AN ATTENTION, TACTICS, OPERATIONS, SOLUTION, UNIQUE CONDITIONS GUIDED ASSESSMENT MODEL FOR TEACHING MATHEMATICAL PROBLEM-SOLVING IN SCHOOLS

A DISSERTATION

Presented in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy in the Graduate School

of The Ohio State University

By

Thomas V. Tinkler, M.A.

The Ohio State University

1998

*****

Dissertation Committee:

Dr. Ayres G. D’Costa, Adviser

Dr. Thomas R. Knapp

Dr. William E. Loadman

Dr. Sigrid Wagner

Approved by:

Adviser

College of Education
This dissertation is dedicated to my grandmothers
Margaret Tinkler and Rose Taaffe
ABSTRACT

One challenge facing advocates of guided assessment is the development of evaluative criteria that represent clear, significant, and useful levels of expertise. This dissertation develops and tests a guided assessment for evaluating mathematical problem-solving called the [A]ttention, [T]actics, [O]perations, [S]olutions, and [U]nique-conditions model. The five ATOSU attributes measure a hierarchical levels of both conceptual and procedural thinking viewed in mathematical problem-solving. Levels of the students' goal-directed verbal, visual, and mathematical activity are judged and rated under a limited number of problem conditions. The ATOSU approach asks teachers to use probing questions to direct problem-solving activity, to cue students when they are confused, and to analogically relate one mathematical topic to another mathematical topic or to a variety of "real world" topics. These criteria or rated attributes are considered useful when they can communicate a range of quality levels that, when analyzed by teachers and parents in decision making, provide insights into learning and feedback that increases performance. Although one study cannot possibly complete the task of validating such a model, this study makes some preliminary attempts at evaluating the quality of these attributes as clear benchmarks of conceptual and procedural growth.

Three sets of research questions are addressed using measurement methods, experimental methods, and qualitative methods. Since the ATOSU model is necessarily complex, it needs to be developed in three stages, answering one set of questions at each stage. In the first stage, the focus of this research is on the validity and reliability of the predicted conceptual and procedural hierarchies. In the second stage, a 2 x 3 factorial design compares the unique contributions of gender and three different treatments (Test-only, ATOSU assessment, and Tutored-only) on conceptual, procedural, and problem-solving
scales of National Assessment of Educational Progress (NAEP). In the final stage, snippets of qualitative data and the students responses are used to illustrate some major themes relative to the ATOSU criteria and whether this qualitative data supports the model.

Correlations of the five attributes suggested a simplex pattern that is indicative of a hierarchy, but these correlations may have been influenced by the dependence among ratings resulting from pooling these ratings across judges when calculating the correlation matrix. Measurement results indicated that the ATOSU attributes fit the Many-Facets Rasch model, and indicated a hierarchical relationship among attributes for all but the [U] attribute. Student, task, judge, and attribute facets all demonstrated reliabilities coefficients over 0.90; however, inter-rater reliabilities computed using Guilford's method varied from moderately high to low for each attribute.

Experimental results yielded a multivariate interaction effect between levels of gender with levels of treatment on the conceptual, procedural, and problem-solving dependent variables. Graphs of the interactions were each interpreted for the conceptual, procedural, and problem-solving scales of the NAEP. A large interaction effect on the univariate test on the conceptual scale seemed to explain most of the multivariate interaction. Simple effects measured by these tests revealed large advantages in conceptual performance favoring males in the Test-only group; however, ATOSU females had significantly higher NAEP scores when compared to the Test-only and Tutored-only females.

Qualitative data was finally used to illustrate some major themes and demonstrate how the ATOSU attributes were judged. Such illustrations are left for the reader to accept or reject.
Acknowledgments

I would like to express a special thanks to all my committee members and teachers at The Ohio State University for their encouragement and support during my doctoral program and this dissertation project. First and foremost among the people to be acknowledge is Dr. Ayres D’Costa, my advisor, who continued to push me to improve these ideas and work to higher professional standards. Without his guidance, this project might still be stuck at the “think about it” stage. He generously lent a much-needed hand when I felt overwhelmed or unsure. I am proud to be called “his student,” and I would hope that I never let him down.

I am also grateful to Dr. Thomas Knapp who spent inordinate amounts of time reading my drafts and commenting on my thinking. Thank you for your hard work and continued assistance in the face of my pleas for help. Dr. Knapp is a scholar any student would wish to emulate, since he possesses both insight and wisdom. Dr. Knapp sets the standard for excellence in education.

Many thanks too to Dr. William Loadman who never seemed to be satisfied with an average performance. He possesses superior expertise in measurement and always manages to ask the tough questions. His Socratic philosophy of teaching and guiding students has influenced my thinking about making classroom assessment more responsive to the needs of students when teaching problem-solving. I believe any reader will readily recognize his influences in this work.
Dr. Sigrid Wagner deserves a special acknowledgment for her expert handling of the many details related to mathematics education and qualitative inquiry found in this dissertation. She is an excellent listener and problem-solver who cares about her students. She took particular care to press me when my performance was waning, and to praise me when my performance was waxing. Her affection for teaching and her commitment to research in mathematics education is contagious.

Always in my heart will be the memory of Dr. Edwin Novack, my first mentor and advisor at Ohio State, who first helped me to get accepted into the doctoral program. I owe him much gratitude and thanks that I could never repay. After his recent death, he is still missed and always in my prayers.

Finally, I would like to express my thanks to my family and friends who believed in me. Most particularly, my parents, Thomas F. Tinkler and Elizabeth Taaffe, who always encouraged me to do my best; my dear brother Dennis who supported me unconditionally; my close friends, Tim Graham and Eileen Kennedy, who always listened and encouraged me; and my good friends, Mark Eifert, Paul Tucker, and Tim Kirchbaum, who always enjoyed a good discussion about education.
Vita
Thomas V. Tinkler

Fields of Study

Major Field: Education Policy and Leadership
Area of Specialty: Quantitative Research, Evaluation, and Measurement
Cognate Area: School Finance

Education:

School: The Ohio State University, Columbus, Ohio
Attended: Spring 1991/ present, Ph.D. Candidate

School: The Ohio State University, Columbus, Ohio
Attended: Fall 1981/Winter 1983
Graduated: M.A., Educational Administration, 1983

School: Youngstown State University, Youngstown, Ohio
Attended: Fall 1979/Summer 1981

School: Hiram College, Hiram, Ohio
Attended: Fall 1973/ Spring 1976
Graduated: A.B., Education, 1976

School: University of Tampa
Attended: Fall 1972/ Spring 1973

Research Positions:

Position: Research Consultant
Department: Family Medicine, Ohio State Hospital
Duration: Summer 1993/ Fall 1995

Position: Research Associate
Department: Educational Research Services, College of Education, The Ohio State University
Duration: Winter 1993/ Summer 1993
Position: Research Associate
Department: Life Span Processes, College of Nursing,
The Ohio State University
Duration: Fall 1991/ Summer 1992

Teaching Positions

Position: Middle School Math and Science
District: Columbus Public Schools
School: Linmoor Middle School

Position: Instructor
School: The Ohio State University
Department: College of Education, Professional Education
Duration: Fall 1982/ Spring 1983

Position: High School Learning Disabilities
District: Champion Local Schools, Champion, Ohio
School: Champion High School
Duration: Fall 1978/ Spring 1981

Position: Math and Science Teacher
District: Diocese of Youngstown, Youngstown, Ohio
School: St. Luke School, Boardman, Ohio
Duration: Fall 1976/ Spring 1978
TABLE OF CONTENTS

Dedication .......................................................................................................................... i

Abstract ........................................................................................................................... ii

Acknowledgments .............................................................................................................. iv

Vita ................................................................................................................................... vi

List of Tables ...................................................................................................................... xi

List of Figures ................................................................................................................... xiii

Chapter 1: A New Guided-Assessment Model ................................................................. 1
  1.1 Introduction ................................................................................................................ 1
  1.2 Statement of the Problem .......................................................................................... 3
  1.3 Background of the Problem ..................................................................................... 3
  1.4 Rationale for the ATOSU Guided-Assessment Model ............................................... 6
  1.5 Mathematics as Problem-Solving ............................................................................ 10
  1.6 Purpose of the Study ................................................................................................. 20
  1.7 Justification of the Study ......................................................................................... 21
  1.8 Limitations of the Study ......................................................................................... 22
  1.9 Definitions of Relevant Terms ............................................................................... 24
  1.10 Summary ................................................................................................................. 26

Chapter 2: Review of the Literature ............................................................................... 27
  2.1 Total Scores and the Mastery Approach .................................................................. 28
  2.2 Item-based Scoring and the Guided-Assessment Approach ....................................... 31
  2.3 National Assessment of Educational Progress (NAEP) ........................................... 36
  2.4 The ATOSU Scoring Model as Guided-Assessment ................................................. 39
  2.5 How Guided Assessment Impacts Standardized Achievement ................................ 43
      2.5.1 Scoring Approach ............................................................................................. 43
      2.5.2 Graded Task ..................................................................................................... 48
  2.6 Promoting Gender Equity in Assessment ................................................................. 52
  2.7 Summary ................................................................................................................... 56

Chapter 3: Methodology .................................................................................................. 57
  3.1 Overall Purpose of the Study ................................................................................... 57
  3.2 Design ....................................................................................................................... 59
      3.2.1 Measurement Design ....................................................................................... 59
      3.2.2 Experimental Design ....................................................................................... 60
      3.2.3 Observation Design ......................................................................................... 63
3.3 Sample .................................................................................. 63
3.4 Experimental Treatments ...................................................... 65
3.5 Instrumentation and Data Sources ......................................... 68
  3.5.1 ATOSU Assessment Protocol .......................................... 68
  3.5.2 Achievement Test .......................................................... 70
  3.5.3 Audiotapes of Lessons ..................................................... 72
3.6 Experimental Procedures ...................................................... 73
3.7 Proposed Analysis .................................................................. 74
  3.7.1 Measurement Analysis ................................................... 74
  3.7.2 Experimental Analysis .................................................... 78
  3.7.3 Qualitative Analysis ....................................................... 81
    3.7.3.1 Description ............................................................. 82
    3.7.3.2 Interpretation ......................................................... 83
3.8 Summary ................................................................................. 86

Chapter 4: Analysis of the Results .............................................. 87
4.1 The Research Questions ......................................................... 87
  4.1.1 Validity and Reliability of the Model ................................. 87
  4.1.2 Treatment and Gender Effects ....................................... 87
  4.1.3 Qualitative Themes ....................................................... 88
4.2 Measurement Phase ............................................................... 88
  4.2.1 Measurability of the Model ............................................. 88
  4.2.2 Hierarchy of the ATOSU Model .................................... 92
  4.2.3 Inter-rater Reliability .................................................... 96
  4.2.4 Reliability of the Dependent Measures ............................ 102
4.3 Experimental Phase ............................................................... 104
  4.3.1 Descriptive Statistics .................................................... 104
  4.3.2 Testing the Assumptions of the MANOVA ...................... 107
  4.3.3 Testing the Three hypotheses ....................................... 110
4.4 Analysis of the Qualitative Data ............................................ 115
  4.4.1 Description to Interpretive Prescription ........................... 116
  4.4.2 Interpretative Levels of Attention ................................ 137
  4.4.3 Interpretative Levels of Tactics .................................... 139
  4.4.4 Interpretative Levels of Operations ............................... 141
  4.4.5 Interpretative Levels of Solution ................................. 144
  4.4.6 Interpretative Levels of Unique-Conditions .................... 146
4.5 Summary ................................................................................ 147

Chapter 5: Conclusions, Implications and Suggestions ................. 149
5.1 Conclusions .......................................................................... 149
5.2 Implications ................................................................. 155
5.3 Suggestions for Future Research ........................................ 159
References. ................................................................. 166

Appendices. ........................................................................ 175
A: ATOSU Scoring Protocol .............................................. 175
B: ATOSU Tasks and Goals ............................................... 179
C: NAEP Achievement Test .............................................. 180
D: Item Analysis of the NAEP Achievement Test ............... 190
LIST OF TABLES

Chapter 3 .............................................................................................................. 57

Table 3.1 Scoring Matrix ................................................................................. 61

Table 3.2 Judge scoring matrix for the Facets Analysis ................................. 61

Table 3.3 Factorial Experimental Design .......................................................... 62

Chapter 4. ............................................................................................................. 87

Table 4.1 Task Facet ............................................................................................ 91

Table 4.2 Attribute Correlation Matrix ............................................................. 93

Table 4.3 Attribute Facet ................................................................................... 95

Table 4.4 Students represented the object of measurement in the Facets analysis ........................................................................... 97

Table 4.5 Judge Facet ........................................................................................ 97

Table 4.6 Guilford's Inter-rater Reliability for 2 and 4 judges ......................... 101

Table 4.7 Descriptive statistics used to describe the distributions of individual cells ............................................................................... 106

Table 4.8 Box Test of Equality of Covariance Matrix ..................................... 110

Table 4.9 Multivariate tests compare treatments, gender, and interaction vectors on three dependent variables simultaneously ........................................... 114
<table>
<thead>
<tr>
<th>Table 4.10</th>
<th>Univariate tests of significance compare treatment, gender, and interaction means one at a time</th>
<th>114</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4.11</td>
<td>Snippets of discourse provided as illustrations of levels of attention</td>
<td>133</td>
</tr>
<tr>
<td>Table 4.12</td>
<td>Snippets of activity representing levels of tactics</td>
<td>134</td>
</tr>
<tr>
<td>Table 4.13</td>
<td>Snippets of discourse used to illustrate levels of operations</td>
<td>135</td>
</tr>
<tr>
<td>Table 4.14</td>
<td>Snippets of discourse used to illustrate levels of mastery viewed in a solution</td>
<td>136</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>FIGURES</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chapter 1</strong></td>
<td></td>
</tr>
<tr>
<td>1.1 The Rule of Addition</td>
<td>12</td>
</tr>
<tr>
<td>1.2 The Rule of Multiplication</td>
<td>13</td>
</tr>
<tr>
<td><strong>Chapter 4</strong></td>
<td></td>
</tr>
<tr>
<td>4.1 Facets graphical report</td>
<td>99</td>
</tr>
<tr>
<td>4.2 Histograms of the treatment by gender cell distributions</td>
<td>108</td>
</tr>
<tr>
<td>4.3 Graphs of the treatment and gender interactions for each dependent variables</td>
<td>113</td>
</tr>
<tr>
<td>4.4 Female response to task 2.</td>
<td>122</td>
</tr>
<tr>
<td>4.5 Male response to task 2.</td>
<td>123</td>
</tr>
</tbody>
</table>
Chapter 1
A New Guided-Assessment Model

1.1 Introduction

Teachers need to understand how their students solve mathematics problems, especially as the thought processes required become increasingly more complex. One method teachers commonly use to gain insight about their students is routinely employed classroom assessment. In mathematics, these assessments typically rely on teacher-made or published tests that ask students to solve a number of questions similar to those found in the exercises listed in the textbook. This popular approach fosters two problems: first, teachers receive little information about how to improve or modify their instructional and curricular practices (Stiggins & Conklin, 1993); and second, students are given little direction on how they might improve their understanding when they solve mathematics problems (Wolf, Bixby, Glenn & Gardner, 1991).

Wolf, et. al. (1991) have noted that the "changing views of classroom assessment have to be situated in the larger framework of views on learning and education" (p. 33). Useful classroom measurement methods should generate evidence that encourages students to reflect on and clarify their own thinking about the quality of their work (Wolf et al., 1991). Used jointly, both quantitative and qualitative methods could provide a more complete picture of the learning process (Wolf et al., 1991). By strengthening the focus of attention and evaluating over a short series of well-planned tasks, teachers can provide students a guided assessment with appropriate intervention and the extended support necessary for better mathematics understanding.
Guided assessment is the educational act of improving student performance across a variety of contexts, using evaluative strategies designed to enhance understanding while problem-solving (Shrier & Hammadou, 1994). In mathematics teaching, the primary foci are verbal, visual, and mathematical communications that result in problem-solving and promote invention (Hiebert & Carpenter, 1992). Teachers interested in guiding students through complex problems need scoring criteria designed to focus students on relevant aspects of learning (Quellmalz, 1991).

Since one goal of the National Council of Teachers of Mathematics (NCTM) is to promote gender fairness in classroom assessment (NCTM Assessment Standards, 1995), the proposed guided-assessment strategies should help males and females equally well. While gender differences in mathematics performance have typically been observed, these differences are not always the result of the instructional methods utilized or the assessment practices employed by the teacher. Gender differences may also be attributed to cultural, social, and other environmental factors. Accordingly, it is necessary to analyze the outcomes of any assessment procedure for possible gender differences and to promote beneficial changes in learning and instruction (Harris & Carlton, 1993).

In this chapter, a comprehensive scoring and guided-feedback technique is proposed utilizing five attributes: [A]ttention, [T]actics, [O]perations, [S]olution, and [U]nique-conditions (ATOSU). Given a mathematical problem-solving task, a teacher rates each student using the scoring rubrics found in Appendix A representing the five attributes. ATOSU data provide specific information on the five attributes defined as follows:

1. **[A]ttention**: The student descriptively writes or verbally communicates his/her understanding of the problem with varying degrees of observable effort. Levels of attention may range from nonexistent to deliberate in nature.
2. [T]actics: The student is encouraged to visualize the problem and express these insights as pictures, diagrams, tables, flow-charts, or other appropriate means. Students are rewarded for noting patterns, extending these patterns, or recognizing new patterns by accurately combining previously learned tactics or inventing new tactics.

3. [O]perations: The student is rewarded for mathematical moves that are accurately identified and procedurally carried out. Students are asked to justify their procedures as related to their tactical representations. More points are awarded for inventive moves that require insight.

4. [S]olution: The student is rewarded for mathematical accuracy when achieved both with and without assistance. Additional points are earned when the student works backwards through the problem to check an already correct answer.

5. [U]nique-conditions: The nature of activity is assumed more complex when the task is presented in unfamiliar contexts or when it requires the student to extend and invent procedures in order to solve the problem. Scoring criteria separate the degrees of judged inventiveness given the complexity of the task.

1.2 Statement of the Problem

Current assessment methods provide few opportunities for promoting student thoughtfulness, imagination, and pursuit of mathematical understanding (Wolf et al., 1991). The purpose of this study is to evaluate the utility of a newly proposed [A]ttention, [T]actics, [O]perations, [S]olutions, [U]nique-conditions (ATOSU) framework as an assessment and teaching tool designed to promote guided problem-solving in schools.

1.3 Background of the Problem

Mathematical assessment is defined as "the process of gathering evidence about a student's knowledge of, ability to use, and disposition toward mathematics, and of making inferences from that evidence for a variety of purposes" (National Council of Teachers of Mathematics [NCTM], 1995, p. 3.). A small but growing body of knowledge is beginning
to provide a picture of the nature, quality, and impact of classroom assessment (Fleming & Collins, 1983; Stiggins 1991; Stiggins, Conklin, & Bridgeford, 1986). Classroom assessment, utilized more frequently in schools than its standardized testing counterpart, has the potential to provide more diagnostic information, and can emphasize the evaluation of higher-order skills (Stiggins, 1991). However, research indicates that teachers are inadequately trained to assess students (Schafer & Lissitz 1987; Stiggins, Griswold, & Wikelund, 1989), lack the time to do a thorough job (Stiggins et al., 1986), and are uncomfortable when constructing classroom assessment instruments (Carter, 1984).

Against this environment, one notes the paucity of quantitative and qualitative procedures for conveniently scoring, indexing, monitoring, and evaluating learning problems given the emerging new interest in assessment. The popular classical or closed-ended scoring procedures require dichotomous or partial-credit scoring for multiple-choice, true-false, or matching formats. Typically, teachers calculate and rank total scores by counting the number of points earned by each student. These classical methods provide teachers with little information beyond patterns of success or failure on responses over the domain of the test. Competency or mastery-based methods are criticized as “reductionistic” because they segment the curriculum into specific objectives and skills that are seldom related (Darling-Hammond & Falk, 1997). In the face of educational reforms to develop student's thinking and to open the curriculum to challenges beyond the mere recall of facts, classical scoring procedures have come under attack (e.g., NCTM, 1995; Wolf et al., 1991).

In reply to such criticisms, some measurement experts in the 1980's and 1990's (see Moss, Beck, Ebbs, Matson, Muchmore, Steele, & Taylor, 1992) have promoted the development and investigation of alternative forms of scoring assessment. Rather than
employing item formats that have students choose from among fixed alternatives, another way to systematize scoring is to develop rules for judgment that all scorers can follow. These free-response formats often adopt partial-credit or rating-scale schemes when scoring a predefined set of criteria. Increasingly, free-response assessment techniques are being developed to provide additional information about the student’s problem-solving processes.

The ATOSU protocol represents a free-response scoring procedure employing a set of within-task ratings of attributes modeling the problem-solving process. Teachers rate the student’s responses using the ATOSU attributes viewed in Appendix A. Qualitative and quantitative observations of voluntary [A]ttention, visualized [T]actics, connective [O]perations, accurate [S]olutions, and [U]nique-conditions are believed to focus learning. The ATOSU approach assumes that a set of five attributes may be applied in ways that better guide or direct the learner. With this approach, the evaluator can provide a quantitative and qualitative portrait of the learner’s progress. The teacher works as a facilitator who draws out student thinking and monitors progress. At the end of the assessment, students are asked to state their observations, clarifying whether they understood the feedback.

Systematic reform rests on much careful analysis of what information is necessary for improving learning and instructional intervention. Assessments that scores students’ practice as they perform challenging tasks improves any state or district’s chances at achieving educational reform.
1.4 Rationale for the ATOSU Guided-Assessment Model

Students construct meaning or understanding by experiencing learning both as individuals (Anderson, Reder, & Simon, 1996; Davis, 1992) and within groups (Greeno, 1997; Lave & Wenger, 1991). Evaluating such forms of active inquiry does not mean that the teacher solves math problems for the students as a means of passing on his/her knowledge. Rather, guided assessment encourages teachers to evaluate learner progress as the students engage in problem-solving, both as individuals and within groups. To foster learning within rich environments of activity, new approaches for scoring mathematical inquiry need to be devised that observe and support the basic dimensions of conceptual and procedural knowledge (Glaser, 1979; Hiebert & Carpenter, 1992; Hiebert et al., 1996). By integrating their conceptual knowledge with their procedural knowledge during problem-solving activity, students can extend or construct new conceptual and procedural representations.

No concept is formed from the induction of a single example to a class (Murphy, 1988). It takes multiple examples across multiple contexts before a conception actually forms and presents itself in language, strategic behaviors, or mathematical operations. Activity produces enriching experiences where the learner “makes separate discriminations related to the concept (Larcombe, 1985),” characterizing each experience as contained by or complementary to its stipulated meaning (Larcombe, 1985; Murphy, 1988; Piaget, 1970). The development of higher-order concepts depends upon generalizations made from lower-order concepts – from the concrete or standard example to the abstract case. However, these
higher-order generalizations seem confined to specific contexts and do not always transfer to tasks in other contexts (Perkins & Solomon, 1989). Hence, a hierarchical order of concepts is postulated where the highest order concepts are dynamically constructed from the primary concepts during activity, although these same levels of higher-order conceptualization are subject to conditions or constraints viewed in the task.

Concepts permit the learner to abstract some common experience from a world of particulars (Leon'tv, 1981). In other words, conceptual knowledge expands with the student’s ability to perceive and generalize properties of an object or an arrangement of patterns underlying an event or series of events. These abstractions are always made within the context of activity, whether performed in social contexts or privately. During activity, students reorient their thinking as the conditions of a task change. This relieves frustration and allows new activity when progress is blocked. Such activity may lead to the abandonment of some definition for another, more appropriate, meaning (Vygotsky, 1981). Although stable, conceptual representations are metaphorical in character (Lakoff & Johnson, 1980), so knowledge often evolves by functionally mapping or comparing qualities of one set of objects to another set of objects. However, although conceptions seem bounded by definition at points in time, logical procedures are assumed to permeate these boundaries during times of activity. Since concepts become more "general" or restricted with "use and application" (Dewey, 1971), the relationship between qualities of objects or events dynamically changes or fluctuates over time. Higher scores on ATOSU scales are believed to represent ordered and more generalized interpretations of conceptual thinking.
Dewey (1971) maintained that thinking is a process: it flows, it is endless, and it is in a continual state of change and adjustment. From an early age, activity draws inferences, connects, maps, and transforms native ability into expert knowledge. As Dewey pointed out, knowledge of procedures is not necessarily understood and applied as a set of formal rules viewed in logical syllogisms or mathematical properties; rather, the uses of formal rules are means of testing, validating, or proving what occurs naturally in reasoning. The natural recognition of knowledgeable procedures is viewed in the way students connect, interpret, order, and derive meaning from the evidence in actual practice. Outcomes are the products of the reasonable and reflective flows of thought. To facilitate the teacher’s understanding of his/her students’ thinking, assessment should promote, evaluate, and track the natural progress of activity as it moves toward a solution.

A process consists of actions “subordinated to the idea of achieving some result” (Leont’ev, 1981 p. 60), subject to some task or problem condition. Procedures are structured sequences or algorithms that are identified, changed, and performed to solve a problem. Procedures carry out rules through activity among conceptual representations. Activity attempts to reconfigure procedures as the condition of the task changes. From a developmental perspective, the establishment of a combinatorial set of rules plays a central role when linking corresponding ideas (e.g., Flavell, 1963; Piaget, 1957). A subject’s ability to "link a set of base associations" or correspondences with each other in all possible ways is first regulated by an adult or other participants. Adults and peers examine procedural representations and attempt to publicly make sense of them. Through his/her social interactions with others, the learner eventually understands how to mediate his/her own
activity by learning how to identify what is important and significant in memory for determining future action (Blumer, 1969; Vygotsky, 1981). The ATOSU scales may be examined in sequence to monitor procedural actions from one problem/task to the next.

Hiebert and Carpenter (1992) concluded that conceptual and procedural representations of a problem are connected. For this reason, it is often difficult to delineate where in thinking one begins and another ends. To better understand how concepts and procedures are related, Hiebert and Carpenter (1992) proposed definitions for both constructs that identify such activities with an observable set of behaviors. To clarify and distinguish conceptual and procedural activity, they equated conceptual knowledge as "connected networks" internally integrated by the learner that increase in sophistication as the person gains understanding (Chi & Koeske, 1983). In contrast, for these authors, procedural knowledge consists of actions taken in sequence using the practical as well as formal rule-systems formed in mathematics or logic.

Mathematics educators agree that instructional and assessment programs that meet the needs of each child never sacrifice one thinking skill at the expense of another (Hiebert & Carpenter, 1992). Experts believe that when conceptual and procedural skills are integrated, they encourage invention and transfer (Hiebert et al., 1996); when students are permitted to reflect upon their mathematical conceptions and encouraged to think in unfamiliar ways, they improve on their routine problem-solving procedures (Carpenter, Fennema, Peterson, Chiang, & Loe, 1989; Cobb, Wood, Yackel, & McNiel, 1991); when students are encouraged to solve problems rather than memorize, they better organize information in ways that make them sensitive to what they are doing (Carpenter et al., 1989;
Hiebert & Wearne, 1993); and when students work from a comprehensive knowledge base, they gain the depth of the expert's schematic knowledge (Larkin, McDermott, Simon, & Simon, 1985).

From these research-based propositions comes a viable starting point for an assessment technique for scoring mathematical problem-solving. In the last section of this chapter, a new measurement approach is advanced in an effort to capture more of the complexity of problem-solving in mathematics to enable teachers to provide guided feedback to their students.

1.5 Mathematics as Problem-Solving

A new view of what problem-solving is has evolved in recent years. This new view has led to changes in the way problem-solving is taught in classrooms. The proposed [A]ttention, [T]actics, [O]perations, [S]olution, and [U]nique-conditions model is developed as a general set of five attributes by which one might assess the conceptual and procedural structure of problem-solving activity. Following the substantive theories of both Polya (1957) and Russian psychologists (cf., Leont'v, 1981; Vygotsky, 1981; Zinchenko & Gordon, 1981), these five attributes respectively represent three kinds of communications activities, one over-all problem solution/goal, and one unique set of problem/task conditions; and are operationalized in the instrument viewed in Appendix A. This section will now discuss the theoretical and operational development of this instrument.

