems and their ability to evaluate these systems. Auditors are human beings though, and cannot be expected to make subjective evaluations with the accuracy of a machine. This research did not attempt to isolate the causes for the observed differences in utility functions for the determinants of the subjective evaluations made, yet it can be expected that experience, personal risk attitudes, fatigue, and other factors do have an effect on auditor judgments (i.e. consistency and consensus). The finding that subjects do have different levels of risk tolerance (as evidenced by the conditional utilities in Table 3) and perceptions about internal control system reliability indicates that the effect
of these variations may need to be moderated by developing explicit procedures (where possible) for various auditing tasks to ensure audit quality control.

5.4 SUGGESTIONS FOR FURTHER RESEARCH

This research has provided preliminary findings concerning the applicability and validity of multi-attribute utility theory in an auditing context. An obvious direction to move from this first step is to study MAUT in other areas of auditing as well as other professional contexts. From the information gathered in studies which supplement and extend this one, greater insights into audit decision-making as well as decision-making in other professions. The evidence of this study relates to a small, non-random sample of subjects and pertains to only one phase of the audit process. It cannot be used as the basis for grand generalizations about the validity of MAUT.

Analysis of results on an individual subject basis revealed many differences in scaling constants, degree of the models' predictive abilities, and factors influencing evaluations. The implications of these differences should be investigated in hopes of finding the linkages among risk attitudes, training, experience, and the perceptions auditors have of the accounting systems they evaluate. Findings in these areas could have a significant impact on
the auditing profession: its training methods, documentation, hiring practices, and methods of auditing.

Expansion of the sample size to allow analysis of results across levels of experience as well as among C.P.A. firms could yield useful information related to quality control of audits and could provide data from which information concerning the determinants of risk attitude and preferences for multi-attribute alternatives may be garnered.

5.5 LIMITATIONS

The research was limited to studying the internal control system evaluations of 15 auditors in one C.P.A. firm, and within this setting only the transaction controls of the internal controls over credit sales were considered. These restrictions of context and subjects constrain the implications of the results to inferences about the behavior of the participants and do not allow statements of a general nature.

Without minimizing the fact that there is evidence against the hypothesis, the non-supportive results produced by the other nine subjects may be attributed to a variety of experimental limitations. One cause for findings which contradict the second research hypothesis is "model misspecification" due to the unfamiliar task of specifying
subjective probabilities on the part of the subjects. During elicitation of the MAUFs, subjects were asked to specify probabilities for outcomes in a lottery which would make them indifferent between the lottery and a certain consequence. While almost all subjects (80%) had some familiarity with probability theory, there could have been misspecification of the probabilities due to the fact that most subjects do not explicitly consider probabilities in the normal course of their work. Giving "wrong" probabilities would result in inappropriate conditional utilities and scaling constants. If a subject specified scaling constants which did not sum to one, the quasi-additive model was considered the appropriate form of the MAUF. If in fact the constants were misspecified and would have summed to one if properly given, the additive form is most appropriate. This situation could result in the scale-weighted additive model having the highest correlation with subject ratings if the subject actually used an additive decision rule.

The researcher attempted to ask questions in a manner which would not lead subjects to preordained answers. This was done by posing very general questions which would allow subjects to tell about their methods of internal control evaluations and their perceptions of the firm's prescribed procedures. Even though precautions were taken, model misspecification may have resulted from subjects' desire to
"cooperate" to the extent of telling the researcher "what he wanted to hear." This could have occurred at the attribute-specification stage of MAUT elicitation. All subjects stated that firm policy determined the attributes they examined during internal control evaluations, but if in reality the subjects considered other cues in their assessments, there is no reason to expect the MAUT model to predict ratings better than the simpler models. Consequently, either (or both) of the simpler models may produce ratings more highly correlated with subject ratings than the MAUT model simply because the auditor uses a decision rule which does not correspond to the MAUT model elicited (nor the other utility models used for comparison purposes).