Given the nature of most probability and statistics tasks, this research assumes that human beings transmit and interrelate ideas and thoughts through three forms of
communication that comprise mathematical activity: 1) a language or verbal activity 2) a visualization or imaging activity, and 3) a mathematical activity. These three forms of communication may be represented in a student's answer to an open-ended question (see Figures 1.1 & 1.2). Each of these communication activities -- together with their resulting actions, levels of solution, and constraining task conditions -- are represented by a category specifying the variables of the ATOSU model. Each of these variables will now be explained, justified, and operationalized in more detail.

The first form of communication, language or verbal activity, is broader than the mere semantics of words. In spoken language, a word in itself is a "generalization" of many characteristics of an object (Vygotsky, 1986). Likewise, a sentence combines words in ways where the literal meanings of individual words are less emphasized, while the "sense" of what is being expressed within the phrase or sentence becomes more critical. More generally, as Vygotsky concludes, sentences build into thoughts or arguments that constitute even more generalized reflections of the verbal thought that constructs and reconstructs reality (Vygotsky, 1986). Likewise the mathematical communication necessary for solving written word problems is understood by sensing their more general meanings rather than focusing on key words or phrases. Mathematics learning needs to be grounded in a wide variety of concrete experiences; and students need ample opportunities to generate ideas and processes (Katterns & Carr, 1986; Silver & Smith, 1990).

Vygotsky (1981) asserts that there are two forms of attention development: The first relates to neurological mechanisms and "the organic processes of growth, maturation, and development of the neurological apparatuses and functions of the child" (Vygotsky, 1981,
Verbal Communication

What is the probability of drawing a face card or black card in a single draw?

Visual Communication

Spades 2 3 4 5 6 7 8 9 10 J Q K A
Clubs 2 3 4 5 6 7 8 9 10 J Q K A
Diamonds 2 3 4 5 6 7 8 9 10 J Q K A
Hearts 2 3 4 5 6 7 8 9 10 J Q K A

Mathematical Communication

\[
\frac{3 \times 4}{4 \times 13} + \frac{2 \times 13}{4 \times 13} - \frac{2 \times 3}{4 \times 13} = \frac{32}{52} = \frac{8}{13}
\]

or more generally expressed as the probability of the union of two events,

\[
P(A \cup B) = P(A) + P(B) - P(A \cap B)
\]

Figure 1.1: The Rule of Addition may be taught by encouraging connections between verbal, visual, and mathematical forms of communication.
Verbal Communication

What is the probability of drawing both a face card and a black card in a single draw?

Visual Communication

Spades  2 3 4 5 6 7 8 9 10 J Q K A
Clubs   2 3 4 5 6 7 8 9 10 J Q K A
Diamonds 2 3 4 5 6 7 8 9 10 J Q K A
Hearts  2 3 4 5 6 7 8 9 10 J Q K A

Mathematics Communication

\[
\frac{4 \times 3}{4 \times 13} \times \frac{2 \times 13}{4 \times 13} = \frac{3}{26}
\]

or more generally expressed as two independent events if,

\[P(A \cap B) = P(A) \times P(B)\]

Figure 1.2: The Rule of Multiplication may be taught by emphasizing verbal, visual, and mathematical forms of communication.
p. 193); the second form, called voluntary attention, is described as a qualitatively distinct process of attention emerging from the social interactions of children with adults. The ATOSU model is primarily concerned voluntary attention over neurological mechanisms of attention.

Voluntary attention utilizes "external auxiliary means" in the effort to mediate attention (Wertsch, 1985). In probability and statistics exercises, for example, a teacher might use playing cards as objects or referents to be manipulated and verbally related during instruction. In other words, the student participates in a problem situation where the teacher questions the student about the relational properties and moves viewed while solving a problem. Using verbal indicators, the teacher encourages the student to see alternative patterns, and to develop useful verbal interpretations for characterizing these patterns. Such feedback helps the student learn that generalizations apply to broad mathematical concepts that may be abstracted from patterns. This feedback also communicates the levels of thinking and completeness of explanations expected from the problem-solver.

Voluntary attention implies that some conscious level of effort needs to be exerted to carry out and direct attention. Mastery of attention processes provides an increasing opportunity for "subordinating" attention under the control of the student’s own authority. Vygotsky (1986) emphasized that language plays a major role in the continued adult/child interactions that lead to the development of self-regulated thought. As learners become better at mediating their conceptual and procedural representations during the social communication of instruction, they slowly begin to internally mediate thought using inner language and self-reflection. Therefore, it would appear that voluntary [A]ttention within
the context of a language system is potentially a viable link to cognitive development and understanding.

The second form of communication is visual or imaging activity. Imaging activity forms, manipulates, and transforms visual symbols and patterns on the intrapsychological plane. Through social activity, the learner understands how to visually represent, interpret, and make sense of these processes (Eisner, 1982). Children visualize, interpret, and make sense of their experiences when playing with their erector sets, when watching the pictorial representations of Walt Disney, or when paging through a storybook with their parents. These external objects are represented by signs and forms that are contextually understood and made sense of (Blumer, 1969; Hiebert & Carpenter, 1991). As with language, such images are operated on, interpreted, synthesized, or transformed within activity. Heuristics or [T]actics are considered projective or expressive techniques in that they suggest aspects of the person's visual cognition while he or she is representing the activity as drawings, tables, flow-charts, or other forms. By learning to generate [T]actics with adult guidance, the learner becomes adept at representing visualized forms using their tactics. With time, new patterns may emerge that are considered very useful when fluently extending or reworking their imagined forms.

The impact of conceptual representations on imagined entities is termed "structured imagination" (Finke, Ward, & Smith 1996). A great deal of experimental work on structured imagination and creative invention can be found in the cognitive science literature. The purpose of many of these studies was to demonstrate that creative invention is predictable. Although the exact form of the creative entities may not be predictable (Finke & Slayton, 1988), one can identify the features and conditions under which the
creation is structured. For example, college students were asked to draw imaginary creatures from a planet in another galaxy similar to our own. The students’ creations possessed properties similar to those found in earth creatures, like bilateral symmetry, sensory organs, and functional appendages. Therefore, rather than being idiosyncratic, the students’ creations appear to be highly structured by characteristic attributes of known categories.

Once more, teachers might indicate where a student’s tactical representations may be improved. In the playing cards example, students learn how to attribute tactical representations of problems communicated to them in verbal form. Students are encouraged to construct their own tactics by drawing pictures, making flow-charts, building tables, or in ways not taught. Here, the best feedback is descriptive, specific, relevant, timely, and encouraging. The teacher is encouraged to listen to the learner and to make suggestions rather than provide dictums. The teacher examines how parts of the tactical representation communicates a general direction for utilizing the learner’s math conceptions and procedures.

The third form of communication, mathematical activity, has concepts and procedures tantamount to those viewed in imaging and language communication. Numbers or symbols are used to represent mathematical respectively, while mathematical [O]perations transform a given domain of values for these traits into a range of [S]olutions. Mathematical systems are learned in communities of mathematical practice, where mathematical power is the understanding to apply mathematics to unfamiliar situations, as well as having the confidence to do so. As language and imagery builds to these more holistic structures, mathematical thought constructs, reconstructs, and interprets reality in similar ways.
As one language may be translated to another, mathematical communication similarly translates the images and verbal thinking of students into mathematical reasoning. Intercommunication between verbal and image activities may be noted by describing some problem verbally, asking the students to imagine the problem through some mediated tactic or heuristic, and drawing the visualized picture on paper. In Figures 1.1 and 1.2, the intercommunication between visual and mathematics activities may be noted by presenting a drawing to the student and asking the student to write an equation or [O]peration that addresses the problem and is represented within the drawing. For many problems, there might not be a single procedure or a clear answer. However, students would need to demonstrate that their mathematical interpretation of the diagrams, pictures, tables, or manipulatives seems reasonable and appropriate. Intercommunication between verbal and mathematics activities may likewise be accomplished by asking students to write equations representing the question. However, asking some students to generate mathematical inferences without some heuristic or tactic might prove difficult when students lack the experience or fluency with mathematics. Moreover, students are encouraged to become more efficient with their operations by utilizing mathematical principles that need to related and applied to problems after they have been taught (Kennedy & Poplin, 1991).

The overall goal or [S]olution is viewed as the result of interrelated conceptual and procedural activity. Students are asked to explain why their mathematical solutions follow from their procedures. They are further encouraged to work the problem backwards to demonstrate that they are correct. However, students are not permitted to flounder in their search for a solution. They are encouraged to mediate their activity through the questioning patterns of the teacher, eventually sharing their thoughts with the teacher or with others.
within the group. The aim of this communication is to motivate students, and to utilize their solutions or successes as opportunities for further or more advanced inquiry.

Goal-directed processes or actions like mathematics are subject to the conditions under which the operations are performed. For example, in the case of mathematics, aspects of symbols like numbers are organized into structures by means of directed and undirected thought occurring within these contexts. The conditions of a problem-solving task will influence difficulty by demanding more abstraction as contexts become less familiar. In mathematics, symbols are interrelated and given meaning within equations written as a result of a problem’s conditions. Interfunctional number meanings like these are interpreted within social discourse when teaching via problem-solving, rather than absolute or literal interpretations made when teaching about problem-solving (Schroeder & Lester, 1989). Conceptual understandings are communicated socially and are internally constructed as different aspects of objects are related in systematic-learning activity. The same sense-making within mathematical contexts can be recursively moved to real-world contexts and back again a number of times, so that both conceptual and procedural knowledge are influenced by the (U)nique-conditions presented by some real-world task.

Within the instrument, hierarchical generalizations of problem-solving are subsumed under the [A]ttention, [T]actic, [O]perations, and [S]olution variables (see Appendix A). Given an open-ended problem, the teacher quantitatively rates the level of problem-solving at which the learner is functioning, qualitatively comments on unexpected or informative activity, and qualitatively notes error patterns. When examined individually, each variable provides the teacher with information about the student’s level of problem-solving achievement represented by the five ATOSU attributes. Analyzed individually, each
of these four ratings is viewed as an accomplished level of verbal, visual, and mathematical communication, given the task at hand. By monitoring these four ratings of problem-solving activity, it is believed that teachers will recognize, increase, and broaden student learning. Procedural mastery, as measured by the ATOSU scales, reflects generalized levels of verbal, visual, and mathematical communication posited by this guided assessment model. Mastery is achieved when a student selects and applies these procedures across many task conditions.

By changing problem conditions, the teacher can affect the levels of difficulty in the tasks, and draw on the student’s ability to transfer previously learned skills. Tasks must first be comparatively simple to convey understanding, increase confidence, and encourage discovery. Tasks must become increasingly more complex as students grow more astute and comfortable with their learned concepts and procedures. Tasks become more complex as the conditions upon which they are encountered change. Problem conditions may be changed by formulating tasks that present familiar mathematics concepts and procedures in both familiar and unfamiliar contexts, and by tasks that demand that students modify, extend, or generate new, but related, mathematics concepts and procedures in unfamiliar contexts.

The combination of procedural and conceptual representations helps teachers to help students develop and consolidate his/her problem-solving skills; and, more importantly, to generalize these skills beyond the subject matter and the classroom. This combination can also be utilized to track longitudinal growth for a student, or to compare one student with another, or one task performance to another.
1.6 Purpose of the Study

This study examines the utility of the ATOSU assessment model in enhancing student problem-solving achievement in middle-school level probability and statistics problems. To discover whether this systematic information produces increases in conceptual, procedural and overall problem-solving achievement, this research employs ATOSU as one of three treatment conditions in an experimental study. A collection of items provided by the National Assessment for Educational Progress (NAEP) will serve as the dependent measures.

Gender differences are examined by comparing male and female students in this study using a 3 x 2 factorial MANOVA design. Three scales (conceptual, procedural, and problem-solving) were identified by NAEP for the achievement test, thus making this design a multivariate factorial design.

Since the general purpose of the ATOSU model is to measure how students progress through the problem-solving process, it is necessary that preliminary determinations of the ATOSU instrument's validity and reliability are also made. This knowledge will accrue increased insights into the model's attributes and the appropriateness of its components. The utility of the model is examined in this study by posing three sets of research questions:

1.6.1 Are the variables measured in this study valid and reliable?

1. Are the five ATOSU attributes measurable? Are the tasks content valid and reliable?
2. Is the hierarchical nature of the five ATOSU scales validated in this study? Is the hierarchical structure reliable?
3. Can the ATOSU model be implemented reliably by the judges?
4. What is the reliability of the conceptual, procedural, and problem-solving scales used as the dependent variables?
1.6.2 Are there treatment and gender differences? Are there interaction effects between treatment and gender?

1.6.3 What are some major themes described within the qualitative data relative to the ATOSU attributes? Do the qualitative data support the use of ATOSU protocol?

1.7 Justification of the Study

Teachers employ guided activity when they direct students toward a goal. A teacher's activities are effective when they are planned in ways that help students acquire increased insights into how to better utilize their abilities. The interaction of planned activity with the predisposed talents of the learner produces learning activity when the teacher is clear and the student makes a sufficient and relevant effort. While classroom instruction is an important step, it should be viewed as part of a cognitive cycle employed with guided feedback and assessment. The key dimensions of this problem-solving activity may be observed and critiqued to improve instruction and to increase student understanding. Research in assessment and guided feedback can provide an understanding of how these practices contribute to individual student problem-solving performance and to examine potential gender differences.

The proposed assessment strategy is employed as a means of observing cognitive activity as a student solves mathematics problems. It is assumed that there are different stages of cognitive activity where students extend, generate, or attribute various mental concepts and processes. Teachers can observe and rate recurring activity as it unfolds during learning. To be pedagogically useful, assessment that flexibly targets, informs, and directs the learner at these points would seem necessary and appropriate.
The various forms of conceptual and procedural representations utilized within each problem-solving activity must be applied flexibly. Assessment is more effective when it encourages students to reconceptualize or reconstruct their thinking when confusion arises. In mathematics, this seems more likely to transpire when students are asked to search for new patterns and relationships between their conceptualization of a problem and the procedures they use to find a solution. To be useful to learners, new ways of scoring and feedback need to be employed that help them surmount obstacles blocking their improvement. Assessment of students' ability to communicate thoughts, to tactically look for patterns, to make connections between tactics and operations, and to think in novel ways seems important. A scoring framework is needed that interrelates the assessment criteria and builds cohesiveness within the evaluative process.

Finally, teachers need to be concerned with assessing children in ways that promote such benefits to both males and females to attain their potential level of mathematical competence. Such thinking leads to an individualized treatment of students. Assessment properly attributed and employed could detect any differences and direct instruction based on these differences.

1.8 Limitations of the Study

The ATOSU assessment plan utilizes a set of evaluative tools designed to work in conjunction with instruction. Test items covering probability and statistics topics drawn from the National Assessment of Educational Progress (NAEP) are used to test the potential benefits of this assessment guided-feedback model. Any gender or treatment differences in
test performance are examined using both qualitative and quantitative techniques. However, several limitations present themselves as a result of the chosen assessment model, and the methods and design utilized in this study.

The ATOSU model produces a profile of five scores designed to structure feedback around the student's problem-solving performance. This creates a complex feedback system for teachers to use. It also provides a challenge for measurement. Furthermore, since the NAEP items included both open-ended and multiple-choice styles, several other complications resulted. Students typically take more time to solve open-ended items and teachers need more time to judge each item. As a result, reliability suffers since fewer items are administered during an assessment session to produce the scores. A long test is generally more reliable than a short test. Furthermore, every item added improves the representativeness of the sample of performance.

Other potential internal validity problems of the proposed experiment include selection bias due to the loss of some subjects (participation was voluntary), halo effect due to the fact that the researcher provided the ATOSU treatment, and memory effects due to the intervals of time among treatments. There is limited generalizability for this study due to the small sample (n=76) and the homogeneity of the population [Title I program at one local middle school]. Hopefully, this study’s benefits will outweigh its limitations, since a new assessment model is being introduced that can give educators a new tool to improve assessment and teaching.

Time and limited resources prohibited a prolonged test of the ATOSU model. Since the treatment group is guided using the ATOSU assessment framework for only a brief time
(three forty-minute sessions), the possibility exists that the treatment effects will be small
due to these time strictures. Typically, most teachers would spend more time teaching
statistics and probability since many seventh-graders are not experienced with these topics.
Moreover, non-significant differences among treatment groups seem possible since the two
of the treatment groups are tutored, and tutoring is also an effective way of producing
achievement (see Bloom, 1984). Treating students one-on-one requires inordinate amounts
of time, which limited the size of the sample. However, to meet the statistical assumptions
of the analysis, the students had to be tutored separately so that each student’s performance
is independent of his/her classmate’s performance.

1.9 Definitions of Relevant Terms

To facilitate understanding and clearly differentiate the technical terms usage
specific to this study, it is important to define each pertinent concept. Ordinarily, these
terms have more than one meaning, and are often dependent upon the context in which they
are used. The terms and definitions listed below delineate how these concepts are generally
used within this research.

Attention: scored verbal effort reflecting the reception and identification of key
points of voluntary information emerging from the socio-cultural
interaction of the child with the adult (Vygotsky, 1981).

Concepts: hierarchical, class-inclusional, and context-influenced definitions that
are symbolized and represented when solving a given problem
(Murphy, 1988).
Conceptual Vector: measured set of scores listed down the rows of the data matrix for a single ATOSU attribute. The rows represent different tasks requiring graded conceptual problems given to a student.


Generative: the act of creating novel thoughts or inferences (Finke et al., 1996).

Measurement: provides rules for assigning numbers to objects in such a way as to represent quantities of attributes (Nunnally, 1978).

Measurement Design: a plan for defining and operationalizing measures of cognitive and task structures (Wright & Masters, 1982).

Observational Design: a plan for collecting qualitative information about cognitive and task structures (Patton, 1990).

Operations: observable mathematical concepts and procedures used by students and scored by judges. Operations describe the dominant and relevant rules employed to solve a problem.

Procedures: activity or combinations of activities used when mapping sequentially applied rules or operations. Procedures are sequential but fluid, since they are orderly and permeate within and between conceptual systems (Hiebert & Carpenter, 1992).

Procedural Vector: measured set of ATOS scores listed across the columns of the data matrix, for a single student given a specific task.

Scoring model: a protocol describing a procedure for rating student problem-solving responses for use in a specific measurement model.

Specific behaviors: anchored behaviors subsumed and ordered under each of the five general ATOSU attributes.

Solutions: judged accuracy of a student's response to a problem/task.
Tactics: observable, written, and scored representations of a visualized problem-solving plan. Mediated practice using tactics is assumed to aid the student in creatively visualizing the problem-solving strategy.

Unique Conditions: judged and scored degree or condition of invention viewed within a given task.

1.10 Summary

This chapter has introduced a new guided-assessment model, called the ATOSU, for teaching mathematical problem-solving in schools, and has proposed to evaluate the utility as its major research goal. This chapter also presents a statement of the problem, a background of the problem, and provides a theoretical rationale for the proposed ATOSU guided assessment. The ATOSU model is used to classify and sequentially order assessment attributes important to problem-solving in order to promote conceptual and procedural thinking in mathematics. This study will attempt to answer research questions pertinent to the development and utility of the ATOSU model. The study's research purposes were specified and justified, and expected limitations were discussed. Chapter 2 will next review the pertinent research literature that influenced this study's development.
Any new assessment model should reflect current professional thinking in the light of past measurement research. Classroom assessment in mathematics must describe a framework of constructs representing cognitive concepts and processes that are utilized by students during their learning activity. Scoring approaches that relate key components of problem-solving are thus prerequisites for building a good guided assessment system. This research examines the utility of a new assessment model called the [A]ttention, [T]actic, [O]peration, [S]olution, and [U]nique-conditions (ATOSU) model. To accomplish this objective, the first two sections (2.1 & 2.2) of this chapter distinguish between total score and item-score approaches for scoring mathematics problems. The next three sections (2.3, 2.4, & 2.5) of this chapter then examine the rationale and the potential for the proposed ATOSU model within the context of the current expectations of the profession and the nation. And finally, since this proposed assessment model is expected to treat both males and females fairly, the last section (2.6) summarizes the relevant research related to gender differences in mathematics performance associated with such assessment procedures. In brief, this review is presented in an effort to conceptualize the relevant theory and justify the research hypotheses and predictions related to the ATOSU model advanced in Chapter III.
2.1 Total Scores and the Mastery Approach

The "total scoring" approach represents the more traditional procedure for scoring responses. One begins by constructing a limited sample of items representing a selected dimension of performance. Standardized scoring rules are then applied to student responses to test several items/tasks, and the total number of correct responses is derived. In the classroom, the customary practice is to use the same scoring rules and item formats as are commonly used in standardized tests. These scores are usually scaled so that they may be ranked along a single continuum, or "yardstick", and comparisons made between students. Several tests and their scores may be presented for each student, thus rendering a profile of scores in a multidimensional context.

For teachers trying to provide feedback to individual students, scoring methods that merely order total scores on a continuum seem fundamentally flawed. This type of assessment information provided to teachers often does not match its major intended purpose — namely, to improve instruction to better serve learners. Assessment information must identify specific patterns or sequential configurations found in the student's thinking pertaining to their interpretations or misinterpretations of the material learned. Because traditional testing is typically performed after instruction, assessment information is seldom designed or utilized to supplement instruction (Resnick, 1994). There is typically a uniform administration of the test after instruction, summarizing the progress of the student based on response patterns and the total number of correct items. Although seldom performed in primary and secondary school classrooms, the technical analysis of difficulty, discrimination indices, reliability, and error patterns can take place only after
the test is administered. Even if these indices were regularly computed and utilized, such analyses fall short since they seldom capture the structure or process of learning. Sources of difficulty and developmental progress go unregulated over the multiple changes of thinking that are understood to transpire during learning and problem-solving (Glasser, 1986). Classical scoring methods of counting an observable occurrence and creating a total score typically do not explain how a person achieves a given response (Lohman & Ippel, 1993). There is no explanation as to how problem-solving traits vary within and among individuals. There is no monitoring of cognitive processes demonstrating how these conceptual traits are related to and elicit the cognitive processes necessary to solve a mathematics problem/task as the conditions of the problem change.

For practical reasons, teachers accept that total scores are very useful in assessments since they are easily derived, convenient, more fair, and easily interpreted (Stiggins, 1991). These scores allow for straightforward and defensible grade calculations that lend an aura of objectivity to student evaluation. However, teacher-made tests are often developed with items from textbooks or are self-designed multiple-choice and short-answer formats that are relatively more memory-oriented (Flemming & Chambers 1983; Stiggins, Conklin, & Bridgeford, 1986). Teachers may vary their standard regarding the percentage of the correct responses expected from test to test; or, they may use a normative approach and compare individual student scores to peer performance within a class. Such testing practices fail not only to hold students to specific and uniform performance standards, but also to provide growth in problem-solving skills.
Mastery-based approaches were developed to improve on such classroom practices. Mastery-based approaches seek to encourage the specification of objectives or targets to be achieved at some predetermined criterion level of performance held to represent "mastery" of a skill or concept. This form of objectives-driven instruction emphasizes the content to be learned over the cognitive aspects of learning. It focuses on learning outcomes instead of learning processes. This emphasis limits the teacher's use of the assessment information to better understand the cognitive progress of the learner. Advocates of measurement-driven instruction argue that if teachers are to teach to their assessments, then the test targets must be worthy of their efforts (see Guskey, 1994). This goal seems unattainable given mastery-based learning in its current form.

Although total-score approaches, like mastery-based learning, are focused and goal-directed, they seem limited in guiding the educational reform movement in mathematics currently underway. Perhaps more than earlier reform efforts, current reform seems to be shaped more by a concern for how students learn mathematics (Hiebert, 1992). Feedback that focuses on the production of correct outcomes on closed tasks atomizes this process, while feedback that causes learners to reflect on their ideas and communicate them to others establishes a cognitive-based process necessary for reflection and cognitive growth. Test boundaries need to be expanded to include the cognitive forms of thought that students employ when solving a problem. Assessment needs to be expanded and improved to include current theoretical perspectives regarding how students learn mathematics.
2.2 Item-based Scoring and the Guided-Assessment Approach

A second approach to scoring views testing as a means for judging problem-solving performance against a set of criteria given a single, contextualized, and “purposeful” task. These item-based scoring procedures inevitably demand a standard set of protocols for rating and scaling a set of problem-solving constructs. The ultimate purpose is to gain an understanding of the learner based upon a “convergence of information” into teaching/learning ideas guided by contemporary developmental and learning theory. Keeping in mind that the scores guide the learner and improve instruction, item-based scoring models are designed across a variety of contexts using evaluative strategies and tasks. Increasingly popular free-response scoring procedures have been developed to provide additional information about a student’s problem-solving practices. Longitudinal changes in information regarding a student’s conceptual and procedural thinking processes are monitored across time. The predominant open-ended formats might use analytic and holistic scoring criteria that a judge applies as consistently as possible for each student response. In a holistic scoring approach, all aspects of performance are considered simultaneously and a global judgment is made. In an analytic scoring approach, judges make explicit and detailed ratings of distinct and specific elements of student performance. As compared to a single total score, these patterns of sequenced ratings can potentially tell teachers a lot more about the student’s thinking processes.

Item-scoring rubrics, like holistic or analytic scoring, attempt to classify or order cognitive phenomena according to a predetermined set of rules, but they do not always relate the criteria in some substantively meaningful way. Better information about a learner
may be gathered using a combination of an observation design and a measurement design. The goal of an observational design is to discover and understand the structure of relationships among a proposed set of constructs and to use these findings to generate ways of scoring the phenomena. Students are studied in a variety of situations so that a general set of categories is hypothetically generated describing this related activity. To accomplish this end, the teacher as a decision-maker and investigator must plan, develop, and arrange tasks so that a problem’s observational conditions reveal more about the conceptions and procedures that generated them. The goal of a measurement design is to develop a scoring framework consisting of individual measures that marries dimensions of some contrived or known substantive theory to the domain of measurement theory. Therefore, as purported here, an assessment should deal with how both observational and measurement designs can be combined to provide information about the cognitive responses that are produced.

With changes in the purposes of classroom assessment come changes in a teacher’s data-collection priorities. Although an observational design includes understanding a student’s primary cognitive relationships, a good teacher is now concerned with understanding how these mental structures change as task conditions are changed. More specifically, this subtle difference could mean that a teacher is now more concerned with understanding classes of conceptions and procedures and how they interact. Using a guided assessment approach, an observational design may be employed where assessment criteria represent or model how learners holistically connect and relate their conceptual and procedural structures during instructional activity. One might begin by examining
observations with both ordered and unordered categories. Descriptions of student behavior are noted and recorded as they naturally occur during problem-solving, but the observer may go beyond the exercise of mere description. Qualitative observations permit the teacher as an expert to interpret deviations from expected patterns, as well as to assign meaning to the conceptual and procedural differences in activity resulting from these endeavors. In other words, teachers use their observations to build theory within practice from the ground up. For example, the teacher's anecdotal record may categorically analyze and interpret expected and unexpected changes in what students do or do not understand, their current level of reasoning and inventiveness, and their strategies for approaching a problem. Likewise, the observational design may also describe the tasks so that defensible inferences about their theoretical complexity may be made from the observations. Comparisons may be made between these task and behavioral structures to note possible arrangements in the order of complexity vis-a-vis a student's understanding.

A measurement design might attempt to arrange variables in ways that construct a structure, so that patterns and connections may reveal themselves, similar to the "nomologic net" concept first proposed by Cronbach and Meehl (1955). A nomological net has a structure or framework that a scoring rubric attempts to capture by looking for constructs that are substantively related. A scoring rubric consists of a fixed measurement scale that is behaviorally anchored using a set of criteria at each point in the scale. At the item level, open-ended questions allow for a multi-faceted evaluation of student activity rather than a single dimension or "bipolar" right-or-wrong scoring. The measurement design is more specific and targeted, asking what is the best way to assign number values
to the respective categories used in the scoring rubrics. The measurement of both conceptual and procedural structures makes it possible to specify multiple objects of measurement, namely, the overall problem-solving progress of the student, as well as the generalized performance skills with any specified component of the structure. For any of these individual, component, or conceptual objects, a measurement design “is used to specify the rules that will be used to score, classify, or combine objects of measurement” (Lohman & Ippel, 1993, p.47). To simulate sequence across thought processes, the measurement design should specify the categories of observations in ways that yield an ordering among the conditions of observation (Lohman & Ippel, 1993). Guided-assessment models, with conceptual scales and procedural strings of measures, are assumed to be more informative because they tie student achievement or progress to specific criteria that are difficulty-weighted and ordered across scales.