The aforementioned possible causes for non-supportive correlations between model predictions and subject-given ratings assumed that subjects used a multi-attribute utility theory model in making their evaluations of compliance test results during the second phase of the experiment but that the MAUT model elicited was not the model they used internally. An additional reason for lack of support for the hypothesized relationships among the correlations rests with the complexity of the experimental task performed by the subjects. Evaluating 42 sets of compliance test results may have overwhelmed the subjects, many of whom may not
have evaluated that many internal control systems in their entire careers (particularly Seniors and Supervising Seniors). Some subjects have only a few large clients and therefore may not have had the opportunity to perform many internal control system evaluations. While the researcher did not attempt to collect hard evidence on the effect of fatigue or mode of case presentation, one may speculate that the relatively large number of cases may have compelled the subjects to use some heuristic for evaluating the cases. Such a simplification may not be used in situations in which subjects are confronted with just a few test results to evaluate, such as an actual audit engagement. The heuristic used may more closely resemble a simple additive model than the quasi-additive decision rule of the MAUT and consequently the subjects' ratings would be more highly correlated with the predictions of the simpler models. (Future research efforts should include controls to prevent such an occurrence in order to more closely simulate actual audit decision-making.)

The speculations made here concerning the reasons for disconfirming evidence imply that further research should be conducted before definitive conclusions concerning the use of MAUT by auditors are made. The limitations cited should be remedied and closer scrutiny of auditor decision processes should be undertaken. Because of the limitations
of the study and the conflicting results across subjects, it cannot be said that the hypothesis is invalid but rather that preliminary results were obtained that indicate that MAUT is empirically valid in an auditing context.

5.6 SUMMARY

The results of the experiment conducted to test the validity of MAUT as a model of auditor preferences in an internal control setting provide preliminary evidence of empirical validity of the theory. The Multi-Attribute Utility Functions elicited from each subject demonstrated a high degree of predictive ability.

Analysis to compare the predictive power of the MAUT model to simpler utility models revealed no statistical differences.

Since MAUT is an individual preferences model, examination of results on an individual subject basis provides insights which cannot be obtained from hypothesis test of aggregate data. Subject-by-subject analysis indicates a wide variety of utility functions and potential causes for such characterizations of auditor preferences. While the aggregate results indicate MAUT is the most appropriate model for only 40% (6 out of 15) of the subjects studied, it appears that task complexity and the rigidity of elicitation procedures may have confounded the experiment such
that the MAUT model did not perform as well as it may otherwise have.

In general, MAUT does appear to be a good model of the internal control evaluations made by subjects in this study although further research is necessary to determine the full extent of its predictive ability in other auditing contexts, descriptive validity as a model of auditors' cognitive processes, and its potential as a decision aid to auditors.
Appendix A

INTERNAL CONTROL EVALUATION POLICY OF A NATIONAL C.P.A. FIRM

One C.P.A. firm supplied the fifteen subjects who participated in the research conducted for this dissertation. The procedures for evaluating internal control systems followed by the firm are described in this appendix. Based on the procedures described, a multi-attribute "utility" model is constructed to illustrate how the firm's policies may be incorporated in the MAUT model. This "firm model" is described in the following appendix.

The Policies

The evaluation of an internal control system involves several phases. One of the "Big Eight" C.P.A. firms begins with an evaluation of the firm's operating environment which is aimed at assessing the degree to which management creates an environment (both physically and in terms of attitude) which is conducive to maintaining a system of controls over the accounting system. This phase includes the documentation, via flowcharts, of transaction points and the controls which are related to these events. The flow-
charts also document the points in the accounting system where the data change form and the control points which should be examined with respect to these events.

The second phase employed by the example firm calls for an evaluation of each transaction control over the transactions with parties outside the corporate entity. An internal control worksheet is prepared for each transaction control documented in the preliminary phase of the internal control evaluation. The worksheet requires documentation of the following items:

1. **Sources of information** - the source documents of the exchange transaction which can be traced directly to the account balance.

2. **Controls for data completeness** - controls aimed at ensuring that all data are captured.

3. **Controls to ensure proper transaction authorizations** - controls to ensure that bogus transactions are not included in the population.

4. **Controls to ensure accuracy of comparisons** - controls to ensure that orders are filled properly, billed to correct party, etc.

5. **Controls to ensure existence of mathematical accuracy** - checks on footings, extensions, etc.