This new form of item-based scoring represents an alternative movement towards assessing student progress using a new generation of performance tasks. These tasks have been frequently described and utilized by developmental theorists like Piaget (1970) and Fischer (1980). These tasks attempt to characterize how a person relates the congruity of one’s ideas to one’s actions by consciously considering one’s experience. As recommended by social psychologists like Vygotsky, such activity may be communicated to other participants, or to an observer, in an effort to validate an idea, clarify one’s reasoning, or rework one’s approach when incongruities exist. When compared with traditional tasks, some experts feel that such extended projects or tasks promote the self-reflective judgments of teachers or students (Resnick, 1994). These tasks might include
more “authentic” situations because they are valuable activities that change the conditions of tasks and involve performances that promote invention (Linn, Baker, & Dunbar, 1991). Consequently, since this broader framework of evaluation attempts to capture how a student’s cognitive structure evolves and is modified, the targets of alternative scoring are focused relatively more on the inventions of the learner and the communicational patterns of the learner with others during problem-solving.

These item-based scoring rubrics reduce the concern of unethical test practices like teaching to the test. The development of such scoring procedures is not undertaken with the standardized test in mind. Resnick (1994) reported that performance-based assessment is unrelated to most forms of standardized testing since normative tests typically do not represent the kinds of knowledge and skills viewed in real-world contexts. Total scores that emphasize classical theories of measurement seek to identify abstract “traits” or dimensions “that are, in essence, unrelated to a given context.” Classical test theory usually “treats these (context-related) conditions as ‘noise’ or ‘error’” (Resnick, 1994, p. 513).“ Thus, this new item-based scoring approach represents a stimulating alternative for discovering systematic patterns and relations in the data that were previously treated as error. Given the teacher’s purposes, these item-based scoring rubrics are potential improvements over the standardized forms of dichotomous (0, 1) scoring, especially since they focus on identifying general and multi-faceted principles for improving problem-solving instruction and learning.

However, alternative approaches to scoring at the item level do have their problems. These approaches are potentially more costly to the district, since they are often
tedious and time-consuming for the practitioner to develop and use. Time for assessment practice is purchased at a high cost in terms of alternative teacher duties like instruction and curricular planning. However, if this assessment effort produces a substantial impact on achievement even for the small number of students, then this large positive effect for a small set of learners can offset the small negative effect due to the time-crunch suffered by the larger group. Administrators and teachers would need to consider how they might reformulate current schools to accommodate such assessment practices. Even with these potential problems, it seems important that researchers continue to investigate how such assessment may be used to serve teachers and learners.

2.3 The National Assessment of Educational Progress (NAEP)

On a grander scale, standardized tests are another form of classical assessment regularly used in the United States to provide information about group and individual achievement. An example of a testing program designed to assess groups rather than individuals is the National Assessment for Educational Progress (NAEP). Based on the total-score concept of measurement, this test is an effort to assess the impact of the nation's educational efforts. In the case of the NAEP, the total score and scale scores measure basic constructs such as conceptual thinking, procedural thinking, and problem-solving across large groups. NAEP makes conscientious attempts to reflect changes in curriculum and educational objectives. NAEP is implemented by Educational Testing Services (ETS) under the guidance of professional experts and government agencies (Mullis, 1990). However, item information about the cognitive structure necessary for
examining progress in conceptual thinking, procedural thinking, and problem-solving is not integrated within the assessment or feedback. Merely reporting a set of unidimensional scales reveals little about the qualitative differences between individual conceptions and procedures that can be captured through the verbal dialogues, visualized strategies, and mathematical operations underlying these scales. Although tests based on total scores are limited for guiding learning, the NAEP does create good tests of conceptual thinking, procedural thinking and problem-solving because it provides for forward-thinking curricular goals.

Although not denying the legitimate role of the NAEP, it must be recognized that standardized testing approaches are often limited to the extent that their form encourages “teaching to the test”, sacrifices higher-order processes to a chosen focus on evaluating basic skills (Shephard, 1990), and ignores the evaluation of inventive thinking inherent in problem-solving (Resnick, 1994). Although the accountability movement has fostered a culture where scores on basic skills have been rising (Darling-Hammond, 1991), all tests induce substantial control over what is taught and learned in schools. Critics view standardized methods for assessing achievement as “products of mass production” represented by a factory-model view of education (Resnick, 1994). While many practitioners bemoan the fact that those who have control over the test have immense impact over instruction and the curriculum, a few others have argued that there may be an advantage in this situation. This latter group of educators has suggested that the utility of classical tests may be pushed a step further, if the results are actually utilized to drive instructional improvement and increase student achievement. This results in what some
educators term “measurement-driven” instruction (Airasin, 1988; McLaughlin, 1991; Popham, 1983; Popham, 1987). Despite these competing views, Silver (1992) has pointed out that tests like the NAEP have a “symbolic function” in that they signal to the students, teachers, and the community what is important to learn and know. No matter what one’s view, standardized tests like the NAEP serve an important role in influencing the educational milieu since they allow all participants a meaningful way to compare and judge progress across large groups.

NAEP testing is designed to determine the progress of education for groups of students in the entire country. NAEP takes great care to construct mathematics items that reflect the most current standards of excellence, as defined by the National Council of Teachers of Mathematics (NCTM) using state-of-the-art psychometrics. NAEP makes conscientious attempts to respond to changes in assessment technology; for example, incorporating refined item-response scaling methods in data-analysis procedures, and adopting innovations in performance testing. Students are selected according to scientific sampling procedures designed to yield nationally representative results for particular subpopulations of students as defined by gender, race, region of the country, and size of the community. NAEP also assesses public and private school students at grades 4, 8, and 12. Test questions are assigned to students on a random basis, so that, within a school, no two individuals are taking the same test.

These standardized tests are normed to allow unit comparisons among individuals, classes, schools, and districts. They allow practitioners to note generalizable patterns of summative improvement attributable to programs, innovations, and teaching. As with
most standardized tests, a set of scales can review the amount of progress one group has made relative to another in a given discipline.

Because of this ability to compare relative performance with similar age-groups across the country, an achievement test consisting of NAEP items appeared a reasonable dependent variable in this study for assessing the overall impact of a research treatment designed to produce improvement in conceptual thinking, procedural thinking, or problem-solving. NAEP presents an appropriate framework of what students should know and can do, since its utility and credibility are based on its ability to change and keep pace with current interests in education. The three NAEP conceptual, procedural, and problem-solving scales reflect constructs similar to those emphasized by the proposed ATOSU model, even though these skills are scored differently. NAEP also makes efforts to ensure that none of its items are biased against any protected population groups. For these reasons, a dependent variable that classifies NAEP items by these problem-solving scales seemed an appropriate choice for this study.

2.4 The ATOSU Scoring Model as Guided Assessment

Glasser (1986) has long argued that what is needed is a theory of measurement that can be used by practitioners to monitor understanding and to facilitate the construction of student knowledge. Assessment should simulate and map the psychological conceptions and procedures of the learner during problem-solving. In mathematics, such an assessment theory would demand feedback that is generic, yet modifiable, given the specific conditions presented by each new task and the abilities of different students. An
assessment theory is generic when it recognizes that humans compare their common and regimented experiences to their current experience as a starting point when encountering a problem. However, experience can also influence the learner in transient ways, producing activity that generates adaptations of new conceptions or procedures. This performance-based approach is operationalized by the [A]ttention, [T]actics, [O]perations, [S]olution, and [U]nique-conditions scoring protocol presented in Appendix A.

The ATOSU criteria are presented as a scoring rubric that may be analyzed to provide conceptual as well as procedural feedback within a general framework. The purpose of this assessment approach is to develop “habits of mind” by unifying the specifics of individual ability, instruction, context, and task with the general standards of educational quality. Such a relationship may be fostered by producing regular feedback that not only examines differences between high and low achievers, but also allows for a common set of standards that may be shared by the learner and the teacher. Standards are not examined at the end of the line but are pervasive throughout the process (see Resnick, 1994).

Wiggins (1993, p. 204) defined competence as a judgment involving “effective adaptation to specific roles or situations.” Since the essence of mathematical competence is problem-solving, Wiggin’s definition of competence suggests that mathematicians focus on habits of good problem-solving. To accomplish this goal, an assessment framework in mathematics needs to challenge, recognize, and reward fluid performance within a variety of specified contexts and experiences (Hiebert, 1992). Any assessment framework worth its value becomes a vehicle by which the criteria may be related, tested, and
reinterpreted. Within the ATOSU framework, this is made possible by using both qualitative and quantitative evidence. Within the qualitative design, qualitative observations need to note individual and contextual constraints that cause confusion or mistakes. Qualitative themes need to capture the holistic observations of the teacher as well as describe any unpredicted activity of the learner. Within the quantitative design, standardized measures are formulated as a general framework to monitor a range of conceptual and procedural activities.

The ATOSU criteria are theoretically oriented and independently score a set of four mutually exclusive dimensions of problem-solving procedural progress -- namely, [A]ttention, [T]actics, [O]perations, and [S]olutions -- which recognizes a hierarchy of problem-solving levels within the context of a U [Unique] task/problem. These procedural criteria seem more appropriate for process assessment than product criteria since they monitor changes of activity over chains of thought (see Guskey, 1996). Procedural criteria present the perspective that grading should describe “how students are performing” rather than merely what they know at some given time. For example, analyzed in sequence, the ATOSU general criteria monitor specific problem-solving activity as it unfolds, taking into consideration the [U]nique conditions under which this activity is observed. These process criteria, like any product criteria, may be followed over time to consider how a student’s thinking processes are improving as a result of his/her learning experiences.

Feedback filters the specific information through the ATOSU general framework so that the student, in a reflexive fashion, constructs or reconstructs structure given changes in the context. Reflexive thinking requires that the learner consider his/her internalized
thoughts as objects to be acted upon (Blumer, 1969). Humans need to organize properties of tasks using their conceptual representations, or they would be overwhelmed by the complexity of their environment. Learners can comprehend competing views, can make sense of similar and different relations between these views, and can express their thoughts about a chosen position. Given an unfamiliar problem, students must possess a starting point from which to begin their search. This starting point constitutes a conjectured and generalized structure derived from previous experience about a given situation that may be refuted, modified, or regenerated. There are times when students must reach beyond an unthinking application of rules and reconfigure their conceptual and procedural structures.

Even though some educators seem averse to some forms of evaluative rankings (see the NCTM Assessment Standards, 1995), the criteria comparisons developed here are viewed as a necessary step in assessment practice. How is it possible to understand anything of a learner’s progress without comparing the individual to cohorts or without comparing the student’s past performance to his/her current performance? Comparisons seem necessary in order to make the judgments necessary for a student’s growth. Some levels of conceptual understanding may be judged superior to other levels of conceptual understanding; some procedural sequences are more advanced than others; some tasks may be more complex than other tasks. When making any comparison, one judges whether conceptions, procedures, or task A are in some way better or worse than other conceptions, or procedures, or task B. Therefore, despite the potential concern for ordering phenomena expressed by the NCTM, comparing or ranking students and tasks still seems necessary for
understanding the quality of student performance. However, in agreement with NCTM, grading and reporting should always be done with reference to the learning criteria that reflect the mental activities in the task performance.

2.5 How Guided Assessment Impacts Achievement

Two principal ways teacher assessment practice are predicted to impact skill achievement within this study are next developed under the headings, Scoring Approach and Graded tasks. Both forms of quality control are developed and adopted here since it seems unrealistic to expect large effects on broadly-based achievement tests from any one assessment practice. This caveat seems necessary since interest in classroom assessment and its promise is at an all-time high. In fact, even though the empirical evidence regarding classroom-assessment’s impact on a district’s or state’s achievement seems limited, many researchers and practitioners consider assessment reform to be at the very foundation of school reform (Cizek, 1995; Wiggins, 1992).

2.5.1 Scoring Approach: First, if utility is defined by gains in a standardized test, then improving the utility of a classroom-assessment seems, in part, dependent upon the quality of the scoring approach. Assessment practices produce gains in standardized achievement when they provide meaningful and usable feedback to teachers, parents, and students. However, due to the concerns over ethical test practices, little research has been done regarding classroom-assessment’s impact on standardized tests. Much of the empirical work examining assessment’s effects on standardized achievement comes from mastery
learning research, which utilizes the traditional theory of dimensional measurement. This research largely recognizes the importance of instructional variables like cues, feedback, participation, and reinforcement as important factors for increasing performance (e.g. see Bloom, 1984 for a review). The basic idea of mastery learning is that the specific learning deficits of the learner need to be remediated immediately rather than being allowed to accumulate. Students are not permitted to advance to the next lesson without demonstrating some predetermined level of understanding (mastery) on the current lesson. Remediation encourages further practice and additional instruction until the learner is better prepared to pass a new test.

Although mastery learning is designed to provide frequent and continuous feedback in which students work at their own pace, there is a question as to whether mastery learning is effective at generating increases in standardized achievement. Some investigators have found modest to nonexistent effects in standardized achievement based on experimental studies lasting four weeks (summarized in Slavin, 1987). Two studies are most germane given the current proposed research. The first study was from a dissertation done by Kersh (1970) using two schools, one from a lower socio-economic-status (SES) class and another from a middle SES class. This study of mastery learning was performed in two elementary schools in which 11 fifth-grade classes were randomly assigned to a Mastery learning or a Control condition. Students were assessed each month and, for those who fell short of the mastery criterion, were remediated by peer tutoring, games, and alternative activities. Control classes were untreated. The study did not yield any meaningful differences between the Control and Treatment conditions on the Stanford
Achievement Test's Concepts and Applications scales. Effects were more positive in the lower SES class school than in the middle SES class school, but no effects were significant.

The second study of mastery learning is the year-long Anderson (1985) experiment. This mathematics study compared students in grades 1-6 in another between-school design, with the treatment and control conditions nested within schools. Students were matched on the Metropolitan Reading Test (grades 1-3) and the Otis-Lennon Intelligence Test (grades 4-6). In the mastery school, classes were presented with a lesson and then specific progress was assessed relative to a stated objective. Group re-learning or review sessions followed, using the test results as a means of focusing on the proper objectives. After every student reached mastery on the unit, the class moved to the next unit. The experimenters employed the Computations, Concepts, and Problem-Solving scales of the California Achievement Test as their dependent variable, administered at the end of the year. The Mastery learning group scored somewhat higher than the Control group on Computations and Problem-solving, but the Control group scored higher on Concepts. Only the Computation differences were significant.

Although there may be alternative interpretations, it seems possible that the disappointing results viewed in the above mastery-based studies can be constructed to the quality of the assessment, especially its effectiveness in improving corrective instruction. If assessment methods—like those developed from structural theory—are to serve beyond their auditing function and produce information that tells participants more about how students solve problems, then it seems necessary to adopt scoring procedures that go
beyond measuring dimensions. With this argument in mind, the next group of studies make a possible case for performance-based assessment that utilizes and goes beyond structural theory.

Item-based forms of assessment rely more on the comparative judgments of teachers over criteria compared with the total-score method of dimensional theories. Hodge and Coladarci (1989) reviewed 16 studies comparing teacher judgments of academic achievement with standardized tests. Teacher judgments of student constructs, such as test performance, and group classification, such as gifted/not-gifted, showed a median correlation with standardized tests of achievement of 0.62. Achievement judgments were regarded as generally accurate, with levels of accuracy varying among teachers. The review concluded that teachers’ judgments be given more weight relative to other forms of evaluation, and that these judgments be further improved with increased teacher development. These findings demonstrate the value of the teacher as an observer and the importance of his/her potential role in improving the assessment process.

A major question that confronts mathematics educators is how teachers can modify their instructional and assessment practices to support and encourage student’s thinking as it changes. Research on cognitive-guided instruction (CGI) conducted over an extended period points to the role of teacher development and increased knowledge in gathering the information necessary to make these adjustments. In a longitudinal study of 21 primary teachers, Fennema, Carpenter, Franke, Levi, Jacobs, and Empson (1996) found that instruction should begin with a student’s informal ideas as a basis for building formal concepts and procedures. With CGI, the teachers are encouraged to ask students to
volunteer their ideas and work cooperatively. Teachers, in turn, encouraged the students to engage in different problem-solving practices, discussing their observations of the children's thinking using a specific model. As a result of these efforts, achievement in problem-solving and mathematical conceptions improved on instruments designed for the study.

Research across four major projects with similar problem-solving approaches was recently reviewed by Fuson, Weane, Hiebert, Murray, Human, Olivier, Carpenter, and Fennema (1997). When students were encouraged to invent procedures or operations and integrate them with conceptual supports (like playing cards, geoboards, or cuisenaire rods) in classrooms, students increased their ability to solve multidigit numbers in all four projects. In each project, teachers did not teach a single algorithm for solving multidigit problems. Rather, the learning of concepts and procedures was accomplished by interaction within groups of students and an occasional teacher intervention. Teachers did not transmit a set of direct rules or operations as to how to solve multidigit problems. Alternatively, teachers expected students in different groups to construct different procedures. In each study, a conceptual framework was then used to describe and classify the different procedures student groups constructed. Errors that arose during learning were identified. It was observed that conceptual structures classified as most sophisticated were products of non-directive instruction where students simply recomposed their strategies. The authors advise that students be encouraged to write down their thoughts because their notations serve as memory supports and extensions of these practices to larger numbers.
When considered in its entirety, the above research left the investigator with some important building blocks for considering the future directions of classroom assessment. Quality, timing, and focus of feedback could be improved by devising new scoring methods that chart conceptual and procedural thinking by better utilizing teacher judgments in both the observation (qualitative) and measurement (quantitative) designs. If classroom assessment is to fundamentally influence instruction in ways that increase future achievement, then effective performance seems unlikely without the necessary good judgment (Wiggins, 1993) formed by classroom-level teacher development (Fennema, Carpenter, Franke, Levi, Jacobs, & Empson, 1996).

Any scoring method designed to provide feedback is ultimately concerned with how it influences decision-makers. Rather than merely administering a posttest, assessment practice could be enhanced if teachers were to observe and record their observations as instruction transpires. By embedding the assessment within instructional activity, assessment “permeates the learning process” (Resnick, 1994), allowing for the timely self-reflection of all participants, as well as the sharing and communicating of ideas. Students should also be encouraged to use written comments, along with their mathematical strategies and computations. Therefore, assessment could be used to certify levels of instructional accomplishment given certain problem conditions, subsequently providing information that redirects the learner and refocuses instruction.

2.5.2 Graded Tasks: The second way assessment practice impacts mathematics achievement recognizes the conceptual quality and gradation of the problem-solving tasks
students are being asked to perform. Students who are given instruction aimed at
cognitive understanding do better on skill tests than students drilled on the skill directly.
Repetition deleteriously affects low-achieving students because they are assigned
indefinitely to dull drill that does not permit them to learn deeper concepts (Levin, 1987).
The assessment should evaluate the connectedness of concepts and procedures as students
employ them during problem-solving given a variety of situations. To accomplish this end,
a teacher needs to increasingly observe and make judgments about a pupil’s ability to carry
out activity when producing an answer. Both traditional classroom and large-scale testing
programs are limited by quality considerations that undermine their importance when
evaluating important learning outcomes. They are both restricted by their emphasis on
knowledge and basic skills, limiting the breadth and depth of the thinking being tested.

Task conditions need to be altered by the teacher in ways that reflect an ability to
think conceptually and procedurally in a variety of novel situations. Although basic
probability concepts, such as the laws of addition or multiplication, can be taught with
reasonable success, students’ knowledge about the use of probability concepts in
unfamiliar contexts is incomplete and not integrated (Konold, Pollatsek, Well, Lohmeier,
Lipson, 1993; Konold, 1989; Nisbet, Krantz, Jepson, & Kunda, 1983). Different problems
access different pieces of knowledge, so altering the content demands of the task impacts
scores. Asked for the most likely outcome of a fair flip of a coin, given four successive
trials on which the coin landed tails up, 47% of the 7th graders and 56% of the 11th graders
selected the correct alternative. However, when the task was altered by changing problem
contexts, only 38% of the 7th graders and 33% of the 11th graders chose the correct answer
(Brown, Carpenter, Kouba, Lindquist, Silver, Swafford, 1988). Thus one question on the National Assessment of Educational Progress (NAEP) may evaluate a student’s understanding of event independence in one way, but another question testing the same idea in another way may note an inconsistency in the student’s mastery. As new features of the problem begin to deviate from the original example, students may revert to nonstatistical and inappropriate ways of reasoning. Thus, correct responses on test items can sometimes be spurious and may not reflect the student’s deficiencies in understanding the concept of event independence in probability problems.

Bruner (1973) argued for levels of mastery given the degree the defining properties or cues are represented by the conditions of the problem (also see Perkins, 1992; English, 1992). Using the acronym, BIG, standing for “beyond the information given”, Brunner described how human cognition must go beyond what is directly learned in instruction. After direct instruction, BIG tasks demand that the learner work through his/her confusion in many ways to extend meaningful inferences to related tasks. These related tasks might demand that similar operations be used in new contexts or that present operations be distinguished from previously learned operations within the same context. The structure of the task is varied in subtle ways to note how learners alter their previously acquired conceptual and procedural structures to reconcile similarities and differences in task conditions. In Appendix A, the ATOSU scoring rubrics utilize this BIG approach to quantify differences between lower levels of conceptualization (cued performance) and higher levels of conceptualization (extended performance) in the [T] and [O] components.
Sensitive to the problem’s features and the procedures by which one’s conceptions are transformed, an ill-structured problem forces the learner to construct a new representation due to lack of direct instruction or previous experience with the task or task situation. In contrast to BIG, WIG stands for “without the information given,” since the learner is not permitted to rely upon previously learned conceptions or procedures (Perkins, 1991). Here the learner is acting more from intuition as a starting point and is forced to invent a conceptual or procedural structure as a condition for solving the task. Therefore, both the WIG and BIG tasks are influenced by how specific conditions of the problem are simply recognized and understood, given the instruction. When no direct instruction is provided, a WIG performance is more likely noted. When previous direct instruction is provided, a BIG performance is observed. Once more, the instrument’s rubrics operationalize differences in extended and generative thought by stipulating and predicting that WIG performances are conceptually more complex than BIG performance.

To be usable, assessment feedback cannot be arbitrary. It must be focused so that specific conditions may be related to a general framework during activity. The idea is that varying the conditions of the task will result in different task demands of the corresponding conceptual and procedural processes. In mathematics, students receiving individual or group feedback do so through verbal, visual, and mathematical communication. This communication is nurtured by the social interactions of the teacher with the student, or the student with other classmates, using external objects to represent situations. Within this context, the student must be pressed to continually defend their thinking. Likewise, it also allows for feedback that is specific to the student’s abilities or a task’s complexity defined
by a problem’s conditions. In this manner, the teacher can observe both repetitive themes of activity as well as the exceptions to expected actions taken with each new task. Understanding and communicating how activity is similar and different, given the conditions of the task, creates a “zone of proximal development” where learners perform within their range of competence, while being assisted in realizing their potential levels of performance (Wertsch, 1985).

2.6 Promoting Gender Equity in Assessment

Equitable judgments about students are made when the assessment accounts for unique differences in how students learn and communicate knowledge (National Assessment Standards, 1995). Although teachers seem to be capable judges (Hoge & Caladarchi, 1989), they are as susceptible to prejudicial behavior as any human being. Part of this may be attributed to differences in training and part to differences in the teacher’s unstated beliefs about what constitutes quality assessment (Bond, 1995). If alternative scoring methods are to be developed, one must ask the question: To what extent are traditional and performance-based methods of scoring influenced by the unintended consequences of gender bias?

In developing scoring models, it is important to examine the scoring rubrics for possible unintended consequences that occur as a result of using the method. Contemporary studies have revealed little evidence of gender differences in arithmetic ability; however, there are still differences in spatial reasoning, figural reasoning, and mechanical reasoning that favor boys as they reach high school age (Feingold, 1993).
These observations were found consistently across norms for the *Differential Aptitude Test* and the *Iowa Test of Basic Skills*. Comparing the means of the 1985 national norms of the *California Achievement Test*, Feingold even found advantages favoring females on all batteries of the test, yet these differences decreased with age. Feingold concluded that advantages (e.g., arithmetic and figural reasoning) once held by males have diminished and parity in mathematics achievement now exists.

Meta-analysis is a statistical technique used to compare replications of similar substantive and methodological research. A review of two recent meta-analyses reveals a more complete picture. Hyde, Fennema, and Lamon (1990) examined age and gender trends in gender differences in three mathematics dimensions: computation, concepts, and problem-solving. In tests of computation, females have an advantage ($d = 0.20$ effect size) until about the age of 15 and than perform about the same as boys. There were no differences in mathematical conceptions throughout the school years, but males held an advantage in mathematical problem-solving ($d = 0.30$ effect size) by age 15 and sustained it into early adulthood.

The national data for average achievement on the NAEP for male and female students in grade 8 for both 1990 and 1992 tests indicate no gender differences. About the same percentages of males and females were estimated to reach proficiency level for eighth grade — about one-fourth. Essentially, the same proportion of males and females (about 60 percent) reached the Basic level for the eighth grade (Mullis, Dossey, Owen, Phillips, 1993).
Linn and Peterson (1985) compared the spatial ability of males and females using process analysis to classify spatial measures into three categories: spatial visualization, spatial perception and mental rotations. Spatial visualization is the ability to interpret visual relationships with regard to objects. Spatial visualization measures include tests of embedded figures, paper form boards, and block designs. The meta-analysis detected no significant differences between boys and girls in spatial-visualization performance. Spatial perception is the ability to detect spatial relationships with regard to one’s point of orientation. The authors found that male superiority in spatial perception varied with age. Among children (under 13) and adolescence (13-18), the effect size was about 0.37. For adults, the effect size was much larger ($d = 0.64$). Mental rotations involves an ability to rotate figures in two or three dimensions. Males again held a small to significant advantage in performing mental rotations, varying across instruments: 0.94 for the Shepard-Metzler and 0.26 for the Primary Ability Test.

Although fundamental differences on numeric measures of achievement and ability once favored boys in mathematics, these advantages are either rapidly diminishing or are altogether disappearing with time. For this reason, since performance on most numeric measures of mathematics are not or are slightly gender-related, any performance differences on performance-based assessments that are only gender related would need to be examined. One notable exception related to mathematics may be in the performance of spatial ability tasks. Important here is that future assessment in mathematics encourages early experiences with visualized tasks with objects and models (erector sets) to simulate these conceptual relations (Linn & Peterson, 1985).
Since performance assessment is more dependent on open formats and judged qualities rather than on closed formats and total scores, there is concern that judging be free from bias. If the ratings of performance are influenced by extraneous factors other than scholastic achievement, than the judged performance is less accurate. For example, one dilemma is over the joint influence of student behavior and gender on teacher judgment. Jussism (1989) found that the teacher’s initial perceptions of student behavior predicted their mathematics grades, even after controlling for previous achievement scores. Likewise, other traits may confound judgments of accurate performance like attendance, age, effort, race, disability, or culture.

It has been observed that perceptions of achievement may be jointly affected by student’s gender and classroom behavior, with boys perceived as less academically adequate than girls in the judgment of teachers (Bennet, Gottesman, Rock, Cerrullo, 1993). Hodge and Coladarci (1989) found three studies in their review that revealed that ratings of academic performance were confounded by gender and student behavior, with boys receiving lower marks from teachers. However, contrary to these results, Bernard (1979) experimentally varied the reported sex and sex role of high school writers’ essays and found that teachers preferred the work of boys and students with masculine behavior over girls’ written work.

Collectively, these results suggest that gender differences need to be monitored with respect to both classical and performance-based assessment. In the present study, gender differences on both the guided assessment treatment and the dependent variables
are examined both quantitatively and qualitatively. Any possible gender differences or gender/treatment differences will be analyzed and discussed.

2.7 Summary

This Chapter has reviewed two dominant forms of measurement theory: classical dimensional theory versus an alternative structural theory, and proposed a new guided-feedback system ATOSU. Dimensional theory is used more in classical forms of scoring, as when a total score is presented in objective testing; structural theory is used to rate and order cognitive processes and tasks as applied in performance assessment. The proposed ATOSU scoring framework is constructed as a structural theory of measurement that can be used to guide student learning to problem-solve. The presentation in this review has argued as follows: An assessment framework that is arbitrary, that does not search for meaningful cognitive patterns and connections, that does not construct structure out of apparent disorder, and that is inequitable to protected groups does little to provide valid information for the teacher who wishes to encourage guided and generalized student learning. Tasks can be made more challenging by altering the task's structural conditions and context.
Chapter 3
Methodology

The [A]ttention, [T]actics, [O]perations, [S]olution, and [U]nique-conditions framework is proposed as a guided-assessment model that can be used to provide feedback to students, faculty, parents, and school administrators. This study examined the utility of this guided feedback using as dependent variables scores from a mathematics achievement test designed from items supplied by the National Assessment for Educational Progress (NAEP) covering probability and statistics problems. Both qualitative and quantitative methods were employed to produce an individualized assessment-based feedback Treatment designed to promote problem-solving understanding for any student, whether male or female. ATOSU-based information was used to provide guided feedback to enhance mathematics achievement. This chapter begins by reviewing the overall research purpose, the specific research questions, and the appropriate research hypotheses necessary to guide the analyses of the data. This overview is followed by a discussion of the design, sample, and instrumentation, treatment, experimental procedures, proposed analyses, and summary.

3.1 Overall Purpose of the Study

The purpose of this study was to examine the utility of the ATOSU model as a feedback and learning tool. Three sets of research questions drive this inquiry as enumerated below. The first set of research questions addresses measurement issues of validity and reliability; the second set of questions addresses experimental hypotheses related to the model’s ability to enhance achievement; and the third set of questions
addresses *qualitative themes* described within the study which can be used to clarify interpretations and cross-check results.

**3.1.1 Are the variables utilized in this study measured validly and reliably?**

1. Are the five ATOSU constructs measurable? Are the tasks reliable and content valid?
2. Is the hierarchical nature of the five ATOSU scales validated in this study? Is this hierarchical structure reliable?
3. Can the ATOSU model be implemented reliably by judges?
4. What is the reliability of the conceptual, procedural and problem-solving scales used as the dependent variable?

**3.1.2 Are there treatment and gender differences in the experiment? Are there interaction effects between treatment and gender?**

Hypothesis 1: Students who are given guided assessment using the ATOSU model will perform better on the dependent conceptual, procedural, and problem-solving scales of the achievement than students who do not receive guided assessment.

Hypothesis 2: In general, there will be no gender differences in conceptual, procedural, and problem-solving scales regardless of the treatment.

Hypothesis 3: There were no differences in conceptual, procedural, and problem-solving performance as measured by the treatment-gender interaction effects.

**3.1.3 What are some major themes illustrated by the qualitative data relative to the ATOSU attributes? Do the qualitative data support the use of the ATOSU protocol?**

What follows is a description of the design, sample, experimental treatments, instrumentation and other data sources, followed by the experimental procedures. The final parts of this chapter present brief descriptions of the measurement, experimental, and qualitative analyses planned for this study.
3.2 Design

Three different designs were employed in this study to obtain the data necessary for analyses. A *measurement design* defined a plan for investigating the reliability and validity of the measures; an *experimental design* defined a plan for investigating the effects of a set of treatments; and an *observational design* defined a plan for obtaining qualitative data so that common themes may be described, interpreted, and verified.

3.2.1 Measurement Design: A scoring matrix was employed to organize the scoring of the ATOSU steps by judges. This scoring entails the rating of the ATOSU five stages in terms of levels of conceptual and procedural thinking — specific to each stage. Each row represents a different task requiring graded conceptual thinking. A *conceptual vector* is associated with a measured set of scores listed down the rows of the data matrix for a single ATOSU attribute. For example, the teacher could compare [A]ttention scores for a student across a graded series of problem-solving tasks or across different students. A teacher can also compare student #1 with student #2 on the [A]ttention stage; noting that student #1 performs at a much higher level (score of 4) than student #2 (score of 2). Similarly, a teacher can also note student differences across tasks — e.g., student #1 scores 4 on the [A]ttention stage on task 1, but scores only 3 on task 2. When examined down a particular column, the teacher could compare levels of generalized growth in the [A]ttention attribute given some task. The four tasks provide for breadth in conceptualization, and represent four different kinds of statistics and probability problems including: the probability of an event, the rule of addition, combinations, and calculating and interpreting the mean, median, and mode (see Appendix B).
A procedural vector is defined by a sequence of ATOSU scores that may be characterized or monitored from task to task or from person to person. These scores are listed across the columns of the scoring matrix in Table 3.1. Teachers can analyze problem-solving progress by examining the procedural vectors derived from the successive columns of the scoring matrix for student #1 in Table 3.1. By examining a series of ATOSU scores across the rows for a student across tasks, or for a task across students, the teacher can observe change in the procedural vector of the matrix. For example, the procedural vector of 43221 for student 1 on task 1 is superior to the procedural vector of 22111 for student 1 on task 3. Likewise, student 1's procedural vector on task 1 (43221) could be compared to student 2's procedural vector on task 1 (22101) to note between-individual differences.

A judging plan was implemented to examine rater bias and error. Table 3.2 displays an unbalanced judge matrix used as the judging plan in the measurement phase of the study. The judge matrix presents three judges rating eight different students undergoing the two treatments. The unit of analysis in this study is the student. All tasks and students were rated by two of the three judges, thus no student is rated twice by the same judge on the same task. As per Table 3.2, four students were judged by the researcher and the Title 1 teacher, while a second group of students was judged by the researcher and the mathematics teacher. An equal number of males and females was judged from the ATOSU and Test-only groups.

3.2.2 Experimental Design: A posttest-only factorial (3x2) experimental design was employed within an experiment as viewed in Table 3.3. This 3x2 factorial design was
### Table 3.1: Scoring Matrix

<table>
<thead>
<tr>
<th>Student</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Task 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>student 1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Task 2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Task 3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Task 4</td>
<td>3</td>
<td>2</td>
<td>3*</td>
<td>2</td>
</tr>
</tbody>
</table>

Conceptual vector
*possible noise

### Table 3.2: Judge scoring matrix for the Facets analysis.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Student/Gender</th>
<th>Judge/Rater</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOSU</td>
<td>1 male</td>
<td>Judge # 1,2</td>
</tr>
<tr>
<td>Treatment</td>
<td>1 female</td>
<td>Judge # 1,2</td>
</tr>
<tr>
<td>(X1)</td>
<td>1 female</td>
<td>Judge # 1,2</td>
</tr>
<tr>
<td></td>
<td>1 male</td>
<td>Judge # 1,3</td>
</tr>
<tr>
<td>Test-only</td>
<td>1 male</td>
<td>Judge # 1,3</td>
</tr>
<tr>
<td>Treatment</td>
<td>1 female</td>
<td>Judge # 1,3</td>
</tr>
<tr>
<td>(X2)</td>
<td>1 female</td>
<td>Judge # 1,3</td>
</tr>
</tbody>
</table>
employed to examine the impact of three levels of treatment (ATOSU assessment with tutoring vs. tutoring-only vs. conventional instruction) and two levels of gender (male vs. female) on a mathematics achievement test with conceptual, procedural, and problem-solving scales. The posttest was constructed using items found in the 1990 and 1992 versions of the NAEP tests provided by Educational Testing Service. Subjects were seventh graders from a school Title I program.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Treatment</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>R X1</td>
<td>ATOSU</td>
</tr>
<tr>
<td>female</td>
<td>R X1</td>
<td>ATOSU</td>
</tr>
<tr>
<td>male</td>
<td>R X2</td>
<td>Tutored-only</td>
</tr>
<tr>
<td>female</td>
<td>R X2</td>
<td>Tutored-only</td>
</tr>
<tr>
<td>male</td>
<td>R*</td>
<td>Test-only</td>
</tr>
<tr>
<td>female</td>
<td>R*</td>
<td>Test-only</td>
</tr>
</tbody>
</table>

*Randomly sampled rather than randomly assigned

Table 3.3: Factorial Experimental Design

During the experiment, the researcher individually instructed each student from the tutored and the ATOSU groups over three lessons covering instructional goals and tasks specified in Appendix B. To perform well, students needed to combine their conceptual and procedural skills to solve mathematics problems. The conceptual, procedural, and problem-solving scales presented on the achievement test (see Appendix C) were assumed to be theoretically similar to the conceptual, procedural, and problem-solving attributes assessed by the ATOSU model.
3.2.3 Observation Design: Finally, an observational design employed qualitative methods to explore expected and unexpected activity as related to the ATOSU framework during learning and problem-solving. The qualitative approach utilized in this study was designed to observe events, behaviors, and activity in ways that organize complex reality. Since qualitative methods do not always focus on the same phenomena as their quantitative counterparts, qualitative data were also used to help evaluate the utility of the ATOSU framework. With qualitative description, one uses the ATOSU framework to categorize and count observed instances of activity during the lesson. Audiotapes of the lessons recorded discourse as it naturally occurred in guided assessment. Snippets of this discourse and the actual student responses were then used to illustrate how the framework was employed. Any trained observer could look for trends and record repeatable regularities and irregularities in instructional events and problem-solving activity.

3.3 Sample

This study employed two sampling strategies: one sampling strategy was chosen to address questions related to the measurement and qualitative phases, while a second sampling strategy was chosen to conduct the experiment. The size of the samples and the number of classes involved were limited in both sampling strategies by the time, money and resources required to do the study.

The first sampling strategy was used to examine the rating scales and elucidate the qualitative themes. A “typical case” sample of the study’s participants was chosen from the original sample for audiotaping based upon convenience, the recommendations of key
informants like the Title I teacher or the classroom teacher, and the individual subject's willingness to be recorded (see Patton, 1990). This subgroup of students represents and highlights what is considered typical, normal, or average with regard to the Title I population. Eight students were audio-recorded, comprising 2 males and 2 females from the Test-only group and 2 males and 2 females from the ATOSU group. This purposive sample was examined to note possible similarities and differences with regard to inter-rater agreement, to examine the learning processes employed by these students, and to illustrate what is typically done with each guided-assessment group to those unfamiliar with the research. For these reasons, this sample was illustrative and not definitive.

A second sampling strategy was employed in the experiment to acquire a representative sample of 76 students from seventh grade classes in a single school. Medina is one of a number of schools described as "conventional" by the Columbus City School System. This minority balanced school is located in a working-class neighborhood and has a high proportion of students receiving free and reduced lunches distributed to seventh graders (8 reduced-cost lunches +107 free lunches/ 223 seventh graders = 52 %). The 52 subjects randomly assigned to the Tutored-only and ATOSU treatments consisted of students participating in a Title I program funded by The 1994 Improving Schools Act at Medina Middle School within the school district. Students eligible for this program must come from low-income backgrounds and demonstrate an educational need in reading, language, or mathematics. A Title I teacher provides supplemental help in small groups to assist each child in meeting curricular standards. Guided assessment, as viewed within the ATOSU framework, seemed like an appropriate treatment for such a population since it provides multiple data sources, structured practice, and focused feedback.
The students were first divided into males and females and then randomly assigned to one of two Treatment groups, designating an equal number of boys and girls for each group. One group of 26 students was tutored, while a second group of 26 students was both tutored and assessed with the ATOSU model. Initial differences between these two groups were assumed to be eliminated by assigning subjects on a random basis to both experimental conditions. A third group of 24 students from the same school was randomly selected, not treated, and merely tested one year later. This meant that the total number of students was 76 — 24 randomly selected and 52 randomly assigned to the three respective groups. Since all statistical tests assume independent observations, both the Tutored-only and the ATOSU students were taught separate from their class, one student at a time. The unit of analysis for this research was the subject.

During the experiment, students were loss in the ATOSU and Test-only groups. Unequal cells are not orthogonal and it is therefore not possible to partition the sums of squares as separate and independent components, so the treatment and interaction effects were correlated (see Pedhazur, 1982). To balance the experimental design after subject mortality, subjects were randomly removed from the overpopulated cells until there were equal cell frequencies (11 students per cell x 6 cells = 66 subjects).

3.4 Experimental Treatments

A unit of three lessons covering probability and statistics was taught to each student in the Tutored-only and the ATOSU group, but not to the Test-only group. A deck of ordinary playing cards was employed to represent concepts and procedures for both groups
Appendix B enumerates goals and specifies targets for each of these lessons.

During the first lesson, students were shown the relationship between fractions and the study of chance. Playing cards were utilized to represent this idea by demonstrating how probability events are subsets of a broader set of possibilities. The number of ways an event may occur were counted and placed in the numerator of the fraction, and the number of possible outcomes were placed in the denominator. Students were shown that the number in the event could never exceed the total number of possibilities. Students were finally shown how to employ the rules of multiplication and addition to operate on and solve problems in probability using the cards or coins to verify the answers.

In the second lesson, students were given a set of nine numbered cards to learn how to calculate the mean, median, and mode. The cards were placed into a box and students were asked to randomly draw three cards and calculate the mean. Students then compared the mean of their random sample of three cards to the population mean of nine cards. The number of cards in the population and sample could be changed as the lesson progressed.

The third lesson combined the two concepts of probability and statistics by practicing their relationship within the cards or within practice problems. Students were asked to practice using the rules of addition and multiplication to solve a variety of problems while employing playing cards as a means of representing the problem space. Students might use only the face cards to symbolically represent the number of possibilities in one problem, or the number cards when symbolically representing the number of possibilities in another problem.
The "Tutored-only" group was shown how to solve a variety of probability and statistics problems. These demonstrations were largely teacher-directed so there was more emphasis on teacher talk at the expense of guided student activity. Relative emphasis was placed more on teaching about problem-solving rather than via problem-solving. Students were taught to methodically employ operations a step at a time. They were not encouraged to discover new approaches or to analyze the situations in light of their existing knowledge.

When using the ATOSU framework during guided assessment, more opportunities were provided for guided and monitored practice that could be evaluated. In contrast with the Tutored-only group, the teacher encouraged students in this "ATOSU + tutored" group to discover or observe patterns while working with the cards. By teaching via problem-solving, students were encouraged to demonstrate their attention by verbally describing and identifying relevant aspects or dimensions of the problem. Students were next asked to tactically represent the problem with pictures, diagrams, tables, or graphs. The teacher then modeled how these verbal and tactical concepts connect to the operational or mathematical procedures used to solve probability problems. Since the student's activity must lead to learning, students were challenged to query, reason, and make connections, while the teachers provided feedback. The ATOSU group tried some additional guided practice over some mathematics problem examples at the conclusion of each lesson. These activities were assessed by the researcher and, on at least eight occasions, by an observing teacher, and the results were shared with the students.

The "Test-only" group received no instruction beyond what was learned about probability and statistics in the classroom or in the Title I program. For this reason, the
group's achievement is expected to measure conceptual, procedural, and problem-solving knowledge learned without the tutoring treatment. Additional instruction may have occurred in previous classroom lessons or in previous experiences learned within real-life situations. In many ways, the timing and intensity of these previous lessons and experiences may be different from student to student.

3.5 Instrumentation and Data Sources

Three sources of evidence combined data collection procedures over the course of this study. The first two sources are the ATOSU instruments and achievement test, and these are considered primarily quantitative in nature. A third source of data is strictly qualitative since it consists of audio recordings of students receiving instruction that were selectively transcribed and analyzed. These data sources were triangulated to check out the consistency of the findings generated by different data-collection methods.

3.5.1 ATOSU Assessment Protocol: Feedback is structured by an assessment protocol (see Appendix A) to fit the purposes of guiding problem-solving. This protocol is designed to supplement ongoing communication between the instructor and student, thereby monitoring probability conceptualizations and the problem-solving procedural progress made with the playing cards. Feedback begins with the student's own understandings of fractional concepts as related to partitions of the playing cards, and it proceeds with the interpretation of probability and statistics strategies/rules in relation to these objects. In the beginning, students are cued to use the correct strategy and/or rule. As students become more
comfortable or fluent with problem-solving conceptualizations and procedures involved in the card tasks, students are cued less and less. After time and practice, they were asked to solve new tasks that extend previously taught rules and strategies beyond the familiar form.

The five ATOSU constructs are defined as a sequence of qualitatively different stages of thinking and expression. These scales are behaviorally anchored using specific descriptors that predict progressively more complex forms of activity. Monitored individually, each of the [A], [T], [O], [S], and [U] scores is considered a rating of some procedural ability as the student presents his/her understanding and the teacher assesses and guides it. A is a univariate rating of [A]ttentional skills viewed as language communication; T is a univariate rating of [T]actical skills viewed as visual communication; O is a univariate rating of [O]perational skills viewed as mathematics communication; and S is a univariate rating of depth of [S]olution as forms of communication are transformed. Monitored as a sequence of scores, the ATOS constructs represent an ordered sequence of procedural ability that has a multivariate distribution. The [U]nique-conditions can be altered to produce more challenging tasks (see Appendix B). In this study, there were no practice items where the context condition was dramatically altered. Students lacked sufficient practice to produce the depth of understanding necessary for higher levels of transfer. The researcher scored each of the four tasks with the ATOSU scoring model during the experiment, with two additional raters independently rating eight students on these same tasks, to note judge agreement or disagreement. ATOSU profile scores were used, then, to supplement instruction by promoting better student understanding and decision-making.
3.5.2 Achievement Test: Subjects were posttested on an achievement test consisting of 19 items covering probability problems provided by the National Assessment of Educational Progress (see Appendix C). Federally financed and operated by Educational Testing Service, NAEP conducts a regular “census” of U.S. student achievement in various subject areas. This census has served as an indicator of change in elementary and secondary school achievement over the past two decades (Mullins, 1990). The mathematical abilities described earlier in this chapter for the 1990 and 1992 NAEP assessments (procedural knowledge, conceptual understanding, and problem-solving) specifically address aspects of knowing and doing mathematics. There are three forms of the test that may be administered to fourth, eighth, and twelfth graders. The Grade Eight form of the achievement test was considered most appropriate for the seventh graders in this study since the test was to be administered at the end of their school year. With permission from the U. S. Department of Education, items were chosen based upon the table of specifications and appropriate content categories devised by Educational Testing Service covering probability and statistics for the testing years 1990 and 1992.

NAEP involves many distinct systematic samples of students, thousands of schools, and hundreds of thousands of individual students. For example, in 1992 alone, there were about 419,000 students comprising 12,000 schools included in the sampling procedures (Mullis, Dossey, Owen, & Phillips, 1993). These samples included students who were learning disabled, hearing impaired, blind, limited English Proficiency, and in adjusted curriculum classes. These students were tested in accordance with their Individualized Education Plans (IEP’s) and were dropped when the IEP team determined that the student
was incapable of participating meaningfully in the assessment. Students within participating
schools across the country who are part of today’s Title I programs would have been
considered as part of the NAEP’s population both in 1990 and 1992. However, since
individual items used in this study predate the 1994 Title I Act, item information about this
subgroup is nonexistent. Today, Title I students continue to be included within the NAEP
samples and are never rejected for testing based on their participation in this program.

Conceptual and procedural skills are at the heart of mathematical problem-solving
(Hiebert, 1992) because problems are often represented by both these diverse but related
skills. For the sake of clarity, the framework used by the Educational Testing Service (ETS)
to operationalize conceptual, procedural, and problem-solving abilities is presented here.
Because ETS wished to be consistent over time, the assessment framework is the same for

**NAEP concepts:** This scale measures whether the student can “generate examples
and counter-examples; can use and interrelate models, diagrams, and varied
representations of concepts; know and apply facts and definitions, can compare,
contrast, and integrate related concepts and principles; can recognize, interpret, and
apply the signs symbols and terms used to represent concepts; and can interpret
assumptions and relations involving concepts in a mathematical setting” (Mathematics

**NAEP procedures:** “Students demonstrate procedural knowledge in mathematics
when they provide evidence of their ability to select and apply appropriate procedures
correctly; verify and justify the correctness of a procedure using concrete models or
symbolic methods; and extend or modify procedures to deal with factors inherent in
problem settings” (Mathematics Objectives 1990 Assessment, 1988, pp. 16-17).

**NAEP problem-solving:** This scale examines students’ reasoning and analytic
abilities when they encounter a new situation (Mathematics Objectives 1990
Assessment, 1988, p. 17.) As such, this scale measures student progress under a
variety of changing situations or conditions.
Following the guided assessment, the students in this study were posttested using the achievement test. As indicated, this posttest had a total of 19 questions, and results in a total score of 22 points -- see Appendix C. The 19 questions included a combination of 15 multiple choice and 4 open-ended formats that were all scored correct or incorrect. Question 12 required four responses, which increased the possible total score to 22 points. Altogether, the total score was comprised of 11 conceptual responses, 4 procedural responses, and 7 problem-solving responses resulting in 3 scale scores. Impacts on the conceptual, procedural, and problem-solving scale scores due to the three treatments were compared to partially judge the appropriateness and adequacy of the ATOSU scoring model.

3.5.3 Audiotapes of Lessons: Four students from the ATOSU group were audiotaped during their guided assessment session; four additional students from the test-only group were also audiotaped at a latter date. The taped discourse and written comments record, index, and clarify emergent instances observed by the teacher and the students. These were unplanned and distinct to the task or student. The recorded discourse adds another dimension to the analysis when examining how students solve problems during a guided-assessment session. One advantage of using audio-recordings is their potential for reviewing lessons after they are taught. Also, with audio transcriptions, other researchers can independently evaluate the assessment procedures. Qualitative information was used as a trial and check strategy of the ATOSU scales. Discourse noted in the audio transcriptions helps triangulate item-response patterns on ATOSU variables. Thus, limitations in one method can be compensated by the strengths of a complementary one.
3.6 Experimental Procedures

1. The researcher tutored 7th graders in the Tutored-only groups. He also tutored and assessed 7th graders in the ATOSU group using the ATOSU protocol in Appendix A. Topics, activities and classroom procedures primarily focus on understanding probability and statistical concepts. In the Tutored-only group, students were simply shown how to do the problems. In the ATOSU group, instead of telling students how to do the problem, the overall goal was more to challenge students during the lessons and guide them through the experience.

2. Assessment feedback sessions on probability and statistics were held individually between the researcher and students in the ATOSU group. Ongoing feedback was provided in verbal form as the feedback session unfolded. The researcher then evaluated the individual subject's performance in the treatment group. Open-ended questions were scored by the researcher using the ATOSU assessment during this individualized feedback session. The Title I teacher observed four students receiving ATOSU feedback and provided judgmental ratings along with the researcher. Audiotapes of these guided assessment sessions recorded four separate lessons using two male and two female students.

3. The researcher recorded his observations and insights during each student's feedback session. Written observations were made and a score assigned for each ATOSU scale on each task. These observations were used to provide further direction to the student.

4. The posttest was administered as per Table 3.3. The posttest was given to the third group — the Test-only group — one year latter. The Test-only group consisted of a new population of Title 1 students in the 7th grade and from the same school. After the test, a third teacher judged the performances of four additional students. Audiotapes of these guided assessment sessions record four separate lessons using two males and two females.

3.7 The Analyses

The proposed analyses occur in three phases: 1) measurement analysis, 2) experimental analysis, and a 3) qualitative analysis. These three phases are now discussed in detail.
3.7.1 Measurement Analysis: According to Wright and Masters, measurement begins with the idea of a variable represented on a continuum along which objects can be positioned and ordered (Wright & Masters, 1982). Before one measures, one must indicate events that are believed to be ordered in some preconceived fashion. These events are interpreted as steps in an "intended direction," where a score of three means the learner has passed through steps one and two. A measurement design is employed to check whether anticipated orderings match empirical orderings when judges employ the [A], [T], [O], and [S] scales to independently rate the students' performances (Wright & Masters, 1982, p. 3). Although not part of the problem-solving process, the [U] condition may also be rated to determine the judges' conceptions of the difficulty of the task conditions. The purpose of this measurement analysis is to make preliminary determinations of the validity and reliability of the ATOSU scales and the post-test.

The ATOSU constructs were considered to be measurable by experts who reviewed the ATOSU instrument and determined it to be content valid. To determine content validity, the expert compared the scoring model with what is commonly maintained and described by learning and developmental theory. There may be some judgmental differences; however, how much this impairs the usefulness of the ATOSU protocol is a matter for analysis. Identification of a domain, implementing it carefully, and obtaining agreement on it are critical in content validation. Tasks used in teaching units were selected as representative of probability and statistics problems commonly found in this content domain.

The ATOSU constructs were also theorized as hierarchical scales. One way of examining a set of measures for a hierarchy is to inspect the correlation matrix presenting
pair-wise relationships between the constructs. The adequacy of order was judged by how each empirical result relates to one anticipated by many-facets Rasch theory (see Linacre, 1989). Since these [A], [T], [O], [S] and [U] constructs predict a preconceived order of difficulty, the constructs should correlate in a manner consistent with a hierarchy. If the [A], [T], [O], [S], and [U] constructs follow a monotonic order, the constructs correlation matrix will exhibit higher correlations between attributes closer in proximity within the matrix, and lower correlations between attributes that are farther apart in the matrix. This pattern of correlations is often called a simplex matrix.

A second way of inspecting the order of the constructs was to employ Rasch-model based Facets analysis. Facets analysis employs a measurement design at the item or task level to calibrate variables and make scales (Linacre, 1989). The Facets analysis advantage is due to its ability to handle certain forms of unbalanced data matrices (Linacre, 1989). Ordinarily, a proper judging plan must include judges-items-students combinations that share similar linkages. However, in the Facets design, judges need not rate all students or all tasks, as when the judging plan is properly as in Table 3.2. Since every level of every facet can be linked directly and unambiguously with every other element, a Facets analysis is possible.

In this unbalanced situation, there are four factors that one can be concerned about: the ability of the students taking the tasks (Bn), the difficulty of the tasks or problems upon which they are judged (Di), the severity of the judges (Cj), and ATOSU constructs rated
(Fk). The data can then be treated with a four-facet model, using equation 3.1 below (Linacre, 1991):

$$\log \left( \frac{P_{njk}}{1 - P_{njk}} \right) = B_n - D_i - C_j - F_k \quad (3.1)$$

where $P_{njk}$ is the probability of observing student (n) over task (i) for judge (j) with scale (k), divided by the probability of all other possible Facets rankings.

The ATOSU model tracks a series of observable problem-solving activities assumed to be generalizable over some preconceived order or hierarchy. Facets analysis (see Wright & Masters, 1982) can be used to verify the model's hierarchy and to determine the reliability of separation among the raters, items, students, and construct scales. A major assumption in Facets analysis is that the data fit the Rasch model. This is tested by using a set of fit tests. When the local independence and unidimensional assumptions hold, the observed results may be compared to the theoretical or expected results. The unidimensionality assumption is validated when the fit and order of the constructs reproduces themselves again and again across judges, students, and tasks.

The estimate of the judge reliability using Facets analysis was compared to Guilford's classical approach for determining inter-rater reliability. The Guilford approach demands that one first take the ATOS ratings and order them from high to low, and then the following formula may be applied to those ranks (see Dick & Haggerty, 1971).
The Guilford formula

\[ \bar{r} = 1 - \frac{k(4N+2)}{(k-1)(N-1)} + \frac{12S^2}{k(k-1)N(N^2-1)} \]

where

\[ \bar{r} = \text{average intercorrelations among individual judges and, importantly, is therefore the reliability of one rater} \]

\[ k = \text{number of judges, where } k > 1 \]

\[ N = \text{number of individuals, where } N > 1 \]

\[ S = \text{sum of ranks of ratings for any individual} \]

The step-up formula (Spearman-Brown Prophecy formula) was used to determine the inter-rater reliability for \( k \) raters in each of the 3 groups.

The step-up formula (Spearman-Brown Prophecy formula)

\[ r = \frac{k\bar{r}}{1 + (k-1)\bar{r}} \]

where

\[ r = \text{inter-rater reliability} \]

\[ k = \text{number of raters} \]

\[ \bar{r} = \text{the average reliability coefficient for one rater} \]

By comparing the Rasch approach to the classical approach, one notes the robustness of the results using different methods. When compared to classical approaches, Rasch methods generally yield higher reliability coefficients, making Rasch methods more robust.
Classical-test-theory-based Cronbach’s alpha was calculated for each of the three dependent variable scales representing conceptual, procedural, and problem-solving ability. When there is high reliability of the dependent measures, there is increased power for the statistical tests for the experimental analysis (see Appendix D).

3.7.2 Experimental Analysis

The purpose of the experimental analysis was to test a series of three hypotheses enumerated under the second research question earlier in this chapter. A 3 x 2 factorial design -- three levels of assessment (assessment plus tutoring, tutoring, and no tutoring) and two levels of gender (males and females) -- was used to test the three hypotheses.

In the preliminary stage of the experimental analysis, one should test the assumptions upon which the Factorial MANOVA model is based. There are three assumptions regarding the error term that must be checked before proceeding with the analysis. The “multivariate normal” assumption, the “homogeneity of covariance” assumption, and the “independence” assumption of the MANOVA are each examined as preliminary stages in the analysis. These three assumptions must hold for the error effects if one is to have confidence in the results.