6. **Segregation of duties** - assesses the impact of segregation (or lack thereof) on the audit program.

7. **References to the audit program** - ties the internal control worksheet to compliance test worksheets and compensating tests (when compliance testing is not performed; done during substantive testing).

8. **Conclusions** - summarizes the findings of compliance and/or compensating tests and their impact on the remainder of the audit program.
Within each of these subsections of the questionnaire are questions related to the reliability of the control procedures. The questions ask for yes/no responses as well as "how," "what," "when," and "by whom" type questions. There is generally a question to summarize each section which asks: "Effective if properly performed?" The answer to this question will have an effect on the type of testing that will be done later in the audit.

Evaluation is performed on those population and accuracy controls which are judged to be 1) effective if properly performed, 2) missing, or 3) ineffective if properly performed. For those controls which the auditor feels would be effective if properly performed, the auditor plans for compliance or compensating tests based on an informal cost/benefit analysis. Boundary controls which are documented are always compliance tested by this firm. For the controls which are missing or judged to be ineffective if properly performed, the firm policy specifies that compensating test should be undertaken. This form of testing is also conducted on controls which are shown to be ineffective as indicated by compliance testing.

Subjects' preferences over quantitative factors are presumed to be more accurately captured by the MAUP than are their preferences for qualitative (subjectively scaled) factors. The multi-attribute utility function is a
mathematical model of an individual's preference structure. Quantitative variables fit readily into the model whereas a subjective scaling would need to be developed to quantify a qualitative variable. Since compliance testing allows for some quantitative factors (e.g. number of irregularities in a compliance test) to be used in the internal control system evaluation, it seemed important that the experimental task of this study involve the evaluation of compliance test results.
Appendix B

FIRM MODEL OF INTERNAL CONTROL EVALUATION

Assumptions (Related to procedures of the internal controls evaluation)

1. Assume the firm's control environment has been reviewed and given a satisfactory rating.

2. Assume the internal control worksheet is completed (to the point of cross-referencing to compliance test worksheets) and the compliance tests are to be performed for each control documented.

3. Moderate reliance on the transaction controls is planned. A sample of size 40 for compliance testing has consequently been chosen. (An audit Manager in the firm stated this is consistent with the firm's policy for testing boundary controls.)

The compliance test worksheet provides a guideline for attributes to be tested and measures to be assigned. From a discussion with an audit Manager in the firm whose procedures are herein described, the researcher learned that subjects could be expected to be fairly consistent in the attributes they would specify as being important to evaluating the internal control system. (In fact this was found to be true.) This is due to firm training and audit guidelines provided by the firm's management.

Assumptions: (Related to the MAUT model)
1. The C.P.A. firm is risk neutral with respect to all factors of the internal controls evaluation.

2. Additivity of individual utility functions is warranted.

3. Equal weighting of attributes is appropriate.

**The Model**

\[ U = .25u_P(x_P) + .25u_A(x_A) + .25u_C(x_C) + .25u_M(x_M) \]

where

- \( x_P \) = number of irregularities in compliance test sample related to POPULATION COMPLETENESS
- \( x_A \) = number of irregularities in compliance test sample related to AUTHORIZATION REVIEWS
- \( x_C \) = number of irregularities in compliance test sample related to COMPARISONS ACCURACY
- \( x_M \) = number of irregularities in compliance test sample related to MATHEMATICAL CHECKS
- \( U \) = overall utility of the boundary Controls over Sales
- \( u_i(.) \) = conditional utility function of the C.P.A. firm for attribute \( i \) (where \( *i* \) is \( P, A, C, \) or \( M \))

The range of the values for each attribute/control type is the same for \( P, A, C, \) and \( M \) according to the compliance test worksheet. This document prescribes rating points based on the number of deviations found in compliance tests for various sample sizes.

The compliance test worksheet provides the following interpretation of error frequency (for a sample of size 40):
<table>
<thead>
<tr>
<th># of errors</th>
<th>Rating Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

The ratings indicate the level of restriction of substantive testing that can be done, based on the compliance test worksheet and the degree to which the auditor plans to rely on the internal controls. (In the case of transaction controls, the firm’s policy is to plan for a moderate degree of reliability, requiring at least two rating points.)