If these assumptions hold, Wilk’s Lambda was calculated for the treatment effect, the gender effect, and the gender/treatment interaction effect. Rao F is an approximate statistic derived from Wilk’s Lambda and is calculated to test the two main effects and the interaction effect viewed within the model. Any significant effect produces an obtained value that is greater than the critical value and associated with a predetermined alpha level.
(e.g., alpha = .05). All three multivariate tests may be followed by a series of univariate tests that examine the main effects and the interaction effect on each dependent variable, one variable at a time. If an approximate multivariate F is significant for any main effect, post hoc comparisons are done to explore this result, using Tukey’s Test of orthogonal pairs, across the dependent variables. However, a significant interaction effect makes the interpretation of the treatment or gender effects impossible using Tukey’s Test, but graphical methods may be employed to examine the effects of the treatments in combination with the effects of gender.

The first main effect has three levels representing each group in the experiment – namely, test-only, tutored-only, and guided-assessment with tutoring. If one disregards the effects of gender, Hypothesis #1 compares the treatment factor effects over linear combinations of the conceptual, procedural, and problem-solving scales. Even if there were no real differences among the three comparison groups, one would expect a slight difference in the sample means due to chance alone, so one would test the null hypothesis that the three linear combinations are random samples from populations with identical conceptual, procedural, and problem-solving mean vectors. A main effect that favors the performance of the Tutored plus ATOSU assessment group on the dependent variables leads to the rejection of the null hypothesis.

The second main effect of interest is gender, which obviously consists of two levels: males and females. Hypothesis # 2 concerns the existence of a gender main effect regardless of the treatment, and postulates that no gender differences will exist. No gender differences would imply that boys and girls perform about the same over linear combinations of the
dependent scales, regardless of what comparison group they were assigned. For a sufficient size sample, test results that fail to reject the null (i.e., no gender differences with respect to the treatment) would make the results easier to interpret and accept. But when the sample is small, such a result may prove to be inconclusive. A gender effect suggests disparities in achievement between boys and girls on the dependent variable, regardless of the treatment. A significant gender main effect is therefore potentially problematic, but there are alternative explanations that do not suggest inequities attributed to any measure. For example, if either the male or female members have been socialized in ways that encouraged gender differences in mathematics learning or that influenced the development of prerequisite skills necessary for thinking mathematically, a test using the NAEP items could highlight these differences. Three sessions of instruction using most treatments would not obliterate the effects of a lifetime of differential socialization.

Hypothesis #3 posits that there is no interaction effect, meaning that the conceptual, procedural, and problem-solving performance of the males and females are consistent for each respective group tested. In the case of an interaction, a problem arises when levels of gender performance vary over levels of treatment performance. This scenario is strongly suggested when a disordinal interaction exists, since treatment differences are directly related to the gender of the subjects. For example, females may do better than males when given assessment feedback using the ATOSU assessment (the females may get better ratings because they are better behaved) while the males do better than the females under tutoring conditions without the assessment feedback. Although not as dramatic, these effects are also present for an ordinal interaction, with one level of gender gaining or declining less
dramatically when compared to the other level of gender as a result of the treatment. Given either type of interaction, one is required to examine the interaction effects using graphical plots of the cell means. Interaction can be defined as a significant departure from the parallel relationship of the three lines. In the overall analysis, one prefers to fail to reject the null when testing interaction effects, since no interaction maintains clarity of interpretation and meaningfulness of the results.

3.7.3 Qualitative Analysis

Qualitative data from the observation design are analyzed by using the ATOSU framework as a classification scheme for student reponses, evaluative comments, and the instructional discourse transcribed from the audio recordings by the researcher. The same cases used in the measurement phase were audiotaped during instruction and assessment; and these tapes, along with each student's respective profile, were used in this analysis. To focus this qualitative analysis of the ATOSU framework, the 8 student cases -- 4 males and 4 females from both the ATOSU and Test-only groups — were asked to attempt four tasks and explain how he/she solved each task. These typical cases were selected based upon the richness of the information each case may yield as recommended by the key informers (in this case, teachers). A case consists of all the raw data and information collected about a particular student for which this analysis is conducted. The researcher assembled 8 case records containing instances of events or activity as described by Patton (1990). Instances are "condensations or snippets" of the raw-case data that may be edited, classified, and
organized into aggregated records or manageable packages (see Erickson, 1986) using coding techniques. Instances may vary in length and they may overlap scenarios within the data. Efforts are made to link judged ratings of student activity using the ATOSU model to these qualitative instances. Patterns of common and uncommon instances are described, interpreted, and verified in an effort to explain their substantive significance.

3.7.3.1 Description: As in quantitative analysis, the first goal of the observation design is description, and efforts are made to describe the aggregated cases or records. ATOSU open-ended responses, written observations, and audio recordings provide a rich data-base for constructing such records. Traces of evidence are found in the observations recorded by the researcher on the ATOSU instrument and in transcripts of the audio recordings. These student activities or event patterns are thematic instances representing the basic units of qualitative analysis that can be coded by student, task, judge, and ATOSU category. Instances or exemplars of students' responses may also be coded and used as thematic instances in an effort to draw comparisons and detail points. For example, any qualitative discourse could also be classified by category and compared between cases or between gender. Qualitative snippets detail subtle and unique activity or events that make a difference between possible judgments made with the ATOSU scales.

General description is presented in a format that is "synoptic" — that is to say, the data are organized in ways that afford a general view of analogous instances of "broader themes" (Erickson, 1986). The 8 students judged by the teachers and the researcher will be
included in a qualitative analysis of their audio taped assessment sessions. From the converging patterns of data found in these 8 cases, a set of assertions may be generated largely through induction and guided by the research questions as described by Erickson (1986). These assertions are organized in a systematic attempt to confirm or dismiss their relevance in understanding the significance of these observed activities, events, or instances, and to search for theoretical linkages between these instances and the scores that lead to further insights.

3.7.3.2 Interpretation: The second goal of the qualitative analysis involved the interpretation of any deduced assertions. All 8 cases summarize these thick descriptions as a holistic portrayal of each student as their activity unfolds using inductive analysis. Inductive analysis "means that the patterns, themes, and categories of analysis come from the data" (Patton, 1990, p. 390). The discipline and rigor of qualitative analysis depends on presenting solid descriptive data, what is called "thick description" (Denzin, 1989; Geertz, 1973), in such a way that others reading the results can understand and draw their own interpretations. Interpretation involves explaining the findings by putting patterns viewed within the qualitative data into the ATOSU framework and attaching significance to particular results. Interpretive commentary that precedes and follows description is necessary to frame how any analytic instance is a concrete example of the overall assertion that is being advanced. Interpretive commentary fills in information and clarifies the details regarding the qualitative exemplars that are presented. This permits the researcher to be an "analyst" as well as a "reporter" (Erickson, 1986), to demonstrate patterns of instances
aimed at answering the questions, and commentary aimed at explaining the retrospective interpretations made by the investigator. Published research was used to bring focus to these interpretations of this learning activity.

Rating students on hierarchical scales implies that there are important differences in the quality of their performances and their understanding of the lessons. However, measurement scales are static and may be too gross or too specific in capturing the meaning of these differences in quality. In qualitative analysis, the analyst looks for natural variation in the data as it occurs during problem-solving activity, with particular attention to how the 8 cases differ in knowledge, instruction, and assessment. These particular instances may be flexibly construed over observable patterns of behavior and stated as assertions. The students’ thinking concepts and processes are bound to be affected by treatment and gender variations, and the researcher’s goal is to interpret these variations in purposive activity across these different tasks. When applying the ATOSU categories, variations in thought may be very different from case to case, so qualitative analysis would attempt to establish non-overlapping instances between categories and to homogenize instances within behavioral categories. Even though groups of students may have the same ratings, there may be great variation in their individual case records. To understand the individual learner within the context of this variation, the observer would need to focus on nuance, on detail, and on the changing perspectives of the learner (see Patton, 1990).

Regular patterns of instances are classified and coded, even though some instances may potentially overlap in their characterization. Similar instances are aggregated within coded categories to signify their convergence; different instances are assigned to a variety
of separate categories to indicate their discriminatory properties. These patterns may then be related to both the ATOSU scales and their anchored behaviors and the item performance viewed on the dependent results. The ultimate goal is to explain these discriminating and converging formulations in a defensible manner that may be verified, using evidence from several sources judged by teachers. Interpretations offered in this fashion apply the multi-trait/multi-method approach, but these relationships do not rely exclusively on a quantitative correlation matrix as originally proposed by Campbell and Fiske (1959). However, this approach seems more suited to the purpose of any guided assessment — that assessment should be a means of fostering growth to higher expectations.

Likewise, convergent and divergent instances both within a case and among cases can then be reexamined for irregular or incongruent instances not explained by the ATOSU measures. During any instruction, patterns of activity and response are simply what the student mostly does. Priorities in teaching need to be determined as a result of this activity — what the student usually does at a given level of complexity. It is more helpful to talk about common themes of difficulty as opposed to the occasional lapse or an impulsive inference. However, it is expected that many of these uncommon instances will have common themes that may be interpreted. The researcher will also attempt to describe how such themes relate to guided performance and NAEP performance. Any discoveries may be used to improve the model and help provide future directions for research.
3.8 Summary

This chapter described a methodological plan for testing the research hypotheses proposed in Chapter 1. Data are gleaned in ways that maximize the evidence such that the researcher gains a clearer sense of the ATOSU model's impact on problem-solving. Overall, determining the quality of any assessment practice requires that one understands and evaluates the impacts of such assessments on students. A methodological mix of qualitative and quantitative approaches was used to test hypothetico-deductive and holistic-inductive interpretations of the data. Different methods were required to assess different types of targets. Care was taken to meet the independence assumption by treating students individually rather than in groups. At the end of the assessment feedback, it is hypothesized that students receiving such guided feedback perform better those taught in more traditional ways.
Chapter 4
Analysis of the Results

This chapter presents the quantitative and qualitative results of this study in three phases: the first phase describes the results of the measurement analyses of the ATOSU and achievement test scales; the second phase discusses the descriptive and inferential analysis of the experimental results; and the last phase illustrates the ATOSU criteria using qualitative data, and then applies the model by presenting 8 cases describing students receiving guided assessment. The purpose of these analyses is to answer questions and related hypotheses presented in Chapter 3. Assumptions of the methods are tested or examined where appropriate, and interpretations are made. Briefly, the research questions are as follows:

4.1 The Research Questions

4.1.1 Are the variables utilized in this study measured validly and reliably?

1. Are the five ATOSU attributes measurable? Are the tasks content valid and reliable?
2. Is the hierarchical nature of the five ATOSU scales validated in this study? Is this hierarchical structure reliable?
3. Can the ATOSU model be implemented reliably by the judges?
4. What is the reliability of the conceptual, procedural, and problem-solving scales used as the dependent variables?

4.1.2 Are there treatment and gender differences in the experiment? Are there interaction effects related to ATOSU treatment and gender?

Hypothesis 1: Students who are given guided assessment using the ATOSU model will perform better on the conceptual, procedural, and problem-solving scales of the achievement test than students who do not receive guided assessment.
Hypothesis 2: In general, there will be no gender differences in performance on the conceptual, procedural, and problem-solving scales regardless of the treatment.

Hypothesis 3: There are no differences in conceptual, procedural, and problem-solving performance as measured by the treatment-by-gender interaction effects.

4.1.3 What are some major themes illustrated by the qualitative data relative to the ATOSU criteria? Do the qualitative data support the use of ATOSU protocol?

4.2 Measurement Phase

In this study increases in mathematics problem-solving performance were to be affected by using a form of guided assessment called the [A]ttention, [T]actic, [O]peration, [S]olution, and [U]nique conditions (ATOSU) model. The first set of research questions addressed the validity and reliability of the various measurement scales developed for this study.

4.2.1 Measurability of the Model: The ATOSU model was determined measurable through a process of content validation (Cronbach, 1984). Scrutiny of the content is especially important in a test that certifies competence or that evaluates cognitive progress in problem-solving. The ATOSU model, as a form of guided assessment, attempts to develop competence in problem-solving. The model, then, ought to assess those characteristics, not something else. For a scoring procedure to be content valid, the model developer needs to define the domain or range of attributes and tasks appropriately, and the kinds of behaviors the teacher or judge is expected to count. This is delimited by a scoring protocol that allows any trained user to decide whether the scoring procedure aims at the proposed targets. The protocol that defines the scoring procedure is available for review in Appendix A. Members
of the investigator's doctoral committee were asked to review the quality of the conceptual and procedural measures, and to judge the content validity of the tasks relative to the probability and statistics skills scored by the protocol.

[A], [T], [O], and [S] attributes are assumed to measure increasing levels of conceptualization skills under various [U] conditions. For students to recognize more sophisticated conceptual patterns, they need to develop the ability to learn and generalize verbal, visual, and mathematical patterns taught in their lessons. These conceptual generalizations were noted when the judge rated superior forms of problem-solving when the student responded in more sophisticated ways, and rewarding this performance with a higher number for a particular attribute. For example, when scoring a response, a score of 3 on the [A]ttention attribute was assumed to be a more generalized form of conceptualization than a score of 2. This seems appropriate when a score of 3 corresponds to a specific level of attending behavior that is considered more complex than a specific level of attending behavior associated with a score of 2, especially when the student is confronted with unrelated situations. Scores on a marginal attribute conditioned on judges constitutes a univariate frequency distribution of student scores for one judge on all the tasks. Because there are five such different univariate distributions -- one for each of the ATOSU marginal attributes -- each attribute has a corresponding scale that delimits specific levels of increasingly more sophisticated conceptual thinking.

In learning to generalize procedures, this protocol assumed that students first mimic the actions of teachers, while learning to control their own learning activity. Initially, teachers will cue students when their thinking strays but, through varied and guided practice, students learn to function independently. This protocol further assumed that procedural
thinking is combined and recombined during varied practice and is necessarily connected to conceptual thinking. A procedural hierarchy is employed by grouping the marginal [A], [T], [O], and [S] attributes and by considering the relative ordering of these scores subject to changing [U] problem conditions. When given a word problem, the ordered structure of the scale predicted that students first use verbal communication [A], then visual communication [T], next mathematical communication [O], that finally leads to mastery during mathematical problem-solving [S] under varying [U] conditions.

So considering the content validity of the protocol, the researcher’s committee members agreed that the ATOSU measures function separately as 5 scales representing generalized concepts. The higher the rating on a given scale, the more a concept is said to have generalized in thinking. The committee members further understood that, when considered together, these 5 scores were predicted to follow an order where the probability of a high rating on the [A] attribute is greater than the [T] attribute, the likelihood of a high rating on the [T] attribute is greater than the [O] attribute, and so on. If this ordering of scores is reproducible, it invited a sequential interpretation of the attributes. This generalized sequence is said to model procedural thinking.

The content validity of the tasks was next analyzed, but first Facets results are presented for the task facet to note whether their ordering makes sense. The pattern of task calibrations provide a description of the reach and ordered difficulty of the task. The Facet’s analysis report (see Linacre, 1989) related to tasks presented in Table 4.1 summarizes the difficulty of the four tasks reflecting various levels of complexity in problem-solving in probability and statistics. The task difficulty values (calibrated in logit units) suggested no overly difficult or easy tasks, with all four tasks being approximately within one standard
<table>
<thead>
<tr>
<th>Obsvd Score</th>
<th>Obsvd Count</th>
<th>Obsvd Average</th>
<th>Fair Average</th>
<th>Calib Logit Error</th>
<th>Infit MnSq Std</th>
<th>Outfit MnSq Std</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>145</td>
<td>80</td>
<td>1.8</td>
<td>1.9</td>
<td>-0.21 0.15</td>
<td>1.4</td>
<td>2</td>
<td>1.4   2</td>
</tr>
<tr>
<td>176</td>
<td>80</td>
<td>2.2</td>
<td>2.4</td>
<td>-0.91 0.15</td>
<td>0.9</td>
<td>0</td>
<td>0.9   0</td>
</tr>
<tr>
<td>126</td>
<td>80</td>
<td>1.6</td>
<td>1.6</td>
<td>0.26 0.16</td>
<td>1.0</td>
<td>0</td>
<td>1.0   0</td>
</tr>
<tr>
<td>104</td>
<td>80</td>
<td>1.3</td>
<td>1.3</td>
<td>0.86 0.17</td>
<td>0.6</td>
<td>-2</td>
<td>0.6   -2</td>
</tr>
<tr>
<td>137.8</td>
<td>80.0</td>
<td>1.7</td>
<td>1.8</td>
<td>0.00 0.16</td>
<td>1.0</td>
<td>-0.4</td>
<td>1.0   -0.3</td>
</tr>
<tr>
<td>26.4</td>
<td>0.0</td>
<td>0.3</td>
<td>0.4</td>
<td>0.65 0.01</td>
<td>0.3</td>
<td>1.9</td>
<td>0.3   1.8</td>
</tr>
</tbody>
</table>

Root mean square error (RMSE) 0.16  Adjusted Std 0.63  Separation 3.95  Reliability 0.94
Fixed (all same) chi-square: 66.45  d.f.: 3  significance: 0.00
Random (normal) chi-square: 3.00  d.f.: 2  significance: 0.22

Symbols/ Abbreviations Used:
Obsd = Observed  Avrge = Average  Calib = Calibration  MnSq = Mean Square  Std = Standard deviation

Table 4.1: Task Facet

deviation of the mean difficulty. Task difficulties were evenly distributed over this range, with each of the four tasks separated and ordered at approximately equal logit distances on the line representing the latent variable. The reliability of the separation of tasks is 0.94 and the RMSE (root mean square error) is 0.16. Tasks separated themselves in an appropriate way given the sample of 8 students, so no task was considered too difficult or too easy. However, students had more success with tasks administered earlier in the study than later in the study. This poorer performance suggested fatigue or a loss of interest or motivation. This decline of scores in time was especially apparent in the case of the boys. Yet, despite this trend, these tasks permitted a useful means for following the changes in task difficulty and ordering them on a line. Once ordered along the line, the students’ positions were interpreted in terms of the kinds of problems the students were expected to get correct or incorrect based on their estimated ability level.
All four tasks were considered representative of basic probability and statistical skills as judged by the committee and the teachers. Teacher’s ratings on the [U] scale suggested that tasks 1, 3, and 4 demand thinking under relatively more unfamiliar and discriminating ways. One teacher-judge rated some of these tasks more complex than the other judges; however, the judge agreed that task 2 was decisively more difficult than any of the other tasks.

Although not affirming any assumptions made by the ATOSU protocol without additional evidence, the committee agreed that the protocol and administered tasks attempt to advance and measure some fundamental forms of verbal, visual, and mathematical communication related to probability and statistics. Teachers who judged this problem-solving activity during the measurement phase of the study generally agreed with these views.

4.2.2 Hierarchy of the ATOSU Model: Having gained the approval of experts and teachers regarding the content validity of the tasks and the ATOSU instrument, the next step ascertained whether a hierarchy exists among the [A], [T], [O], [S] and [U] attributes. Levels of attributes differentiating a hierarchy may not appear to be universally appropriate over all persons and tasks. Error or noise may be observed when unpredicted patterns do not reproduce the assumed order. For this study, a representative pattern was hypothesized to reproduce itself, such that the attribute scores maintain the following predicted hierarchy [A] >/ [T] >/ [O] >/ [S] >/ [U] as observed across a significant majority of tasks. A correlation matrix was generated and examined to note whether there was a monotonic (descending or
ascending) pattern of correlations among the attributes. An overall pattern presented itself within any vector of the correlation matrix that may be summarized with the following rule:

The greater the distance between any two attributes within the data matrix, the lower the correlation; the smaller the distance between two attributes within the data matrix, the higher the correlation. Such patterns would be indicative of a hierarchical order between the [A], [T], [O], [S] and [U] attributes, and this type of hierarchical pattern is called a simplex matrix. For example, if one compares the correlated pairs of [A]ttention with other attributes in Table 4.2 above, one notes a declining pattern such that the correlation between [A]ttention and [T]actic (r=0.771) is higher than the correlations between [A]ttention and [O]peration (r = 0.670), or [A]ttention and [S]olution (r = 0.645), or [A]ttention and [U]nique Conditions (r= -0.022); If one compares the correlated pairs of [T]actic with other attributes, the correlations between [T]actic and [A]ttention (r=0.771) and between
T]actic and [O]peration (r = 0.822) are higher than the correlations between [T]actics and [S]olution (r = 0.749) or [T]actic and [U]nique Conditions (r = 0.135); and so on for most of the [O], [S], and [U] correlations. The only exception to this pattern is found when one compares the correlated pairs of [S]olution with other attributes. Even though the [U] attribute is close in proximity to the [S] attribute within the data matrix, the correlation of the [S]olution attribute with the [U]nique conditions attribute (r=0.287) is lower than correlations between [S]olution and [T]actic (r=0.749) or [S]olution and [A]ttention (r=0.645).

However, one limitation presented itself in terms of this simplex explanation. When the correlations were calculated, the attributes were pooled across judges and tasks. This aggregating most likely introduced some dependency between the attributes that undoubtedly influenced these correlations. Despite these limitations, the correlation matrix illustrated in Table 4.2 mostly represents a simplex pattern with correlations of a chosen attribute with all other attributes decreasing as the distance between vectors in the correlation matrix increases.

The structure of the calibrated rating scales conforms to the unique order of the [A], [T], [O], and [S] attributes, but not the [U] calibration, as examined using the Many-Facets Rasch model. The order or direction established by the ATOSU scales is a function of how the raters employ it, thereby producing attribute separation as a reflection of hierarchically-ordered levels of procedural quality. The Root Mean Square Error (RMSE) calculated differences between the actual and estimated Item Characteristic Curves (ICC's) by
<table>
<thead>
<tr>
<th>Obsvd</th>
<th>Obsvd</th>
<th>Obsvd</th>
<th>Fair</th>
<th>Calib Logit</th>
<th>Model Error</th>
<th>Infit MnSq Std</th>
<th>Outfit MnSq Std</th>
<th>Scale attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Score</td>
<td>Count</td>
<td>Average</td>
<td>Avege</td>
<td>-1.16</td>
<td>0.16</td>
<td>0.8</td>
<td>-1</td>
<td>0.8</td>
</tr>
<tr>
<td>150</td>
<td>64</td>
<td>2.3</td>
<td>2.5</td>
<td>-0.13</td>
<td>0.17</td>
<td>1.0</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td>113</td>
<td>64</td>
<td>1.8</td>
<td>1.8</td>
<td>0.34</td>
<td>0.18</td>
<td>0.9</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td>98</td>
<td>64</td>
<td>1.5</td>
<td>1.6</td>
<td>1.04</td>
<td>0.20</td>
<td>1.2</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>78</td>
<td>64</td>
<td>1.2</td>
<td>1.3</td>
<td>-0.10</td>
<td>0.17</td>
<td>1.0</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>112</td>
<td>64</td>
<td>1.8</td>
<td>1.8</td>
<td>0.00</td>
<td>0.18</td>
<td>1.0</td>
<td>-0.1</td>
<td>1.0</td>
</tr>
<tr>
<td>110.2</td>
<td>64.0</td>
<td>1.7</td>
<td>1.8</td>
<td>0.72</td>
<td>0.01</td>
<td>0.2</td>
<td>0.2</td>
<td>0.9</td>
</tr>
<tr>
<td>23.6</td>
<td>0.0</td>
<td>0.4</td>
<td>0.4</td>
<td></td>
<td></td>
<td>0.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root means square error (RMSE) 0.18  Adjusted Std. 0.69  Separation 3.91  Reliability 0.94
Fixed (all same) chi-square: 81.47  d.f.: 4  Significance: 0.00
Random (normal) chi-square: 4.00  d.f.: 3  Significance: 0.26

Symbols/Abbreviations Used:
Obsd = Observed  Avrge = Average  Calib = Calibration  MnSq = Mean Square  Std = Standard deviation

Table 4.3: Attribute Facet

computing the squared distances between the curves at a number of points. In Facets Analysis, only non-extreme scores are used in calculating the RMSE. By dividing the adjusted standard deviation (standard deviation of the non-extreme calibrations) by the RMSE, a separation index was calculated indicating the degree of spread or discrimination between the scale values. Scale calibrations in Table 4.3 demonstrated good separation with the [A], [T], [O], and [S] scales separating in their monotonic order. However, the predicted calibration order was not demonstrated for the [U] attribute. When the procedural order is reproducible as indicated by the scale calibrations, their procedural states of development are scalable and comparable across judges, persons, and tasks. Despite the [U] calibration not following the predicted order, the low RMSE of 0.18 for the scaled attributes suggests that the predicted hierarchy is accurate and precise in relation to the Rasch model. The reliability of this hierarchy and its predictive fit to the Rasch model was also analyzed and

95
interpreted using Facets analysis. Reliability of the separation of scaled attributes was reported as 0.94 in Table 4.3.

4.2.3 Inter-rater Reliability: Having demonstrated some preliminary evidence of a hierarchy, the next question addresses the inter-rater reliability and judge accuracy as determined by both the Facets model and the Guilford method. Based on Rasch Theory, the Facets Analysis method reports the proportion of observed sample variance which is not due to measurement error as the reliability with which the ratings separate the judges (Linacre, 1989). Unlike classical methods for computing reliability, this estimate is not an average intercorrelation among individual judges, but is particular to the score the judge actually assigns. A Rasch method like Many-Facets is less sensitive to extreme ratings assigned by extremely lenient or severe judges because they are rendered “sample free” in the Facets Analysis. Accuracy of the facet dimensions, as determined by their fit to the Rasch Facets model, verifies their consistency with the idea of a dimension along which attributes have a unique order. To compare these Rasch-based methods with classical methods, Guilford’s method was employed to estimate inter-rater coefficients of judge reliability that are averaged over judges. Classical methods like Guilfords are influenced by sample variance so they are sample specific.

The judge matrix (see Table 3.2 in Chapter 3) proposed an analysis of an unbalanced design with judges rating eight different students on four tasks stipulating students as the unit of analysis. All tasks and students were rated by two of the three judges, meaning no one student is rated twice by the same judge on the same task. Estimation of the model with
<table>
<thead>
<tr>
<th>Obsvd Score</th>
<th>Obsvd Count</th>
<th>Obsvd Average</th>
<th>Fair Avrge</th>
<th>Calib Logit</th>
<th>Model Error</th>
<th>Infit MnSq Std</th>
<th>Outfit MnSq Std</th>
<th>Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>103</td>
<td>40</td>
<td>2.6</td>
<td>2.4</td>
<td>1.03</td>
<td>0.21</td>
<td>0.8</td>
<td>0.8</td>
<td>0</td>
</tr>
<tr>
<td>81</td>
<td>40</td>
<td>2.0</td>
<td>1.8</td>
<td>0.10</td>
<td>0.21</td>
<td>1.1</td>
<td>1.1</td>
<td>0</td>
</tr>
<tr>
<td>61</td>
<td>40</td>
<td>1.5</td>
<td>1.3</td>
<td>-0.84</td>
<td>0.23</td>
<td>0.8</td>
<td>-1</td>
<td>0.8</td>
</tr>
<tr>
<td>42</td>
<td>40</td>
<td>1.1</td>
<td>0.9</td>
<td>-1.97</td>
<td>0.26</td>
<td>1.8</td>
<td>2</td>
<td>1.7</td>
</tr>
<tr>
<td>49</td>
<td>40</td>
<td>1.2</td>
<td>1.3</td>
<td>-0.91</td>
<td>0.25</td>
<td>0.4</td>
<td>-2</td>
<td>0.4</td>
</tr>
<tr>
<td>88</td>
<td>40</td>
<td>2.2</td>
<td>2.4</td>
<td>0.98</td>
<td>0.20</td>
<td>1.2</td>
<td>1.2</td>
<td>0</td>
</tr>
<tr>
<td>59</td>
<td>40</td>
<td>1.5</td>
<td>1.6</td>
<td>-0.34</td>
<td>0.23</td>
<td>0.9</td>
<td>0</td>
<td>1.0</td>
</tr>
<tr>
<td>68</td>
<td>40</td>
<td>1.7</td>
<td>1.8</td>
<td>0.11</td>
<td>0.22</td>
<td>0.9</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td>68.9</td>
<td>40.0</td>
<td>1.7</td>
<td>1.7</td>
<td>-0.23</td>
<td>0.23</td>
<td>1.0</td>
<td>-0.2</td>
<td>1.0</td>
</tr>
<tr>
<td>19.2</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.94</td>
<td>0.02</td>
<td>0.4</td>
<td>1.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Root mean square error (RMSE) 0.23 Adjusted Std. 0.92 Separation 4.05 Reliability 0.94
Fixed (all same) chi-square: 133.61 d.f.: 7 Significance: 0.00
Random (normal) chi-square: 6.98 d.f.: 6 Significance: 0.32

Symbols/Abbreviations Used:
Obsd = Observed  Avrge = Average  Calib = Calibration  MnSq = Mean Square  Std = Standard deviation

Table 4.4: Students represented the object of measurement in the Facets analysis.

<table>
<thead>
<tr>
<th>Obsvd Score</th>
<th>Obsvd Count</th>
<th>Obsvd Average</th>
<th>Fair Avrge</th>
<th>Calib Logit</th>
<th>Model Error</th>
<th>Infit MnSq Std</th>
<th>Outfit MnSq Std</th>
<th>Judges</th>
</tr>
</thead>
<tbody>
<tr>
<td>259</td>
<td>160</td>
<td>1.6</td>
<td>1.6</td>
<td>0.27</td>
<td>0.11</td>
<td>0.7</td>
<td>-2</td>
<td>0.8</td>
</tr>
<tr>
<td>164</td>
<td>80</td>
<td>2.0</td>
<td>2.2</td>
<td>-0.72</td>
<td>0.15</td>
<td>1.3</td>
<td>1</td>
<td>1.3</td>
</tr>
<tr>
<td>128</td>
<td>80</td>
<td>1.6</td>
<td>1.5</td>
<td>0.45</td>
<td>0.16</td>
<td>1.1</td>
<td>0</td>
<td>1.1</td>
</tr>
<tr>
<td>183.7</td>
<td>106.7</td>
<td>1.8</td>
<td>1.8</td>
<td>0.00</td>
<td>0.14</td>
<td>1.1</td>
<td>-0.0</td>
<td>1.1</td>
</tr>
<tr>
<td>55.3</td>
<td>37.7</td>
<td>0.2</td>
<td>0.3</td>
<td>0.51</td>
<td>0.02</td>
<td>0.2</td>
<td>1.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Root mean square (RMSE) 0.14 Adjusted Std. 0.49 Separation 3.44 Reliability 0.92
Fixed (all same) chi-square: 35.21 d.f.: 2 significance: 0.00
Random (normal) chi-square: 1.99 d.f.: 1 significance: 0.16

Symbols/Abbreviations Used:
Obsd = Observed  Avrge = Average  Calib = Calibration  MnSq = Mean Square  Std = Standard deviation

Table 4.5: Judges Facet
maximum likelihood methods produced a linear rating scale that ranks the students' answers as to their quality. Figure 4.1 displays a facets plot that vertically ranks students by their problem-solving ability, tasks by their difficulty, judges by their severity, and scaled attributes by their order of complexity on graphical lines according their respective calibrations. Graphical results depicted the object of measurement (students) relative to the range of the three facets: tasks difficulty, judge severity, and attribute complexity. As seen here, students 1 and 6 had ability levels that were high enough to satisfy the severity levels of all 3 judges, all 4 tasks difficulty levels, and the highest ATOSU attribute complexity. Tables 4.4 and 4.5 present the Facets Analysis for 8 students and the 3 judges. The student calibrations demonstrate the most variation, with four logit calibrations of student elements exceeding zero and four less than zero. Table 4.4 also presents pertinent statistics describing reliability and predicted accuracy in relation to the Rasch model. The student separation reliability describes the stability of the linear order of students in terms of their order of performance. Employing this notion, the reliability of student separation was 0.94. The separation index was largest for Students as compared to the other facets, but this was caused mainly by the number of students and Student # 4's very low score. The student facet measurement also had the largest RMSE of 0.23.

Judge calibrations varied the least, although one judge was more lenient than the other two. The reliability of judge separation was 0.92 in Table 4.5, so the separation of judges into lenient and severe levels seems reasonably stable. The root mean square error (RMSE) of the judge facet was 0.14.
Figure 4.1: Facets graphical report
Tables 4.1, 4.3, 4.4, and 4.5 summarized the fit of the person, judge, task, and scale calibrations respectively. From the probabilistic nature of the model, one begins to expect certain values to emerge as judged ratings are made. A chi-square test of "fixed effects" hypothesized that the elements can be regarded as the same measure allowing for measurement error. For example, the test is whether judges are homogeneous in their severity and leniency. In Table 4.5, the null hypothesis was tested against alternative hypotheses that the judges are not homogenous, with $p < 0.05$, rejecting the null hypothesis that the judges shared levels of severity. In fact, all four tests of fixed effects in Tables 4.1, 4.3, and 4.4 also rejected the null that specific elements of students, tasks, or scale attributes were homogeneous. These tests indicate that the elements were heterogeneous in their student ability, task difficulty, and attribute complexity.

A second chi-square fit test functions in ways similar to the fit test in loglinear analysis. The "random effects" test hypotheses that the elements were sampled from a normal distribution. Essentially, one is testing the null hypothesis that the data fit a normal distribution. In this case, when the model fits, one fails to reject the null hypothesis. By examining the four random chi-square tests in Tables 4.1, 4.3, 4.4, and 4.5, it was concluded that the basic structure of the person ability, ATOSU scale, task complexity, fit the normal distribution.

Guilford's method of calculating inter-rater reliability represents a classical approach which first assigned a rank to the ratings, and then the Guilford formula was used on these ranks to determine the inter-rater agreement. In Table 4.6, Guilford's method of calculating inter-rater reliability was applied to each of the A, T, O, and S attributes for both pairs of judges on the four tasks. This was done twice for each task since different judge-pairs rate
<table>
<thead>
<tr>
<th>Treatment, Task</th>
<th>judge pairs</th>
<th>attributes</th>
<th>2 judges</th>
<th>4 judges</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATOSU task 1</td>
<td>judges 1 &amp; 2</td>
<td>attention tactics</td>
<td>.62</td>
<td>.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.88</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.72</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>solutions</td>
<td>.33</td>
<td>.50</td>
</tr>
<tr>
<td>ATOSU task 2</td>
<td>judges 1 &amp; 2</td>
<td>attention tactics</td>
<td>.52</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.82</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.88</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>solutions</td>
<td>.75</td>
<td>.86</td>
</tr>
<tr>
<td>ATOSU task 3</td>
<td>judges 1 &amp; 2</td>
<td>attention tactics</td>
<td>.52</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.52</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.18</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>solutions</td>
<td>.52</td>
<td>.68</td>
</tr>
<tr>
<td>ATOSU task 4</td>
<td>judges 1 &amp; 2</td>
<td>attention tactics</td>
<td>.52</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.18</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.09</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>solutions</td>
<td>.09</td>
<td>.17</td>
</tr>
<tr>
<td>Test-only task 1</td>
<td>judges 1 &amp; 3</td>
<td>attention tactics</td>
<td>.92</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.92</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.33</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>solutions</td>
<td>.33</td>
<td>.50</td>
</tr>
<tr>
<td>Test-only task 2</td>
<td>judges 1 &amp; 3</td>
<td>attention tactics</td>
<td>.52</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.92</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.62</td>
<td>.77</td>
</tr>
<tr>
<td></td>
<td></td>
<td>solutions</td>
<td>.18</td>
<td>.31</td>
</tr>
<tr>
<td>Test-only task 3</td>
<td>judges 1 &amp; 3</td>
<td>attention tactics</td>
<td>.52</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.52</td>
<td>.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.18</td>
<td>.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>solutions</td>
<td>.18</td>
<td>.31</td>
</tr>
<tr>
<td>Test-only task 4</td>
<td>judges 1 &amp; 3</td>
<td>attention tactics</td>
<td>.57</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.57</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>operations solutions</td>
<td>.09</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>solutions</td>
<td>.18</td>
<td>.31</td>
</tr>
</tbody>
</table>

Table 4.6: Guilford’s Inter-rater Reliability for 2 and 4 judges.
different students. Guilford’s method required that each judge rating of the individual attributes was reliability for one judge. The Spearman-Brown prophecy formula was then used to determine the inter-rater reliability if the number of judges was increased from two to four. Results indicate high to moderate reliability for [A]ttention and [T]actics attributes and low reliability for the majority of [O]perations and [S]olution attributes. Generally speaking, there was lower reliability for task attributes whose individual ranks were commonly repeated because all or most students received the same ratings. A lack of variation among ranks more frequently occurred on the [O]perations and [S]olutions attributes because this sample of students seldom earned higher ratings. Even though the raters often agreed on the rating, a lack of rank variation on the assigned ranks by both judges depresses the reliability coefficient because there was less information to distinguish between student performances. In task 2, as students began to get higher ratings on [O]perations and [S]olution attributes, reliability for these traits improves while the [A]ttention reliability declines since students were no longer scoring on the low end of this scale.

Since the task conditions remain constant from student to student attempting each task, the rankings of the [U]-condition ratings do not vary. Consequently, when utilizing the Guilford formula, a set of tied rankings for the U-condition produces a reliability coefficient of -1.00 for each task. Therefore, the U-condition was not included in the above analysis.

4.2.4 Reliability of the Dependent Measures: Reported reliability of the achievement test’s three scales enables one to answer the question: What is the reliability of the
dependent variable's Conceptual, Procedural, and Problem-solving scales? A possible raw score of 11 points may be earned on the Conceptual score. The Cronbach reliability coefficient represents the correlation between parallel items when the items measuring a specific characteristic correlate with each other. Individual items do not correlate, revealing that the scale was not internally consistent or that it does not measure the same characteristic. The heterogeneity of the items comprising the Conceptual scale accounted for the relatively low reliability of this variable (Cronbach's alpha = 0.325).

There were only 4 Procedural items administered on the posttest, but each item correlated moderately with the other items, positively influencing the reliability of the Procedural scale (Cronbach alpha = 0.522). However, even though this scale was moderately precise, the reliability of the Procedural scale still suffers due to an insufficient number of items. Of the 19 possible items in the posttest, there were four Procedural items which represent 21% of the total number of items administered. Typically, ETS reports that 30% of the total items were procedural items, so the percentage of items used by the Procedural measure of this study's posttest falls short of the ideal.

There were 7 Problem-solving items administered on the posttest, representing 37% of the total items. The individual items the average inter-item correlation was low resulting in low reliability on the Problem-solving scale (Cronbach alpha = 0.249). Parallel items were intended to be similar in content and nature and designed to measure the same trait. Even though the percentage of Problem-solving questions used on the posttest is representative of the percentage used on the NAEP, the items were not parallel.

In summary, these results suggested that the ATOSU model as viewed in the protocol demonstrates some promise in monitoring mathematical problem-solving. Evidence and
arguments were laid out about the cognitive processes required for conceptual and procedural success to help the reader make sense of the model’s scores. ATOSU hierarchy was demonstrated in two ways — using a correlation matrix to show a simplex pattern and examining Facet scale calibrations and the scale separation index. Reliability based on Facets’ attribute separation and its predicted accuracy seems more than adequate. After ranking scale scores, reliability of these ranked scores was higher for attributes with fewer tied ranks. This order of predicted separation of the attributes was consistent for all tasks and for all judges, whether they knew about it beforehand (as with the researcher) or not (as with the teachers). However, more teachers’ ratings would seem necessary to draw more definitive conclusions regarding this model and its attributes.

4.3 Experimental Phase

In the second phase of the analysis, three hypotheses were addressed by the experiment. A between-subjects design examined treatment and gender differences in conceptual, procedural, and problem-solving abilities as measured by the dependent variable. Descriptive statistics, tests of assumptions, and tests of three hypotheses of the treatment, gender, and interaction effects are now reviewed and analyzed.

4.3.1 Descriptive Statistics: Descriptive statistics of sample mean, standard deviation, skewness and kurtosis were first used to describe how typical values are concentrated and to test whether assumptions of the inferential analysis were met. Likewise, graphical presentations of the data are used to describe the shape and spread of the variables to
increase the clarity of the research findings. This approach of examining sample statistics has obvious import since it allowed for useful ways of describing a group of scores.

The mean score for the Conceptual scale in Table 4.7 varied from a high of 6.19 points to a low of 4.18 points correct out of 11 possible points. Standard deviations for each of the six cells in Table 4.7 were approximately equal, with the exception of the ATOSU females whose Concept score distribution were leptokurtotic and whose standard deviation is relatively small. Concept score means were generally higher for both the ATOSU and the Test-only groups as compared to the Tutored-only group. Males in the Test-only cell outperformed all other gender-by-treatment cells.

Examining Table 4.7, Test-only males and ATOSU females had moderately larger mean scores on the Procedural scale as compared to the others. Males generally outperformed females in Procedural thinking. Standard deviations for each of the six Procedural cells only varied by moderate amounts, seemingly leaving these small differences in variation to chance alone. However, possible floor and ceiling effects existed for the procedural scale since the standard deviations of each cell are large relative to their respective means. Two standard deviations above or below each cell mean often exceeded the range of possible raw score values.

Table 4.7 displays mean performances on the procedural scale of approximately 1.55 correct and a standard deviation slightly over 1 for females, revealing another possible floor effect for the females on the entire procedural scale. Standard deviations across problem-solving cells varied little relative to the size of the mean, except in the case of the Test-only distribution.

105
<table>
<thead>
<tr>
<th>ACHIEVEMENT SUBSCALES</th>
<th>GENDER</th>
<th>TEST ONLY</th>
<th>ATOSU</th>
<th>TUTORED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>STD</td>
<td>Mean</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.63</td>
<td>3.65</td>
<td>9.37</td>
</tr>
<tr>
<td>CONCEPTS</td>
<td>MALE</td>
<td>6.19</td>
<td>1.89</td>
<td>5.36</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>5.24</td>
<td>STD</td>
<td>SKEWNESS 0.77</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>2.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONCEPTS</td>
<td>FEMALE</td>
<td>5.18</td>
<td>1.89</td>
<td>5.45</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>4.96</td>
<td>STD</td>
<td>SKEWNESS 0.77</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>1.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROCEDURES</td>
<td>MALE</td>
<td>2.82</td>
<td>1.07</td>
<td>1.54</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>2.09</td>
<td>STD</td>
<td>SKEWNESS -0.15</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>1.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROCEDURES</td>
<td>FEMALE</td>
<td>1.18</td>
<td>0.751</td>
<td>2.27</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>1.55</td>
<td>STD</td>
<td>SKEWNESS -0.329</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>1.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROBLEM- SOLVING</td>
<td>MALE</td>
<td>2.82</td>
<td>1.08</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>2.30</td>
<td>STD</td>
<td>SKEWNESS 0.75</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>1.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROBLEM- SOLVING</td>
<td>FEMALE</td>
<td>1.36</td>
<td>1.12</td>
<td>1.91</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>1.76</td>
<td>STD</td>
<td>SKEWNESS 0.16</td>
</tr>
<tr>
<td></td>
<td>STD</td>
<td>1.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.7: Descriptive statistics used to describe the distributions of individual cells.
4.3.2 Testing the Assumptions of the MANOVA: Before employing statistical procedures of inference, assumptions on which the chosen analysis is based must be examined. For the chosen MANOVA procedures, three assumptions are viewed: the dependent variable must be normally distributed, the covariances of each pair of the dependent variables must be equal across groups, and the observations must be independent. Descriptive analysis of moments like skewness and kurtosis, along with graphical procedures, aid in this process (see Table 4.7 and Figure 4.2). Since the multivariate normal distribution is important to MANOVA, it is necessary to examine the assumption that the data come from a multivariate normal distribution. In practice, the joint (multivariate) distribution is likely determined by methods directed at examining the assumption of normality for each of the marginal (univariate) distributions. To examine the assumption of normality, graphs of the cell distributions were employed for each of the three dependent variables. A necessary condition of any multivariate normal distribution is that each of the variables comprising the general case must be univariate normal (Stevens, 1986). The conceptual, procedural, and problem-solving distributions must each be distributed normally for every group and gender.

Kurtosis and skewness of the distribution has an effect on the power of statistical tests assuming normality. Kurtosis estimates the height of the density of scores in the distribution; skewness estimates the degree of symmetry or asymmetry in the distribution. By using the kurtosis and skewness coefficients, one can test whether the distributions vary from normality. Smith described the tests and published a set of critical values for kurtosis and skewness (see Smith, 1986). D’Agostino & Tietjen (1971, 1973) developed these critical values using simulations. Two statistical tests at the 0.01 level for each cell, times
Figure 4.2: Histograms of the treatment by gender cell distributions
the 18 cells in Table 4.7, produced an overall alpha less than 0.36 (2 x 0.01 x 18), which was considered liberal. Therefore, four or five spurious tests were expected to occur by chance alone.

Kurtosis was tested first because of its ability to affect type II error. The obtained value is found in Table 4.7 (this value is obtained by adding 3 to kurtosis). Only 1 of the 18 obtained values exceeded the critical interval of 1.425 and 5.10 at alpha = 0.01 and n=11 students per cell. So in all but one case, one accepted the null that the sample distributions are from the normal distribution. Skewness coefficients were next tested where the critical value is 1.54 at alpha = 0.01 and n=11 (c.f. Stevens, 1986, p.215, Table 6.5). Examining Table 4.7, no value of skewness exceeds 1.54, so overall, one was reasonably confident that the univariate distributions making up the joint distribution were normal. Therefore, by appealing to these tests as well as the general robustness of the multivariate test, one was reasonably confident in assuming a multivariate normal distribution.

If the normality assumption is tenable, the assumption of equal covariance matrices must next be met. This means that the covariance matrix for the three dependent variables for Group 1 must be equal to the covariance matrices for Group 2 and Group 3. To test the null hypothesis that the groups come from populations with the same covariances, one uses the Box Test, which can be conducted in SPSS 8.0 statistical software utilizing the GLM procedure. This test uses the determinants of the within covariance matrices or the generalized variances. In Table 4.8, the Box test evaluates the null hypothesis that the observed covariance matrices of the dependent variable are equal across groups. If the observed probability is smaller than 0.05 for the three dependent variables, one can reject
the null hypothesis that all covariances are equal. In the study, since the observed probability level (0.879) was large, the dependent variables' covariances matrices were reasonably equal; therefore, the null hypothesis cannot be rejected and the assumption seemingly holds.

<table>
<thead>
<tr>
<th>Box's M</th>
<th>24.432</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>0.708</td>
</tr>
<tr>
<td>df1</td>
<td>30.0</td>
</tr>
<tr>
<td>df2</td>
<td>8136.0</td>
</tr>
<tr>
<td>Prob.</td>
<td>0.879</td>
</tr>
</tbody>
</table>

Table 4.8: Box Test of Equality of Covariance Matrix tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

Independence describes a condition where one observation is in no way influenced by another observation. Regrettfully, there is no statistical test to determine whether observations are independent of one another. However, by randomly assigning subjects to each experimental condition and treating the subjects separately, one increases the ability of the experimental design to produce independent observations. This assumption is more likely violated by the Test-only group, whose subjects were randomly selected, rather than the randomly assigned Tutored-only and ATOSU groups.

4.3.3 Testing the three hypotheses with MANOVA: The purpose of employing inferential statistics was to test the hypotheses about treatment, gender, and treatment-by-gender differences previously made. When the assumptions of this analysis hold, the null hypothesis contends that any observed difference in a sample or a measure is regarded as a chance differences resulting from sampling or measurement error alone. Three hypotheses
associated with the experiment are of interest in this study: Does the ATOSU group outperform tutored or test-only groups on the three dependent variable scales? Are there gender differences with respect to the three scores? And is there an interaction between the effects of treatment and gender? The statistical technique used to test these hypotheses is MANOVA.

MANOVA tests multivariate differences for the treatment effect, the gender effect, and the interaction effect on three dependent variables. Three treatment groups crossed with males and females were compared to determine whether they differ on average in a set of Concepts, Procedures, and Problem-solving scales. Total variation on three levels of treatment and two levels of gender is divisible into two components: the difference in deviations of raw scores from their group means (within variance) and their deviation of the group means from one another (between variance). The multivariate analysis of variance produces a wide variety of multivariate tests, whose numerators represent the variation between the groups being compared, and whose denominators represent the variation within the groups. Wilk’s Lambda is defined in terms of the determinants of the W and T covariance matrices and is called the “generalized variance” for each set of variables (Stevens, 1986). Wilk’s Lambda has an approximate F statistic called Rao’s F. Multivariate F’s were estimated to test overall differences in the two main effects and the interaction effect. The obtained F approximation can be compared to a critical value on an F table at some predetermined level (usually alpha = .05).

The Rao F’s associated with Wilk’s Lamda are now presented (treatment: \( F = 1.651, p < .0139 \); and gender \( F = 2.02, p < .121 \); treatment by gender: \( F = 2.268, p < .042 \)).
Since the interaction was significant, the analysis must focus its interpretations on it. An interaction effect means that the effect of each treatment has on the dependent variables was not the same for all levels of gender. Because of this treatment by gender interaction, treatment and gender main effects were not interpretable using MANOVA.

Interpretations of the gender-by-treatment interactions were made by graphing multiple levels of treatment and gender over combinations of gender, treatment, and the achievement test scales. Graphs of the interactions may be viewed in Figure 4.3. Examining the differences in the Concepts scale, the Test-only group clearly outperformed both the ATOSU and Tutored-only groups. This was particularly true for male subjects in the Test-only group. Females in the Test-only group do at least as well as any other group included in the original round of findings. Graphs of the treatment groups for the Problem-solving scale revealed a distinct advantage favoring the males in all groups. Males generally did better in Procedures in all treatment groups but the ATOSU group. The ATOSU females demonstrated significantly better Procedure skills than females in the Test-only or Tutored-only groups.

Interaction effects on the multivariate tests were largely attributed to the procedure scale. In Table 4.10, by examining the univariate tests of each dependent variable on the joint effects, significant effects were noted for only the Procedure scale and not the Concept, or Problem-solving scales. The ATOSU females outperformed the ATOSU males, but this advantage was reversed in the Test-only and Tutored-only conditions. The ATOSU’s males diminished performance in Procedure performance was puzzling and discouraging. Perhaps the ATOSU males are not benefitting from the assessment feedback or are not motivated by the extra points earned for demonstrating more advanced levels of problem-solving.
Figure 4.3: Graphs of the treatment and gender interactions for each dependent variable
<table>
<thead>
<tr>
<th>Effect</th>
<th>TYPE OF TEST</th>
<th>VALUE</th>
<th>F</th>
<th>Hypoth. df</th>
<th>Error df</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>PILLAI'S TRACE</td>
<td>.095</td>
<td>2.02</td>
<td>3</td>
<td>58</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>WILK'S LAMBDA</td>
<td>.905</td>
<td>2.02</td>
<td>3</td>
<td>58</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>HOTELLING'S TRACE</td>
<td>.104</td>
<td>2.02</td>
<td>3</td>
<td>58</td>
<td>.12</td>
</tr>
<tr>
<td></td>
<td>ROY'S LARGEST ROOT</td>
<td>.104</td>
<td>2.02</td>
<td>3</td>
<td>58</td>
<td>.12</td>
</tr>
<tr>
<td>Treatment</td>
<td>PILLAI'S TRACE</td>
<td>.151</td>
<td>1.609</td>
<td>6</td>
<td>118</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>WILK'S LAMBDA</td>
<td>.849</td>
<td>1.651</td>
<td>6</td>
<td>116</td>
<td>.14</td>
</tr>
<tr>
<td></td>
<td>HOTELLING'S TRACE</td>
<td>.178</td>
<td>1.690</td>
<td>6</td>
<td>114</td>
<td>.13</td>
</tr>
<tr>
<td></td>
<td>ROY'S LARGEST ROOT</td>
<td>.177</td>
<td>3.486</td>
<td>3</td>
<td>59</td>
<td>.02</td>
</tr>
<tr>
<td>Interaction</td>
<td>PILLAI'S TRACE</td>
<td>.204</td>
<td>2.228</td>
<td>6</td>
<td>118</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>WILK'S LAMBDA</td>
<td>.801</td>
<td>2.268</td>
<td>6</td>
<td>116</td>
<td>.04</td>
</tr>
<tr>
<td></td>
<td>HOTELLING'S TRACE</td>
<td>.243</td>
<td>2.306</td>
<td>6</td>
<td>114</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>ROY'S LARGEST ROOT</td>
<td>.217</td>
<td>4.259</td>
<td>3</td>
<td>59</td>
<td>.01</td>
</tr>
</tbody>
</table>

Table 4.9: Multivariate tests compare treatments, gender, and interaction vectors on three dependent variables simultaneously.

<table>
<thead>
<tr>
<th>Effect</th>
<th>ACHIEVEMENT SCALES</th>
<th>SUMS OF SQUARES</th>
<th>df</th>
<th>MEAN SQUARE</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>CONCEPTS</td>
<td>1.227</td>
<td>1</td>
<td>1.227</td>
<td>.418</td>
<td>.521</td>
</tr>
<tr>
<td></td>
<td>PROCEDURES</td>
<td>4.909</td>
<td>1</td>
<td>4.909</td>
<td>3.62</td>
<td>.052</td>
</tr>
<tr>
<td></td>
<td>PROBLEM-SOLVING</td>
<td>4.909</td>
<td>1</td>
<td>4.909</td>
<td>3.94</td>
<td>.062</td>
</tr>
<tr>
<td>Treatment</td>
<td>CONCEPTS</td>
<td>26.303</td>
<td>2</td>
<td>13.15</td>
<td>4.47</td>
<td>.015</td>
</tr>
<tr>
<td></td>
<td>PROCEDURES</td>
<td>2.545</td>
<td>2</td>
<td>1.273</td>
<td>1.02</td>
<td>.366</td>
</tr>
<tr>
<td></td>
<td>PROBLEM-SOLVING</td>
<td>.212</td>
<td>2</td>
<td>0.106</td>
<td>.078</td>
<td>.925</td>
</tr>
<tr>
<td>Interaction</td>
<td>CONCEPTS</td>
<td>4.364</td>
<td>2</td>
<td>2.182</td>
<td>.742</td>
<td>.480</td>
</tr>
<tr>
<td></td>
<td>PROCEDURES</td>
<td>15.636</td>
<td>2</td>
<td>7.818</td>
<td>6.28</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>PROBLEM-SOLVING</td>
<td>3.364</td>
<td>2</td>
<td>1.682</td>
<td>1.25</td>
<td>.297</td>
</tr>
</tbody>
</table>

Table 4.10: Univariate tests of significance compare treatment, gender, and interaction means one at a time.
Although less conclusive, other univariate effects revealed significant differences of performance on the Concepts scale and significant gender differences in performance on Procedure and Problem-solving scales. These simple effect differences favored mainly the Test-only group. The effects demonstrated that the Test-only group outperformed both the ATOSU plus tutoring and the tutoring approaches, which was diametrically opposite of the predicted result. This humbling empirical result was attributed to the Title I teacher’s use of guided instruction and one mathematics teacher’s use of a innovative program called Connections. Published by McGraw-Hill, the Connections program emphasizes the use of methods and resources that promotes communication and problem-solving, and was presently being observed by the Rand Corporation at Medina Middle School. Overall, both males and females seemed to benefit from these instructional methods, but males generally outperformed females.

In summary, there was no current evidence that supports any of the three experimental hypotheses. These results were possibly influenced by a number of unexpected factors like the poor reliability of the three dependent variables and the year-long treatment of the Test-only group with guided instruction and the Connections program.

4.4 Analysis of the Qualitative Data

The purpose of this analysis was to illustrate whether the qualitative data support anticipated or spontaneous activity that occurred during problem-solving. For this analysis two males and two females were chosen from the ATOSU treatment group, and two males and two females were chosen from the Test-only group and asked to try the four practice
tasks after taking their achievement test. The teaching-learning sequence -- verbal effort, visual representation, and mathematical operations -- measured by the ATOSU attributes was tested against teacher/student instances of problem-solving activity. However, as trained professionals, teachers were free to spontaneously comment as new assessment information emerged, whether their observations supported the model or not. The researcher and one teacher observed, rated, and audio recorded evidentiary information about each student. The advantage of describing and interpreting several kinds of assessment information is that student’s evolving understanding can be continuously monitored.

Eight cases were summarized using this information. These 8 cases each suggested an assessment theme and illustrated, by means of examples from the audiotapes, how teachers might interpret activity in an assessment situation to improve the learning environment. Assertions about the ATOSU model were generated and described at the end of the next section. To examine these assertions in relation to the qualitative data and the judge ratings, the researcher cites instances from the audiotapes illustrating the use of the ATOSU protocol.