While it is feasible to have up to 40 errors in a sample of size 40, the company specifies that any number of deviations beyond 2 out of 40 indicates the control cannot be relied upon without extending tests. Consequently the range of errors which indicate a control is reliable is: 0–2. (If any attribute/control type is found to have more than two errors, the transaction control cannot be relied upon to the extent planned. The other compliance tests cannot compensate for a deficiency of such a magnitude in an attribute/control type.)

Since all attributes in the hypothetical model are identically scaled, the "utility" function for one attribute will serve as an example for all the conditional utility functions.
The utility of "number of deviations" is a discrete-valued function (as there cannot be fractional errors).

Using $M$ (MATHEMATICAL CHECKS) as an example, define $u_M("2\ errors") = 0$  
$u_M("0\ errors") = 1$

as these are the worst acceptable and best possible outcomes respectively.

Assuming the C.P.A. firm is risk neutral,  
$\Delta utility\ of\ #\ of\ errors = c = .5$ (in this example)  
$\Delta #\ of\ errors$

$\begin{array}{c|c}
\hline
\text{Utility} & \text{# of errors in compliance test of Mathematical Checks (n=40)} \\
\hline
1.0 & \\
.75 & \\
.5 & \\
.25 & \\
0 & \\
\hline
\end{array}$
The utility functions for the remaining three attributes would be identical to that above except that the horizontal axis would be relabelled appropriately to correspond to the attribute measures. Since an equally weighted additive model was assumed appropriate for the risk neutral C.P.A. firm, no elicitation of scaling constants is necessary.
Appendix C

INSTRUMENT TO TEST PREFERENTIAL INDEPENDENCE

Let:

\[ P = \# \text{ of errors in a compliance test of Population Completeness} \]

\[ A = \# \text{ of errors in a compliance test of Authorizations Review} \]

\[ C = \# \text{ of errors in a compliance test of Comparisons Accuracy} \]

\[ M = \# \text{ of errors in a compliance test of Mathematical Checks} \]

Assume: \( C = 2 \)
\[ M = 1 \]

For each pair of alternatives below, circle the more favorable compliance test report.

1A: (a) \( P = 1 \) error, \( A = 2 \) errors
(b) \( P = 1 \) error, \( A = 1 \) error

2A: (a) \( P = 1 \) error, \( A = 2 \) errors
(b) \( P = 2 \) errors, \( A = 0 \) errors

3A: (a) \( P = 2 \) errors, \( A = 0 \) errors
(b) \( P = 1 \) error, \( A = 1 \) error

***************

Now, assume: \( C = 1 \)
\[ M = 0 \]
Again, circle the more favorable report in each pair of alternatives below.

1E: (a) P = 1 error, A = 2 errors 
    (b) P = 1 error, A = 1 error 

2E: (a) P = 1 error, A = 2 errors 
    (b) P = 2 errors, A = 0 errors 

3E: (a) P = 2 errors, A = 0 errors 
    (b) P = 1 error, A = 1 error 

***************

Now, Assume: C = 1 
    M = 2

Again, circle the more favorable report in each pair of alternatives below.

1C: (a) P = 1 error, A = 2 errors 
    (b) P = 1 error, A = 1 error 

2C: (a) P = 1 error, A = 2 errors 
    (b) P = 2 errors, A = 0 errors 

3C: (a) P = 2 errors, A = 0 errors 
    (b) P = 1 error, A = 1 error 

***************

Assume: A = 1 
    M = 2

Circle the more favorable compliance test report.

4A: (a) P = 1, C = 2 
    (b) P = 2, C = 0 

5A: (a) P = 2, C = 0 
    (b) P = 1, C = 1 

6A: (a) P = 1, C = 2 
    (b) P = 1, C = 1 

***************
Now Assume:  \( A = 0 \)
\[ M = 2 \]

Circle the more favorable compliance test report.

**4E:**
(a) \( P = 1, \ C = 2 \)
(b) \( P = 2, \ C = 0 \)

**5B:**
(a) \( P = 2, \ C = 0 \)
(b) \( P = 1, \ C = 1 \)

**6E:**
(a) \( P = 1, \ C = 2 \)
(b) \( P = 1, \ C = 1 \)

**************

Now assume:  \( A = 2 \)
\[ M = 1 \]

Circle the more favorable compliance test report.