4.4.1 Descriptions to Interpretive Prescriptions: The eight cases that are presented below describe each student’s capabilities and weaknesses using audiotapes of the lessons and the student’s responses. Each case presents a governing theme in italics, and analyzes short excerpts drawn from audiotapes of the ATOSU assessment sessions. Assessment records are employed as a tool for the dual purpose of monitoring progress and providing information for improving the learning environment. Teachers monitor progress by noting how, and by

116
what means, students meet conceptual and procedural levels defined by the ATOSU model. By analyzing and interpreting assessment information, teachers better make decisions that shape the learning environment to accommodate to the needs and learning styles of the student. These teacher decisions influence the ways the teacher and student might build the curriculum, mold instructional strategies, and set standards or expectations as viewed within this learning environment.

Case 1

Some students never appear to find meaning in classroom assignments or tasks. Their culture and world assign little meaning to activity that is independent from their own personal reality. Often times, such students complain that the topic is sterile and has little relevance to their lives. When confronted with a task that has no meaning, there is more of a propensity to experience confusion and resist trying. Grades have little effect in changing the behavior, since the student perceives that the grade is designed to manipulate his choices. Since the student resists adapting his behaviors to the teacher’s expectations, there is a greater chance that, in time, he will ignore direction and fail to acquire the necessary prerequisite skills to deal with the unknown. During questioning, the student might avoid responding or parrot the teacher. To join the students personal reality to formal topics found in the school mathematics curriculum, the teacher may need to offer a more “authentic” learning environment to keep these students motivated.

Playing card tasks have more of a chance of being successful when the student is familiar with the objects and the traditional ways in which they are used. An active search for a solution requires a curious mind and some level of basic understanding. If the student lacks the basic skills or is not motivated to understand the task, the teacher’s job becomes more complicated.

Teacher:...Can you find some patterns in the cards for me? [Pause] Any kind of pattern. Usually what people will do is think of a set of categories for classifying the cards.
Student: In terms of their order.
Teacher: Ok. Show me some order with the cards that you got.
Student: Put them in order. {Begins putting them in order by face value}
Teacher: By their face value. Yeah, is that what you are saying?
Student: Yes, Um... [Pause].
Teacher: Ok, so what kind of order do you have there?
Student: The face order like when its higher or lower
Teacher: Ok, so you have one pattern, do you see another pattern?
Student: Shape.
Teacher: Shape, what do you mean by shape?
Student: I don't mean shape, I mean faces.

The student’s confusion persists through all four tasks. During instruction and practice, this student often seemed confused when searching for patterns, constructing a tactic, or writing a set of mathematics operations even with previous instruction. The student would write down a response when prompted by the teacher, but he would ultimately write down the teacher’s explanations. Otherwise, there would be no response.

He seemed confused by the meanings of basic ideas like “fractional part.” This confusion seemed so extreme that the teacher searches for ways to remain positive despite the lack of progress.

Teacher: If I said to you, what fractional parts of these cards are black, what would you tell me?
Student: The top part.
Teacher: What would be the fractional part of that?
Student: Two by three
Teacher: Two by three is how you get six. I won’t argue that...that’s good. What would be the fractional part here?
Student: That are black?
Teacher: Yes, you have twelve possibilities; how many them are black?
Student: Six.
Teacher: Ok, Six is not a fraction but Six twelfths is. Follow me?

Teachers will often try to “negotiate” meanings of concepts when ideas go astray. Negotiated meaning occurs when the assessor and the assessed try to share meaning. To understand what binds the teacher and the learner, teachers must understand “where the learner is”, and students need to understand “what the teacher wants.” It may be possible that the student does not understand or make sense of what the teacher means by “fractional part.” Often times, teachers will demonstrate their meaning by employing example after example.

Teacher: Ok, let’s try another one. What fractional parts are queens?
Student: Four over four.
Teacher: Four over how many possibilities?
Student: Four, three, I mean two....
Teacher: No, there are four queens and twelve possibilities. Right?
Student: Oh yeah, that.
Teacher: Exactly, what’s a fraction? The number...If I looked at this circle and I looked at this shaded region, what would be my fraction?
Student: So you divide up the circle?
Teacher: What would be the fractional part that is shaded?
Student: Half.
Teacher: Half, right. How did you get half? [pause] There are two possibilities and one of them is shaded. Right.

Even after extended instruction and practice with the teacher on basic probability principles, he fails to make sense of this probability situation since he offers a denominator that would define an improper fraction.

Teacher: Do problem 1.
Student: {Silently reads the problem} Ok, I get ten diamonds.
Teacher: Ok, you got ten diamonds. Ten is number in your event, right. What's the denominator?
Student: Five.
Teacher: And why would five be in your denominator?
Student: I...[Pause]
Teacher: Is that how many you got? How many do you have all together?
Student: Forty.
Teacher: What would be the answer then?
Student: Eight out of Forty.

Even though this student seemed willing to engage very early in the instructional process, the gaps in the student's knowledge were difficult for him to overcome. At several points during the instructional process, he experienced difficulty and confusion with what to do next. He seemed overly dependent on the teacher, often relying on the teacher's cues and prompts for the next step.

Students who perceive little benefit from schooling or little meaning from the topic at hand are less likely to possess the background, interest, and the motivation to learn formal subjects like mathematics. In time, a student who perceives no meaning in schooling or no connection between the subject and his own life-experiences will more likely engage in unconstructive activity like resisting effort, acting out, growing reticent, and procrastinating. If the student remains alienated for extended periods of time, he may not have developed sufficient readiness and prior knowledge to be successful with such tasks.

Environmental Prescriptions: To meet the educational needs of the student, the teacher must reshape the learning environment by changing the instructional approach and assessment task. An important goal of mathematics instruction is to help students learn to communicate about mathematics. Mathematics may be viewed as a form of language. In this case, it is believed that the language of instruction does not appear to be the primary language of the student. Interpreting formal mathematics requires that the student recognize and interpret verbal, visual, and mathematical representations in equivalent ways. However, when students are from a different culture, attempts to promote communication are not always meaningful since the student does not understand ideas in the same way as the teacher or other students.
The teacher must teach the student to take ownership of his actions before academic progress can be made with this student. Educational method must focus on the reason for this helplessness so that the student recognizes the need to reflect on his actions and takes a "clear and lucid grasp" over these actions. The learning of formal mathematics only makes sense when, as a consequence, the student understands that he can transform his world with it. Learning formal mathematics must have meaning for the student, by requiring the student to reflect upon himself and the culture in which he lives. Authentic tasks can play a major role in this process since they join real world problems to formal or theoretical topics to transform lives and communities.

Case 2

When does the teacher know that the student is ready to progress to the next level? As students become more focused and can sustain their regulated or mediated learning activity, the teacher needs to do less prompting and more monitoring of activity. Students mediate their activity when they "voluntarily control" or "consciously realize" relationships that can be mastered as "internalized speech" or concepts. As a student becomes more fluent and insightful in solving problems, they need new, but related, tasks to move them to deeper levels of thinking. By changing contexts, the teacher could note how students apply their concepts and procedures.

In the instruction leading up to this snippet of activity, the student listened well and applied the previously trained skills to situations involving the population of face cards (twelve in a standard deck of playing cards). For task 1, the population of cards was changed from the face cards to the number cards without cuing the student. Although this change does not make for radical changes in context, it does encourage the student to note at least one change in problem conditions.

Teacher: Ok, so let's see if you can answer the question.
Student: [Reading Task 1] Using the number cards as the population, what is the probability of randomly picking a diamond?
Teacher: Can we write out what we're thinking or write a tactic?.....Like we did with previous examples.
Student: Yes, hearts, diamonds, spades...{Begins making the table}
Teacher: Your getting ambitious....
Student: The top number would be eleven.
Teacher: How did you get eleven?
Student: I added all the diamonds together.
Teacher: Ok, show me.
Student: One, two, I just...
Teacher: Ok.
Student: three, four, oh...ten.
Teacher: Ok.
Student: I then take ten times four and get forty.
The teacher should be attentive but encourage the student to think through the process by using questioning techniques. The goal is to move from teacher-talk to more of an emphasis on student-talk. The quality of the student’s paraphrasing of his/her thinking is assumed to be indicative of the quality of the student’s inner speech. Although this example demonstrates some ability to paraphrase, this student indicated an increasing fluency to employ verbal behavior in the form of paraphrasing by task 3 (see Table 4.12, deliberative effort, p. 133).

The student seems to have adapted well to the subtle change in problem conditions presented above, so the teacher might consider future changes that are more extreme in nature. The student needs to be further encouraged to generate her own tactics and write an equation that is represented by the tactic, labeling the elements of the equation. It is further beneficial to ask the student to provide written explanations where possible.

**Environmental Prescriptions:** Motivated and well-prepared students possess sufficient knowledge to readily integrate the activities of the classroom into their academic experiences, but do not necessarily transfer these skills to their real-life experiences. Too often students fail to see the relationship between the theoretical world of academics and the pragmatic world of everyday life. The ultimate goal is to develop a deeper understanding between the strategies of teaching, the structure and organization of a discipline, and the real-world experiences of students.

This student could learn to write the more general algebraic rule for defining the rules of addition, multiplication, and counting. These rules are said to be decontextualized since they describe the applications of operations in all contexts. A set of authentic, or “real world” tasks could be devised and administered to observe whether the student’s “mediated” thinking is sufficient to promote transfer under more extreme conditions.

**Case 3**

_Students often reflect about the problem and construct a model of it using a set of concrete objects. A student who can assign symbolic meaning to the object in an effort to abstract critical dimensions of the problem, separates the relevant from the irrelevant. Searching for and organizing the dimensions of the problem so that it lends itself to a set of equations, operations, or logical moves is often more difficult than carrying them out to reach a solution._

From the start of the first lesson, the student displayed an ability to generate tactics. Note her ability to visualize the relevant dimensions of a problem and represent them in some meaningful way (see Table 4.13, generative tactic, p. 134). In the written response in Figure 4.3 on the next page, the student again demonstrates her ability to generate her own tactic and write an operation for solving task 2 — the probability of selecting a queen or a black card from a deck of face cards in one draw.
Although she attended to instruction earlier in the lesson, she did not solve new problems in exactly the same way as the teacher. Generating her own tactic is abstract since it involves expressing qualities of patterns about objects (e.g., playing cards) in new and meaningful ways. Generating a tactic involves the ability to see and interpret new patterns in previously represented thoughts.

This student was particularly strong on probability tasks. She found short cuts or more direct procedural paths for simplifying operations that were not suggested by the teacher during instruction (see Table 4.14, modified operations, p. 135). Procedural development suggests an ability to connect verbal, visual, and mathematical forms to generate abstract inference on familiar content. Note that her verbal explanation in Table 4.14 is connected to her tactics and operations in the response above.

The student was introduced to statistical concepts by placing all the number cards into a box and asking the student to randomly draw three and calculate the mean. Instruction was designed to teacher concepts like random selection, population value (mean of all the number cards), and statistic (means, median, modes of samples). Despite this instruction, the student did experience more difficulty with statistical concepts like the mean, median, and mode.

Teacher: ...Ok, how do you find the mean?
Student: I know how I get the median, but I forget the mean.
Teacher: Oh, to get the mean I add this plus this....
Student: Oh, I got it now.
Teacher: Ok.
Student: {Works on the problem by adding the numbers}
Teacher: Ok, that your answer, what do you do next?
Student: { Works to divide the number by 6}
Teacher: What's your answer?
Student: Six point one.

The student could not remember how to calculate these statistics unless prompted by the teacher. It seems that she never connects this operational procedure to something meaningful -- i.e., why would you want to calculate sample statistics for random samples of number cards?
Environmental Prescriptions: The student needs to be encouraged to generate her own tactic while looking for patterns, writing an equation that is represented by the tactic and labeling the elements of the equation. She should be encouraged to provide written explanations where possible. The student must be asked to defend her representations — whether they are written, tactical, or mathematical. A set of authentic tasks need to be devised and administered so that the teacher can observe how the student self-regulates her own thinking once she has received guided support.

Case 4

Mathematical operations appear as a special case of problem-solving actions within a broader frame of verbal, visual, and mathematical procedures. A sense of mathematical operations is formed by aggregating related and recombined operations that is regulated and given meaning in procedural contexts. It is maintained that students develop a sense of mathematical operations like numeration, correspondence, estimation, and regrouping by varying the mathematical practice in varying contexts.

This student received the highest score on the National Assessment for Educational Progress (NAEP) from the Test-only group, which pleased but mildly surprised both his mathematics and Title I teacher. He did not complete the last four tasks on the NAEP exam, and those four tasks were the only items he missed. He was the only student to get item #10 correct, which can be described as a complex task with multiple mathematical steps (see Appendix C).

Although this student did not fare as well as others on the four ATOSU tasks, the assessment evidence indicated that, when he visualized, he projected his thinking on paper as equations. The student was reticent, so it became difficult to examine his verbal ability. The student did not employ any other form of visual representations like tables, graphs, pictures, etc. on any of the four assessment tasks. Here is one example of this student’s response to task #2 in Figure 4.4.

\[ \frac{\text{8}}{12} + \frac{\text{4}}{12} = \frac{\text{12}}{12} - \frac{\text{2}}{12} = \frac{\text{8}}{12} \]

Figure 4.4: Male’s response to task 2

The student could conceive the rule of addition without any additional representation but an equation. The rule of addition is inherent from the segmentation of the face cards into events A and B of a sample space, and the union of the two events are then calculated by adding the P(A) (read probability of A) and P(B), while subtracting out P(A) intersect P(B).
If A and B are mutually exclusive events, one merely adds the P(A) and P(B) (see Figure 1.1, p. 12). The student could conceive the proper operations without forgetting to subtract out cards that were both black and a queen. This trap often fools students who properly draw the tactic, yet this student writes the proper equations without the aid of tactical device. His equation writing ability did not help him solve either of the last two tasks, but it also seems that the student did not give as much effort to these problems. With the exception of task 2, his overall performance was average at best, even on tasks he tried to solve.

Environmental Prescriptions: Since the student demonstrated little verbal activity, the exercise of having the student write an explanation of how a problem was solved helps to demonstrate verbal understanding. Unless students frequently and explicitly discuss relations between verbal, visual, and mathematical representations, they are likely to view these representations as disparate ideas that need to be memorized. Connecting ideas is fundamental to the understanding of concepts and procedures. By having the student defend his thinking by paraphrasing orally or in writing, he is more likely to connect verbal communication to his visual and mathematical communications.

Case 5

Students learn by discovering ways to regulate their behavior that are less dependent on context and more dependent on their ability to generate abstractions. Students need to be asked to express their mathematical ideas orally and in writing. Early on, like Socrates, teachers guide this process by drawing the correct answer out of the student by asking questions and providing hints. How a student reacts to the teacher’s questions and hints indicates the nature of the student’s difficulties. Teachers need not dominate the remedial process. An error or misstatement that the student corrects after a few hints is not worth discussing; an error or misstatement that the student does not correct after a few hints is worth discussing. The scope and sequence of the practice problems needs to be defined and is manageable. Two sequences involving the same ideas are better extended and understood when one related problem follows another.

In the following set of examples, this male student learned to regulate his thinking by interacting with the teacher. To convey to students how mathematics works, the teacher has students use concrete objects to model procedures and concepts. The teacher introduces the student to the fundamentals and then attempts to extend them through varied practice. By guiding practice in this manner, the student eventually learns to mediate the process and generate abstractions. Note how the teacher employs questioning and hints. Simply telling students that their answer is wrong is not enough to correct their thinking.

Teacher:  Do problem 1.a
Student:  {Student picks up and arranges the number cards} Ok, I got ten diamonds.
Teacher: Ok, you got ten diamonds. Ten is in your event, right. What’s your denominator?
Student: Five.
Teacher: And why would five be in your denominator?
Student: I.....[pause]
Teacher: Is that how many you got? How many did you have all together?
Student: Forty.
Teacher: What would your answer be then?
Student: Eight out of forty.
Teacher: Eight out of forty. Why don’t you show me how you solved that problem? Show me the math and everything on paper.

As the number of steps increases, so does the probability of making a dumb mistake. This is a commonly used principle when developing tasks. With enough steps, anybody can be induced into making a mistake. Try multiplying 8979877 by 12454332 by hand. Its not likely that most people would get this problem right despite an ability to multiply. When relating tasks, it is more important to extend thinking by improving the student’s ability to make finer discriminations in related, but manageable, problems.

In task 2, to calculate the probability of a task with the rule of addition, one needs to makes finer distinctions than those performed in task 1 — calculating the probability of a diamond. Note that the first inclination of the student is to guess.

Teacher: Try 2-b.
Student: {Reads problem} [pause] . It would be ten out of two.
Teacher: No, show all your face cards.
Student: Set up all your face cards like before.
Teacher: That sounds logical [pause]. Ok now, what’s the event in this problem?
Student: A queen or a black card.
Teacher: A queen or a black card. Ok, show it to me.
Student: {Picks up a card or two}
Teacher: Is that your only queen or black card?
Student: Hearts.
Teacher: A queen or a black card.
Student: {picks up a red queen} Its a queen or a black card.
Teacher: Ok, are there any others?
Student: Here’s another.
Teacher: Ok, that’s a queen or a black card, any others?
Student: The queens.
Teacher: Ok.
Student: There are eight of them altogether.
Teacher: So what’s your answer.
Student: Eight out of twelve.
Current assessment practices need to reward students for overcoming difficulties, rather than penalizing students for having problems with new ideas. Misinterpretations and mistakes differ in the ease at making people aware of what they are doing. In the case of mathematical probability, dumb mistakes may include such things as miscalculating, failing to reduce, or picking up the wrong cards. A selective prompt or question may help the student correct these errors.

**Environmental Prescriptions:** Dumb mistakes and simple misinterpretations are the majority of errors and mistakes made. The students are more likely to blame themselves than the assessment tasks. Students have a tendency to attribute their poor achievement to a lack of ability or motivation, not what really causes the problem, e.g., insufficient practice in the use of new ideas. The teacher must shape the process by employing questioning techniques and providing hints. Oral cues, added to the students' creative thinking, spark new ideas and novel changes of thought. Oral cuing stimulates additional interest in creative thinking, instructional materials, and imaginativeness.

**Case 6**

*Students must be encouraged to adopt a point of view and defend it. Too often, the students mimic the teacher because they think it pleases or because they lack the confidence to try their own thinking. Instructional prescriptions relate objects (e.g., like playing cards) and instructional practices to the student's verbal, visual, and mathematical conceptions. For the purposes of mathematics instruction, teachers develop a point of view in their students by allowing them to define the spatial-temporal qualities of any task; by representing these qualities using pictures, diagrams, pictures, charts, equations, etc.; and by defending their representations as they progress.*

This student seems lacking in confidence and unsure of herself, seldom taking a step without first receiving a prompt. Even with many hints, she still has problems and often requires instruction when compared to peers. The student never seems to develop a definitive point of view; or, if she has acquired a sense of perspective, she is hesitant to defend it. She seems passive during learning, hoping that the teacher can explain all the material to her. After instruction using the face cards, the student answered as follows on task 1 using the number cards (includes Aces).

Teacher: What is the probability of drawing a diamond?
Student: Drawing a nine?
Teacher: A diamond.
Student: A diamond [pause] ten total. Ten total.
Teacher: Ten out of ..... [pause]
Student: Well...[pause]
Teacher: Well, how many diamonds are there?
Student: Ten.
Teacher: How many possibilities are there?
Student: Four.
Teacher: Four. When we were counting face cards, how many possibilities were there?
Student: Twelve.
Teacher: How did you get twelve?
Student: I added up the cards.
Teacher: Ok, so how many possibilities do you have there?
Student: Four.
Teacher: Times.
Student: Two.
Teacher: How many cards?
Student: Times ten.
Teacher: What’s the probability of randomly drawing a diamond?
Student: One out of thirty.

Unlike the student in Case 1, this student seems to bring some skills and prior knowledge to the table, but she lacks confidence and seems anxious. Because she was anxious, the student had difficulty generating any type of unique insights and seldom made sense of the problem. As compared to the student in Case 5, this student seemed to need more than a frequent hint or a question to bring her back on track. Such patterns are indicative of a student who needs instruction more than guided help.

Scoring the lowest of any female on the ATOSU tasks, this student’s anxiety can be viewed in the following snippet of activity.

Teacher: Why 8 plus 4? Tell me why 8 plus 4? Because 8 plus 4 is twelve?
Student: {giggles}
Teacher: How do you get twelve?
Student: I don’t know.
Teacher: There’s four here and four here and four here. What we try to do is fit the operations to the pattern.

A lack of confidence and stress will prohibit reflection and concentration on the part of the student. The teacher needs to help the student adapt to the learning environment. According to many studies, many females attribute their lack of performance to their lack of ability instead of a lack of effort.

Environmental Prescriptions: The student seemed to enjoy the tasks and tried but her efforts seemed tentative and unsure. The teacher should concentrate on building confidence in the student, attempting to slow down her thinking and encouraging her to express her ideas. A warm-up exercise could be used, covering material she has mastered but which is related to probability and statistics, e.g., fractions. The teacher might encourage the student to discover and describe patterns made with the cards. The student could be asked to create a system that demonstrates how fractions work using the playing cards.
Case 7

New scoring methods like the ATOSU approach for assessing students need to be examined for potential problems of bias and unfairness. Traditionally, the patterns of performance scores of any definable group may be compared to the patterns of performances of another definable group. Gender differences in student performance may be attributed to one or more of the following sources: the scoring methods, the socialization process, the teaching methods, or the attitudes regarding the subject or curricular content. When the scoring methods limit or misdirect the usefulness of performance, they give one group an inadvertent advantage over another.

Although only eight students were assessed, the scores on the ATOSU assessment demonstrated advantages in favor of the females over the males. As previously stated, there may be any number of reasons why these advantages in performance exist. However, the current evidence suggests no instances where females received any apparent advantage over males attributed to the scoring methods. And yet, this female far outscores all the males on the ATOSU assessment.

The ATOSU scores multiple facets, allowing students to demonstrate their competencies in a number of ways. Although playing-card tasks are not always familiar to all cultures or groups, there is no reason to believe that such tasks necessarily favor males or females. It was difficult to know whether males and females differed in their attitudes towards the subject or the socialization process they have been living in.

Unlike the four males examined, this student was fluent and insightful in applying previously taught tactics. For example, the student altered and extended her strategy to solve a problem requiring the rule of addition. Although the student was never taught the rule, the student altered her tactic to discover the answer (e.g., see modified tactic, Table 4.13, p. 134). This result contradicts findings in the research that males generally outperform females on tasks involving imagery or visualization (see Chapter 2, p. 56).

Although the tasks were mostly familiar, the student mostly focused and sustained mediated activity while achieving her desired goal (e.g., see deliberative effort, Table 4.12, p. 133). This student mediated activity when she “voluntarily controlled” or “consciously realized” relationships by constructing “internalized speech” to regulate her procedures.

These constructive behaviors seem to explain this student’s success more than her gender. It seems that such skills are advantageous to anyone who might develop them, so the fact that the student happens to be female and that she outscores her male counterparts does not appear to suggest bias in the scoring methods.

Environmental Prescriptions: Assessment should promote equity and fairness. Equitable assessment is achieved by creating conditions that are appropriate to the same extent for all students. In other words, if there are any gender differences, they may not be attributed to the assessment. Educators develop the methods to improve the quality of performance assessments by proceeding systematically.
Case 8

Young students can transfer abstract concepts from one context of application to another by analogy. The teacher's task is to identify parallels between the similar and different learning conditions and the problem, connecting these conditions by analogy with another problem. The teacher accomplishes this task by helping the students recognize and make these connections. The teacher helps students learn to understand and recognize these similarities and differences in problem conditions. The teacher shapes and molds the learning environment to accommodate to the learning styles of a particular student.

The teacher begins with the familiar notion of a fraction and how it relates to the twelve face cards in a deck of playing cards. The researcher points out that segmentation of the population of face cards may be expressed as fractions where one sector describes the number of possibilities in an event and all the sectors describe the total number of possibilities. For example, there are six black cards and this represents one half of the possible face cards.

Teacher: ...How many possible cards do you have?
Student: Twelve.
Teacher: How did you get twelve?
Student: There are three cards and four groups.
Teacher: Good, what fractional part of these face cards are black?
Student: Two fourths.
Teacher: Two fourths, how did you get two fourths.
Student: Two of these [pause]

Since fractional relationships may be tied to problems of chance, fractions may be assigned to the occurrence and nonoccurrence of one or more related events. The elements of these fractional sets may be expressed as unions, intersections, or mutually-exclusive compositions of events. The laws of probability then apply fractional operations to derive an answer. The teacher connects fractions to the meaning of probability.

Teacher: Ok, so you have one plus one is two, and one plus one plus one is four. So what we are going to be doing here is talking about a special kind of task. Remember, it is a task related to probability or the chance of something happening. Let's read question one.
Student: Given the number cards as your population, what is the probability of choosing a diamond on one draw.
Teacher: Ok, what event are you going to try to do?
Student: The number of diamonds out of the number of cards.
Teacher: Alright, show me a tactic.
Student: {Student writes answer}
Teacher: Ok, so what are you telling me.
Student: [Long pause]. Ten out of thirty.
Student: No, I mean ten out of forty.

129
Ann Brown (1989) describes two ways that teachers encourage transfer: "bridging" and "hugging." Bridging involves having students make connections between what is being learned. In this instance, the teacher asks students studying probability to explore analogies with topical events associated with fractions. When students are studying statistical concepts, teachers might encourage them to connect concepts like "sample" or "mean" in their everyday lives and try to identify how these concepts are useful.

"Hugging" involves relating targets of analogies that are relatively close in their relationship during instruction. This is commonplace in many disciplines: one practices how to act on stage or how to swing a baseball bat by playing the game. Seldom does any student directly practice mathematical concepts as they are applied in real life. Students learn a body of knowledge by working on tasks that require that knowledge. Teachers could then ask students to explain what is the same or different from problem to problem.

**Environmental Prescriptions:** There are always ample opportunities to engage students in relating higher level tasks. Teachers not only refer to the best tactics for representing the problem and to the operations students might use to solve them, but demonstrate how to apply these methods under [U]nique problem-conditions. The teacher demonstrates how readily mathematical concepts lend themselves to the solution of practical problems. While the teacher cannot illustrate every problem, he/she can develop logical reasoning which, if properly cultivated, will enable the student to analyze "real life" problems and arrive at a solution by some effective method.

ATOSU assessments attempt to build scaffolds that move each student from the familiar to the new. By monitoring the student progress, the teacher systematically gathers information that better informs the teacher’s decisions about the learner’s strengths and weaknesses. This process of gathering information is shaped and reshaped as assessment continues, depending on the teacher’s insights into the student’s conceptual and procedural knowledge and experiences. Sometimes the collection of information is informal, more impressionistic, and personal in nature; other times the collection of information is formal, measured, and structured. Moreover, students can be included in this process when teachers ask them to participate knowingly and actively in the process by self-evaluating their work. Hence, the data collection of assessment information is more provisional, yet it provides the diagnostic information for feedback necessary to a teacher’s environmental prescriptions.
The student's progress or lack of progress is significantly linked to the learning environment. The learning environment is controlled by the teacher and is ultimately concerned with planning curriculum, shaping instructional inquiry, and setting assessment standards. The ATOSU framework is merely an assessment tool that documents where students need improvement. Information can be used to document the reflections and analysis of the learning environment for a particular student. Teachers, then, can modify and shape the learning environment to meet the needs and learning styles of each student.

A set of assessment measures is apt to give the evaluator a more general reference of the student's progress and the learning environment that produced it. However, the eight cases presented here illustrate how the model may be used with different students to increase performance and improve the learning environment. Although the presented learning activities might be better suited to the specific situation, they may serve to demonstrate how the ATOSU model may be used with different students.

Teachers shape student thinking about mathematics by changing the learning environment, and assessment may be used to reward students for expressing their ideas and making suggestions. With guided instruction and guided assessment, teacher-led investigations encourage students to present and defend their points of view. When a student is having difficulty with a conceptual or procedural skill, the teacher can identify the basis of the problem and determine appropriate materials and activities to help students overcome the problem. If the instructional and learning approaches of teachers are to impact mathematical performance, classroom assessment needs to advance three fundamental assertions:
Assertion 1: Practice should be viewed as an opportunity for children to verbally, visually, and mathematically communicate their thinking about problems. To promote higher-order thinking, classroom assessment needs to focus more on this communication rather than the ability to get the answer.