**4C:**
(a) \( P = 1, \ C = 2 \)
(b) \( P = 2, \ C = 0 \)

**5C:**
(a) \( P = 2, \ C = 0 \)
(b) \( P = 1, \ C = 1 \)

**6C:**
(a) \( P = 1, \ C = 2 \)
(b) \( P = 1, \ C = 1 \)

**************

Assume:  \( A = 1 \)
\[ C = 2 \]

Circle the more favorable compliance test report.

**7A:**
(a) \( P = 1, \ M = 1 \)
(b) \( P = 2, \ M = 1 \)

**8A:**
(a) \( P = 2, \ M = 0 \)
(b) \( P = 1, \ M = 2 \)

**9A:**
(a) \( P = 0, \ M = 2 \)
(b) \( P = 1, \ M = 1 \)

**************
Now Assume: \( A = 0 \)
\[ C = 1 \]

Circle the more favorable compliance test report.

7B:  
(a) \( P = 1, M = 1 \)  
(b) \( P = 2, M = 1 \)

8B:  
(a) \( P = 2, M = 0 \)  
(b) \( P = 1, M = 2 \)

9B:  
(a) \( P = 0, M = 2 \)  
(b) \( P = 1, M = 1 \)

************************

Now Assume: \( A = 2 \)
\[ C = 1 \]

Circle the more favorable compliance test report.

7C:  
(a) \( P = 1, M = 1 \)  
(b) \( P = 2, M = 1 \)

8C:  
(a) \( P = 2, M = 0 \)  
(b) \( P = 1, M = 2 \)

9C:  
(a) \( P = 0, M = 2 \)  
(b) \( P = 1, M = 1 \)

************************

Assume: \( P = 0 \)
\[ M = 1 \]

Circle the more favorable compliance test report.

10A:  
(a) \( A = 2, C = 1 \)  
(b) \( A = 1, C = 1 \)

11A:  
(a) \( A = 1, C = 2 \)  
(b) \( A = 2, C = 0 \)

12A:  
(a) \( A = 0, C = 2 \)  
(b) \( A = 1, C = 1 \)

************************

Now Assume: \( P = 1 \)
\[ M = 2 \]

Circle the more favorable compliance test report.

10B:  
(a) \( A = 2, \ C = 1 \)  
(b) \( A = 1, \ C = 1 \)

11B:  
(a) \( A = 1, \ C = 2 \)  
(b) \( A = 2, \ C = 0 \)

12B:  
(a) \( A = 0, \ C = 2 \)  
(b) \( A = 1, \ C = 1 \)

**************************

Now Assume: \( P = 2 \)  
\( M = 1 \)

Circle the more favorable compliance test report.

10C:  
(a) \( A = 2, \ C = 1 \)  
(b) \( A = 1, \ C = 1 \)

11C:  
(a) \( A = 1, \ C = 2 \)  
(b) \( A = 2, \ C = 0 \)

12C:  
(a) \( A = 0, \ C = 2 \)  
(b) \( A = 1, \ C = 1 \)

**************************

Assume: \( P = 2 \)  
\( C = 1 \)

Circle the more favorable compliance test report.

13A:  
(a) \( A = 2, \ M = 1 \)  
(b) \( A = 0, \ M = 2 \)

14A:  
(a) \( A = 1, \ M = 1 \)  
(b) \( A = 1, \ M = 2 \)

15A:  
(a) \( A = 0, \ M = 1 \)  
(b) \( A = 2, \ M = 0 \)

**************************

Assume now: \( P = 1 \)  
\( C = 0 \)
Circle the more favorable compliance test report.

13B:  (a) \( A = 2, \ M = 1 \)  
      (b) \( A = 0, \ M = 2 \)

14B:  (a) \( A = 1, \ M = 1 \)  
      (b) \( A = 1, \ M = 2 \)

15B:  (a) \( A = 0, \ M = 1 \)  
      (b) \( A = 2, \ M = 0 \)

***************

Assume now, \( P = 1 \)  
\( C = 2 \)

Circle the more favorable compliance test report.

13C:  (a) \( A = 2, \ M = 1 \)  
      (b) \( A = 0, \ M = 2 \)

14C:  (a) \( A = 1, \ M = 1 \)  
      (b) \( A = 1, \ M = 2 \)

15C:  (a) \( A = 0, \ M = 1 \)  
      (b) \( A = 2, \ M = 0 \)

***************

Assume: \( P = 0 \)  
\( A = 1 \)

Circle the more favorable compliance test report.

16A:  (a) \( C = 2, \ M = 0 \)  
      (b) \( C = 1, \ M = 2 \)

17A:  (a) \( C = 0, \ M = 1 \)  
      (b) \( C = 1, \ M = 0 \)

18A:  (a) \( C = 1, \ M = 1 \)  
      (b) \( C = 2, \ M = 0 \)

***************

Now, assume: \( P = 2 \)  
\( A = 2 \)

Circle the more favorable compliance test report.
16B:  
(a) $C = 2, \ M = 0$
(b) $C = 1, \ M = 2$

17B:  
(a) $C = 0, \ M = 1$
(b) $C = 1, \ M = 0$

18B:  
(a) $C = 1, \ M = 1$
(b) $C = 2, \ M = 0$

***************

Now, Assume: $P = 1$
$A = 2$

Circle the more favorable compliance test report.

16C:  
(a) $C = 2, \ M = 0$
(b) $C = 1, \ M = 2$

17C:  
(a) $C = 0, \ M = 1$
(b) $C = 1, \ M = 0$

18C:  
(a) $C = 1, \ M = 1$
(b) $C = 2, \ M = 0$

***************
Appendix D

CASE MATERIALS FOR EXPERIMENTAL TASK - PHASE II

The following test is a brief description of a hypothetical audit client. Following the text are 42 separate "reports" of compliance tests of transaction controls over sales.

Please read the text material and then consider each of the compliance test reports as if it was the result of work performed on the client described.

Each of the 42 reports is different. Evaluate each report in terms of its implications as to the reliability or strength of the internal control system (specifically, the transaction controls over sales). Assign a rating, on a 0-100 scale, to each of the reports to indicate your comfort with respect to each of the systems' reliability. ("0" indicates the least reliable system and "100" indicates the most reliable system.)

If you assign the same value to two or more reports, this is interpreted as meaning that you associate the same feeling of comfort with respect to this transaction control to each of the reports assigned this rating.

- 151 -
Please take whatever time you need to consider the reports. Your careful responses are vital to this study.
Thank you for your cooperation.

***************

Parker-Monroe, Incorporated is a large manufacturer of consumer durables and has been an audit client of your public accounting firm for several years.

The client's organization structure is well-defined and there are clear lines of communications between management and accounting policy makers. The accounting section is headed by a former audit manager who has strived to maintain an environment within Parker-Monroe that is conducive to an effective internal control system.

Due to Parker-Monroe's prominent position in the industry, it has been very successful at recruiting top-notch people throughout the organization. The hiring practices, compensation plans, and working environment have contributed to a low employee turnover rate.
The in-charge auditor has performed a preliminary review of the internal control system. Based on documented transaction controls she has decided to perform compliance tests of details on these controls over credit sales using a sample size of 40. Moderate reliance on the controls is planned. Four separate compliance tests will be performed to evaluate:

(1) population completeness;
(2) authorization reviews;
(3) comparisons accuracy; and
(4) mathematical checks

of the sales data.
Appendix E

SUBJECT DEMOGRAPHICS QUESTIONNAIRE

Your responses to the following items will assist the researcher in the analysis of data collected during the foregoing task. Your anonymity is guaranteed.

Please indicate the following

1. Age

2. Level of Education completed

   Bachelor's degree (A.B., B.A., B.S., B.F.A., etc.)
   Master's degree (M.A., M.S., M.B.A., M.Acctng., etc.)
   Doctor's degree (Ph.D., J.D., Ed.D., etc.)

3. Present rank in your C.P.A. firm

4. Years of service with your C.P.A. firm

5. Have you worked as an auditor for any other C.P.A. firm?

   If yes, how long?

6. Have you have any courses (training) in probability theory (expected value theory, utility theory, game theory, etc)?
Appendix Y

DOUBLE CROSS-VALIDATION AND ANALYSIS

A supplementary analysis was performed to determine whether a regression model based on utilities data would provide better predictions than a regression model based on nominal values. A double-cross validation technique was used. Green and Tull (1978) describe the technique and the information that can be derived from such an analysis. The 42 cases of compliance test results were randomly split into subsets containing 21 cases each. Multiple regressions were performed on each half, both for nominal data and utility data. The regression coefficients for one half of the nominal data were used to predict the ratings for the remaining 21 nominal-value data points. The predicted values for the second half of the nominal data set, using the coefficients of the first half regression is denoted y21(nom). The complementary predictions are denoted: y12(nom). The regression on utility-valued data were similarly used for predictions and these values are similarly denoted, y21(util) and y12(util). Table 6 presents the sum of squared differences between these predicted values and
their corresponding observed values (subject-given ratings).

**TABLE 6**

Sum of Squared Deviations-Cross Validation Regression

<table>
<thead>
<tr>
<th>S</th>
<th>Nominal Data (y21-s)</th>
<th>(y12-s)</th>
<th>Utilities Data (y21-s)</th>
<th>(y12-s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>507.900</td>
<td>522.313</td>
<td>507.818</td>
<td>522.408</td>
</tr>
<tr>
<td>2</td>
<td>3133.626</td>
<td>1246.327</td>
<td>1596.194</td>
<td>1316.862</td>
</tr>
<tr>
<td>3</td>
<td>1003.875</td>
<td>1098.791</td>
<td>822.268</td>
<td>803.324</td>
</tr>
<tr>
<td>4</td>
<td>2899.267</td>
<td>3047.879</td>
<td>6316.704</td>
<td>4698.060</td>
</tr>
<tr>
<td>5</td>
<td>327.255</td>
<td>285.097</td>
<td>330.300</td>
<td>227.168</td>
</tr>
<tr>
<td>6</td>
<td>1718.746</td>
<td>472.616</td>
<td>953.244</td>
<td>472.674</td>
</tr>
<tr>
<td>7</td>
<td>201.869</td>
<td>117.946</td>
<td>145.003</td>
<td>124.639</td>
</tr>
<tr>
<td>8</td>
<td>591.791</td>
<td>770.191</td>
<td>1278.269</td>
<td>935.796</td>
</tr>
<tr>
<td>9</td>
<td>186.766</td>
<td>237.104</td>
<td>186.724</td>
<td>240.293</td>
</tr>
<tr>
<td>10</td>
<td>152.680</td>
<td>186.303</td>
<td>327.145</td>
<td>280.013</td>
</tr>
<tr>
<td>11</td>
<td>265.529</td>
<td>208.409</td>
<td>706.227</td>
<td>830.912</td>
</tr>
<tr>
<td>12</td>
<td>95.960</td>
<td>112.772</td>
<td>330.987</td>
<td>381.909</td>
</tr>
<tr>
<td>13</td>
<td>363.025</td>
<td>269.629</td>
<td>1101.707</td>
<td>1310.839</td>
</tr>
<tr>
<td>14</td>
<td>83.987</td>
<td>77.999</td>
<td>84.830</td>
<td>77.967</td>
</tr>
<tr>
<td>15</td>
<td>1615.637</td>
<td>1074.208</td>
<td>995.677</td>
<td>1028.479</td>
</tr>
</tbody>
</table>

The hypothesis to be tested is:

**Ho:** \((y_{\text{util}}-s) \geq (y_{\text{nom}}-s)\)

**Ha:** \((y_{\text{util}}-s) < (u_{\text{nom}}-s)\)
The F-test value of .4 on 2 and 27 degrees of freedom is related to an alpha=.0713, indicating that Ho cannot be rejected at any reasonable level of significance. Based on this analysis it cannot be concluded that the utilities for the compliance test values form the basis for a regression model that results in predictions with less deviation from subject ratings than do the nominal values themselves. This finding is attributable to the mathematics of least squares regression which determines a "best fit" for the data used as input to the model.