Assertion 2: Verbal, visual, and mathematical forms of communication can be categorized into a generalized scheme that classifies and orders activity into more advanced or less advanced levels. In turn, these levels may be used as definable standards that teachers and students may undertake to achieve.

Assertion 3: Verbal, visual, and mathematical forms of conceptual and procedural thinking are connected, so it seems appropriate to connect them in assessment.

Given these assertions, interpretations are next summarized followed by actual examples illustrating these points. Tables 4.12 through 4.15 present some qualitative instances illustrating increasingly advanced levels of development anchored to a scale underneath each ATOSU attribute. Ratings on these scales can be illustrated with snippets of discourse and student performance from case records maintained for a select number of students chosen from the experiment. These snippets of discourse were chosen because they are representative of these judged levels of performance. To be included in this analysis, two judges must rate the student's performance in a similar fashion. Judges were encouraged to make written observations regarding problem-solving activity and mistakes before rating the response. This current effort represents an important attempt at triangulating this study's quantitative data with qualitative illustrations so that the reader might better understand the assessment procedures.
<table>
<thead>
<tr>
<th>Specific criteria</th>
<th>Qualitative Discourse</th>
</tr>
</thead>
<tbody>
<tr>
<td>no effort</td>
<td>None rated by any judge. All students exhibited some verbal effort.</td>
</tr>
<tr>
<td>minimal effort</td>
<td><em>Girl:</em> {picked up the face cards when the number cards are required}</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> Notice those are not the number cards.</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> Set up a table.</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> Yes, that is one way of doing it, but you don’t have to do it that way.</td>
</tr>
<tr>
<td></td>
<td>You don’t have to do it my way.</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> One twelfth.</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> How did you get one twelfth?</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> Cause {Chuckle}</td>
</tr>
<tr>
<td>effort</td>
<td><em>Teacher:</em> In doing this problem, what are you looking for?</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> Patterns and ....[pause]</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> What’s your goal?</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> Out of forty cards, how many are diamonds.</td>
</tr>
<tr>
<td>sustained effort</td>
<td><em>Boy:</em> How many ways do you combine a face card with a number card?</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> Little bit different problem.</td>
</tr>
<tr>
<td></td>
<td><em>Boy:</em> Do I just pick out the face cards? {Student works with cards}</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> Ok, I see three so far. Do you want to write an operation?</td>
</tr>
<tr>
<td></td>
<td><em>Boy:</em> Not yet.</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> See any others?</td>
</tr>
<tr>
<td></td>
<td><em>Boy:</em> Could you switch them around? {Student works with cards}....I got them.... Fifteen.</td>
</tr>
<tr>
<td>deliberative effort</td>
<td><em>Girl:</em> Ok, to show you how many ways I can combine them. First I put 10 and a J, and</td>
</tr>
<tr>
<td></td>
<td>then I do 10 and a queen, and I do 10 and a king.</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> Ok.</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> Then I just take the 10 away. Now I do 9 and a J.</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> I like the way you are doing this.</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> Then I do 9 and a queen. Then I do 9 and a king. Then I take the 9 away, and</td>
</tr>
<tr>
<td></td>
<td>I do 8 and a J.</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> All right.</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> 8 and a queen....8 and a king. Then I take the 8 away and I do a 7 and J, 7</td>
</tr>
<tr>
<td></td>
<td>and a queen, 7 and a king.</td>
</tr>
<tr>
<td></td>
<td><em>Teacher:</em> Can you tell me how many you got?</td>
</tr>
<tr>
<td></td>
<td><em>Girl:</em> Twelve.</td>
</tr>
</tbody>
</table>

Table 4.12: Snippets of discourse provided as illustrations of levels of attention
<table>
<thead>
<tr>
<th>Specific criteria</th>
<th>Qualitative Discourse</th>
</tr>
</thead>
<tbody>
<tr>
<td>no tactics</td>
<td>no tactic is presented.</td>
</tr>
<tr>
<td>cued tactics</td>
<td>3 Face cards 4 # cards</td>
</tr>
<tr>
<td></td>
<td>![Card Images]</td>
</tr>
<tr>
<td>visual tactics</td>
<td>Diamonds 10</td>
</tr>
<tr>
<td></td>
<td>9 8 7 6 5 4 3 2</td>
</tr>
<tr>
<td>modified tactics</td>
<td>4x4</td>
</tr>
<tr>
<td></td>
<td>![Card Images]</td>
</tr>
<tr>
<td></td>
<td>8/12</td>
</tr>
<tr>
<td></td>
<td>3 x 4 =</td>
</tr>
<tr>
<td>generative tactics</td>
<td>10/40</td>
</tr>
<tr>
<td></td>
<td>![Card Images]</td>
</tr>
<tr>
<td></td>
<td>I got 10 by multiply</td>
</tr>
<tr>
<td></td>
<td>I got 40 by x 10 x 4</td>
</tr>
</tbody>
</table>

Table 4.13: Snippets of activity representing levels of tactics
<table>
<thead>
<tr>
<th>Specific criteria</th>
<th>Qualitative Discourse</th>
</tr>
</thead>
</table>
| no operations     | *Teacher:* Have you ever done that (calculated the mean)?  
|                   | *Girl:* Yes.  
|                   | *Teacher:* How do you find the mean, do you remember?  
|                   | *Girl:* No.  
|                   | *Teacher:* Take all the numbers and add them up. Do you remember now?  
|                   | *Girl:* No. |
| cued operations   | *Teacher:* Ok, now I got a question for you. I got this king, and that king goes with that 8. And that same king goes with that 7. Let’s group these and I bet you see it right away. How many ways will that king go with each number card?  
|                   | *Boy:* Four.  
|                   | *Teacher:* And with that queen?  
|                   | *Boy:* Four more.  
|                   | *Teacher:* How many in all three?  
|                   | *Boy:* Twelve. |
| valid operations  | *Boy:* (written response) “I counted all the cards together, then took out 10 diamond cards. I added all the diamonds by using the ace as one, and knowing it goes to 10… others 30 cards.” |
| modified operations | *Teacher:* Explain how you get your answer.  
|                   | *Girl:* I’m going to add up all the cards and I got 2 kings and 2 jacks and I am going to add up all 4 of the queens. And I’m going to put them over the number twelve……  
|                   | *Girl:* I went 4 plus the 2 black jacks and the 2 black kings and I got 8.  
|                   | *Teacher:* What do you mean 4?  
|                   | *Girl:* The 2 black jacks, the 2 black kings and the 4 queens. |
| generative operations | Although one judge rated 2 student responses as generative, these ratings were not corroborated by a second judge. |

Table 4.14: Snippets of discourse used to illustrate levels of operations.
<table>
<thead>
<tr>
<th>Specific criteria</th>
<th>Qualitative Discourse</th>
</tr>
</thead>
</table>
| no solution or an incorrect      | *Boy:* Ok, I got ten diamonds.  \  
Teacher: Ok, you got ten diamonds. Ten is in your event, right. What’s in your denominator?          |
| solution                          | *Boy:* Five. \  
Teacher: And why would five be in your denominator? \  
*Boy:* I...[pause]  \  
Teacher: Is that how many you got? How many do you have all together? \  
*Boy:* Forty \  
Teacher: What would be your answer then? \  
*Boy:* Eight out of forty. \  
Teacher: Eight out of forty. Why don’t you show me how you solved the problem. Show me the math and everything. \  
*Boy:* I’ll tell you how I did it. I counted them all up and got eight, then I counted all of them and then I put them in a fraction. |
| cued solutions                    | *Teacher:*....Ok, now what’s your answer? \  
*Girl:* Eight. \  
*Teacher:* Eight out of.... \  
*Girl:* Four out of eight. \  
*Teacher:* No. How many possibilities were there? [pause] Eight is definitely in the event. \  
*Girl:* Twelve. \  
*Teacher:* Twelve. What’s your answer? \  
*Girl:* Eight out of twelve. |
| valid solutions                   | *Girl:* If we add up the black cards and the queen. \  
*Teacher:* Yes. \  
*Girl:* Then it would be likely. \  
*Teacher:* Ok, let me see you do that. \  
*Girl:* {Writes the answer} \  
*Teacher:* Ok, can you add them and give me an answer. \  
*Girl:* Eight twelfths. |
| reversed solutions                | There were no students who reversed their responses without help. |
| critical thinking                | There were no students who offered an alternative approach for reaching the solution. |

Table 4.15: Snippets of discourse used to illustrate levels of mastery viewed in a solution.
4.4.2 Interpreted Levels of Attention: Mathematics is partly a verbal exercise requiring the interpretation of language either in oral or in written form. Early learning experience in mathematical understanding, like rote counting or memorizing algorithms, begins with these skills. However, to move beyond rote counting or memorizing algorithms, students must learn to connect or link their verbal experiences to the interpretation of objects or patterns represented by objects. This process requires attentive effort so that students coordinate these connections without merely reciting verse. The [A]ttention attribute classifies verbal information in both oral or written form as levels of effort. Progressive levels of oral or written effort document gradual extensions or generalizations of verbal activity transformed into conceptual knowledge. Students verbally attend in more abstract ways when they can paraphrase their ideas in more sophisticated ways. This goes deeper than the ability to merely restate the question. Levels of attention are predicated on what can be confirmed in oral and written communication as the student endeavors to learn.

**Verification:** In Table 4.12, the specific criterion labeled *minimal effort* is an example of a female who makes an attempt to interpret but does not understand the meaning of the problem. Earlier during instruction, the student was using face cards to understand the concept of probability. The guess of “two twelfths” is similar to an earlier set of answers requiring twelve in the denominator. Because the new problem uses the number cards instead of face cards, the student is required to recalculate the number of possibilities — of course, there are 40 possible number cards when the ace is counted as a number card.

The next attention level in Table 4.12 is described with the label *effort*. In the example provided, the student repeats the question in a declarative form. Restating the problem is indicative of effort, provided the student can proceed to the next step or produces
a possible alternative for solving the problems. When students' verbal reasoning is not yet fully developed, they often have difficulty expressing themselves. Alternatively, the teacher can encourage a student to represent his/her thinking with objects or with a tactic, but the teacher should encourage students to express their thinking first in written form and then orally.

The third specific criterion is sustained effort which requires that the student continues the search for a viable alternative, or systematically modifies the approach once any attempt becomes futile. Students will commonly adjust their thinking when the process is stymied, but they still attack the problem in more conventional ways. This male student's explanation is more his own rather than a restatement of the problem -- a step beyond the "effort" criterion. The student doggedly proceeds down a chosen path toward a solution although he ultimately misses the problem. This process fails because the student's voluntary attention seems focused more on the specific step than on the overall process. As in this snippet of discourse, verbal reasoning often becomes confused or circular when the student does not keep the overall goal in mind.

The final specific criterion is deliberative effort which requires that the student generates verbal reasoning that is insightful, planned, and systemic. This type of verbal thinking might involve verbal insights derived from changing one's point of view or an ability to methodically carry out a long process to its fruition. In the example, a female verbally describes the process she uses from beginning to end. Although the task conditions require little insight, the student's response is characterized as deliberative since it is systematic and the student seems to always keep the goal in sight.
4.4.3 Interpreted Levels of Tactics: When presented with a problem, the student needs to explore the problem's conditions and organize relevant and prominent patterns. Tactics are cued visual tactics early in the learning process, so mathematics teachers will make many of the selections and direct the process of connecting verbal and visual communication. Teachers generally use hands-on materials like playing cards as objects that may be symbolically represented in the mind. Teachers guide students to encourage pattern attribution and connections between verbal and visual forms of communication. When the student experiences confusion, the teacher cues the conceptual interpretations and procedural processes using the objects. Table 4.13 displays specific criteria for judging tactical approaches. Within this context, the \([T]\)actic attribute depicts the student’s ability to externally represent what is imagined in the mind.

Verification: The figure adjacent to the specific criterion labeled cued visual tactic represents an occasion where a student may need guidance. In the portrayed problem, the student is asked to determine the number of possible ways one can combine three face cards with four number cards using two cards at a time. The student discovers three possible pairings with one card left over. The student fails to reuse cards once combined with another card. Typically, the teacher would need to cue an additional move thereby allowing the student to see other possibilities on his/her own.

Eventually, as the student develops, he or she learns to independently select an appropriate visualized tactic to organize these patterns. As in the example in Table 4.13, the student oftentimes uses the very tactic employed in the lesson but the teacher’s assistance is not required. The student’s written explanation was cued after the student pointed out that he did not have to finish the table to solve the problem. This step represents progress but it
seldom develops the contextual readiness necessary for understanding mathematical concepts and procedures. Contextual readiness develops when students become aware of how concepts and procedures are flexibly applied as problem-situations or conditions are altered (Kennedy & Tipps, 1995). Familiar problems may be handled in this manner, but unfamiliar problems demand that students continually change perspectives to solve them.

When problem contexts change or when the problem conditions are subtly altered, visual representations must be modified to accommodate problem-solving. Even when familiar tactics are employed, they often need to be extended or modified. This notion is demonstrated in the tactic displayed next to the specific category entitled modified visual tactics. When presented with the face cards, students were asked to find the probability of randomly drawing a queen or a black card. Students unfamiliar with the rules of addition or multiplication could use previously learned tactics to solve problems using these complicated concepts. In this example, the student alters the “make a table” tactic by drawing circles around both events. He eventually solves the problem by properly counting the circled cards, refusing to count cards circled twice more than once. Such a tactical representation may then be used as an aid for writing mathematical symbols or equations.

Young students have conceptual networks that are less developed when compared to adults (Chi & Koeskie, 1983), and, because these students are less set in their ways, they often are capable of visualizing problem situations in more imaginative ways than adults. Generative thinking may be characterized by the student’s tendencies to change their perspective. Generative attributions are necessary as the problem conditions are altered, but students will often find unique ways of solving a problem even when the problems are
familiar. When these insights are valid, it is important that teachers reward them. In the example next to the specific category labeled *generative visual tactics*, the student draws a picture tactic. The student not only realizes that the number cards are the appropriate cards, but that she need not represent this by making a large table. Instead, she decides to draw the problem as a more compact picture that bundles the cards by their suites. She does not imitate the teacher but generates her own approach for solving the problem. Once more, she can verbally explain her tactic in her own words.

### 4.4.4 Interpreted Levels of Operations:

Student's early understandings of mathematical concepts and procedures develop from informal experience that are assumed to be connected to verbal communication with others and visual communication with objects. As they play they have opportunities to count objects, attribute spatial relations, compare heights and distances, note part/whole relationships and engage in background-building activity. Many Title 1 students know how to perform formal addition, subtraction, multiplication, and division operations by middle school, but few students understand that these operations can describe many situations. To discover the usefulness of these operations, students need to link these operations to their verbal reasoning, to their tactical attributions, and to their previously learned operations as problem conditions change. As a result, as time passes, they are less likely to forget learned concepts and procedures.

**Verification:** Table 4.14 examines levels of mathematical [O]perations that are assumed to be increasingly more developed. The specific criterion labeled *no operation* provides an illustration where a female student remembers that she was taught the concept
of a mean; but she has no sense of what a mean is, so she has no idea how to find its value. Although there is a fleeting memory of a past experience with the material, there is a total absence of meaning. When students verbally and visually conceive the meaning of a mathematics concept, it is believed that they will more likely connect these conceptions to their verbal, visual, and mathematical procedures. Meaningful connections are fostered during varied practice when students compare problem situations and problem conditions. Once a sufficient number of connections are made, forgetting is less of a problem.

Students may be cued to remember an operation by the teacher. The specific criterion labeled *cued operation* is a conceptual level where students are prompted to recognize mathematical meaning represented by a problem’s operations. Although the student essentially remembers the operations, the memory of the operations is seldom identified and selected given the problem situation, the visualized patterns recognized, and the informal language employed. In Table 4.14, the mathematical operation is remembered in bits or pieces because of the student’s limited conceptual network. The teacher guides the step-by-step process because the concepts underlying the process are not fully developed. By expanding the conceptual network through guided practice encouraging connections between verbal, visual, and mathematical communication, interpretations may be made that attribute meaning. Such connections develop operational sense by relating the mathematical syntax and symbols of operations to the syntax and forms of other forms of communication.

Students may independently employ operations once they have developed the operations sense necessary to discriminate between a variety of mathematical concepts presented in familiar problem situations. Students develop operations sense by interrelating
verbal, visual, and mathematical forms of communication. The specific criterion labeled *valid operations* presents an example of this level of mathematical development. Although the example is not fully complete, the student is able to explain his procedures for computation as related to his presented tactic. He can orally describe the probability situation in his own words and apply the appropriate rule. He has begun to build up mental images of probability events as subsets of general sets. In time and with varied practice, he eventually understands how these mathematical operations are distinguished from or related to others and how to potentially use estimation techniques to check his progress.

The specific criterion labeled *modified operations* emphasizes the student’s increasing power to extend the operations or utilize mathematical properties to organize or rearrange the operations. As in the example in Table 4.14, the female student’s response represents an increasing ability to flexibly apply mathematical operations as the problem conditions change in related tasks. This problem required that she discover how to employ the rule of addition— even though it was not taught— given an overlapping event space. The student links her informal, intuitive notions of card patterns to the abstract language and symbols of mathematics. She successfully communicates how she produced her mathematical sentences (shown in the example) while drawing logical conclusions from the operations (not shown in the example).

No student employed *generative operations* as rated by both judges. Students who can apprehend operations in a generative fashion do so with insight and penetration. Although rating this criterion is often a judgment call, such skill is marked by the student’s ability to make mathematical sense of unfamiliar situations — in other words, their capacity
for transfer. In this study, the students’ limitations do not entirely explain this lack of generative thinking. The tasks must be designed to promote such thinking. Due to time limitations, tasks requiring contextual change were not administered to the more successful students.

4.4.5 Interpreted Levels of Solutions: [S]olutions describe levels of mastery that students exhibited in order to solve a problem. The possible success of any student could be described in terms of his/her ability to answer questions at some level of depth. For example, students who can defend their answer are more sophisticated than students who merely give an answer. To build learning skills, students must be encouraged to understand the thinking behind their answer. Constructing meaning that relates a derived solution to the chosen process develops mathematical reasoning. Checking the answer by working backwards or by employing estimation skills develops fluency and understanding of mathematical concepts and procedures.

Verification: In Table 4.15, the specific criterion labeled no solution describes a student who cannot find a valid solution even when prompted. Reasoning and communication are inherent in drawing conclusions when using mathematics. The example indicates a common occurrence where students believe that they are thinking properly but their reasoning is faulty. Teachers can correct faulty reasoning by having students manipulate objects like playing cards to demonstrate their reasoning. Activities and materials used in guided instruction and assessment require that the student search for patterns, organize the problem information, describe mathematical forms and generalize.
about them, and draw conclusions. This form of feedback is believed to be more natural and less damaging to the student’s self-concept and perceived effectiveness.

The specific criterion labeled *cued solution* describes a student who must be cued to recognize the correct answer. In the example, the student’s reasoning is confused since her interpretations are sporadic and incomplete. Although she has some sense of how to work through the problem, she needs to be prompted at critical points in her reasoning. This form of instruction is not antithetical with guided instruction provided that the students are given the freedom to justify their thinking. This requires patient pedagogy and a willingness to be an active listener.

The specific criterion labeled *valid solution* describes a student who can answer the question correctly without the teacher’s help. The student’s response is made with little help from the teacher. This level of mastery represents what is typically counted as correct on closed item formats. Although this level of mastery is important, the educational process should not stop at this point. Poor performance may be increased to average or high performance by requiring students to go beyond correct answers. The ATOSU data indicate that, with the possibility of two exceptions, most students have a long way to go before mastery is achieved.

The deeper specific criteria of *reversed solutions* and *critical thinking* were not met by any student in the opinion of both judges, so no examples of these levels of mastery are available. In reversed solutions the student learns to independently work backwards or make estimations that help the student make sense of the problem. In critical thinking the student can validly identify alternative approaches to the one used to solve the problem. Teachers
encourage this thinking once a solution is reached so that, in time, the student can apply them as checks in the problem-solving process.

4.4.6 Interpreted Levels of Unique-Conditions: Every assessment task implies environmental conditions that are contextual and alterable. While the teacher cannot illustrate every type of problem, he/she seeks to develop logical reasoning which, if properly cultivated, will enable the student to analyze his/her own problems and arrive at their solution by the most efficient methods. For example, students understand fractions so that they become useful to them in their lives. Classroom standards are too low when the goal is to learn fractions that reflect conditions inherent to the classroom and not to real life. This means that the students must take concepts and procedures previously learned in schooling and transfer them to a real, significantly more different task — like engineering a bridge, baking a cake, or when figuring out what wrench to use when assembling that new bicycle.

Classroom tests that do not monitor the students' ability to apply and transfer their knowledge across different contexts are less informative. A broader theme seems necessary since mathematics underlies virtually all phases of a complex society.

This research was limited to simulated tasks with playing cards, rather than the real life or "authentic" tasks stressed by assessment experts (e.g., Wiggins, 1989). But the model does allow for the development of these tasks as a necessary next step for measuring transfer and understanding (see Appendix A, [U]nique-Conditions). The timing of the experiment and the limited number of instructional sessions prohibited a thorough test of the model under both simulated and authentic conditions. This point is demonstrated by the
tasks descriptions of Appendix B which describes no task that can be construed as a practical problem found in some trade or field, such as carpentry, electricity, or medicine. This limitation needs to be addressed in future research.

The ATOSU model presumes an open environment that calls for collaborative and trusting relationships among teachers and students, teachers and parents, and parents and students. Although not directly measured by the model, these relationships ultimately impact upon the progress of the student because they define a shared responsibility. No party can be fully effective without the cooperation and commitment of all the parties. Such unmeasured conditions will also become an important focus of future research.

4.5 Summary

This chapter summarized the three research analyses and the four types of interpretations performed on the quantitative and qualitative data. The first analysis employed a measurement design to examine the validity and reliability of the ATOSU model. Preliminary indications suggested that the measures have potential for producing valid and reliable assessment information. Such assessment information can be potentially used to guide student learning and better direct instruction after more research study.

The second analysis employed an experimental design to compare differences in the test-only, tutored-only, and ATOSU treatment conditions. A significant interaction effect was graphically displayed and interpreted among the three groups. The interaction effects were concentrated mainly in the Procedure scale. Females in the ATOSU group did better in Procedure thinking when compared with females in the other two groups. However, ATOSU males performed at lower levels of Procedure thinking when compared to males in
the other two groups. The lack of scale reliability probably limit the significance of outcomes on the Conceptual and Problem-solving scales due to the increased probability of a Type II error. The third analysis employed an observational design to illustrate how the concepts and procedures provide an evaluation process using the ATOSU model. A set of three assertions was posited as fundamental to ATOSU assessment. Varied practice was viewed as an opportunity for students to verbally, visually, and mathematically communicate their thinking, and assessment needs to focus on these communications. Such communications may be classified in ways that allows the teacher to differentiate between more advanced or less advanced cognitive functioning. Whereas it is useful to differentiate among conceptual, procedural, and problem-solving goals for students in the early stages of their mathematics thinking, these distinctions should begin to blur as students mature mathematically. For this reason, assessment seems more appropriate when it recognizes how conceptual and procedural thinking comes together during problem-solving. Observing and illustrating what happens when the model is used, the variations in the ratings, what students say, and what students do, is necessary to a comprehensive approach to research. The qualitative analysis employed elements of each of these approaches to make interpretations and verify them.
Chapter 5
Conclusions, Implications, and Suggestions

This chapter discusses the conclusions and implications of the conducted research, and presents ideas for future research. The conclusions are framed around the research questions and inferred from the empirical analyses described in Chapter IV. Implications to the field address potential applications of the research results to the classroom. Ideas for future research describe investigation needed before the ATOSU model can be implemented on a mass scale in classrooms. It is hoped that the ATOSU model can be an effective tool in a school reform initiative once the research and development process ends, but this dissertation is merely the start of the developmental process rather than the end.

5.1 Conclusions

Several conclusions can be made given the analyses performed in Chapter IV. These conclusions follow a restatement of the three research questions guiding this study.

5.1.1 How valid and reliable is the ATOSU model?

Although establishing validity and reliability is an on-going process, results of the research indicate that the ATOSU model demonstrates some semblance of validity and reliability. To establish the content validity of the ATOSU framework, experts on the researcher’s doctoral dissertation committee as well as the two participating teachers agreed that the model’s scoring procedures are appropriate and that the tasks were representative of the probability and statistics domain. The predicted validity of the hierarchy was then
established by examining the simplex relationships in the correlations between the [A], [T], [O],[ S], and [U] constructs and by using a Rasch-based Facets Analysis. A simplex pattern was observed in the correlation matrix of the [A], [T],[O],[S], and [U] constructs, indicating a single order or direction down the rows or across the columns of the correlation matrix. A correlation matrix demonstrates a simplex pattern when correlational pairs of a chosen variable with a set of variables can be ordered by a rule. In this case, the rule is that the closer in proximity between a chosen variable and one of the other remaining variables in the correlation matrix, the higher their correlation. The further in proximity between a chosen variable and one of the other variables in the correlation matrix, the lower their correlation. For example, if one examines all correlated pairs of variables formed with the variable [A], the correlation between [A] and [T] should be higher than the correlations between [A] and [S] or [A] and [U], since [A] is closer to [T] in proximity in the correlation matrix, or if one examines correlated pairs of variables formed with the variable [O] as another example, the correlations between [O] and [S] and [O] and [T] should be higher than the correlations between [O] and [U] or [O] and [A], since [S] and [T] are closer to [O] in their relative proximity in the correlation matrix than [U] or [A]. When the correlation matrix in Table 4.2 is examined in its entirety, most of the correlations in ATOSU correlation matrix fit this simplex pattern. However, if one examines all correlated pairs of variables formed with the variable [S], one exception can be observed by examining the lower correlation of the [S] and [U] variables (closer in proximity from [S] in the matrix) when compared to the higher correlations between [S] and [A] or [S] and [T] (further in proximity from [S] in the matrix). Since these are the only exceptions in the correlation matrix, a hierarchical relationship is suggested among the five attributes.
A Rasch Many-Facets model was employed to estimate student, task, judge, or ATOSU attribute parameters (Linacre 1989). If one desires to know something about the parameters in the model — what an estimate is or whether to improve an estimate — one has to observe the problem-solving activity and rate a response in accordance with the parameters of the model. Whether or not the ATOSU attribute, judge, task, and the student observations fit the Rasch model depends, in part, on how closely the pattern of observations match the logistic form of the Item Characteristic Curve (ICC). Analyses revealed that each of the three facets (scaled attributes, tasks, and judges) and the object of measurement (students) fit and support the use of the Facets model.

When one observes the same set of ordered ratings across multiple tasks in an assessment, one concludes that the relative difficulties of the ordered attributes within the framework do not vary from task to task with the exception of the [U] attribute. Rasch evidence demonstrates an invariant hierarchy or common scale across the [A], [T], [O], [S], and [U] scaled attributes developed in this study. The Rasch measures of the ATOS attributes indicated that these learning constructs become ordered and separated in the predicted fashion, while rating of the [U] attribute varied with the judged difficulty of the task conditions. Since the [U] attribute is formulated around task conditions rather than learning constructs, it seems reasonable that future formulations of the Rasch Many-Facets analysis might treat the [U]nique-condition attributes as a separate facet.

The reliability coefficients based on the ATOSU attribute separation, judge separation, task separation, and student separation were all over 0.90. Other Rasch-based estimates, like the Root Square Mean (RMSE) and the Separation Index, confirmed the
predicted order of the ATOSU constructs, lending further credence to a hierarchy. Thus, the verbal, visual, and mathematical procedural order is generalizable for the four tasks and for the eight students assessed during the three lessons; however, these results are limited by the number of students and tasks employed. Moreover, although these results appear promising, they are potentially tainted by the biases of the researcher who performed as a judge across all persons and tasks. While researcher participation in the instruction and judging was always recognized as a limitation of this study, it became necessary due to a lack of teacher time, researcher resources, and funds. Future research should eliminate or limit the principal investigator’s role as a judge and teacher.

5.1.2 Are there treatment and gender differences in the experiment? Are there interaction effects related to ATOSU treatment and gender?

MANOVA was employed to compare the group performances on three dependent variables gathered from a balanced, two-way design. A two-way design enables an examination of the joint effects of two independent variables (treatment methods and gender) on three dependent variables (concepts, procedures, and problem-solving). A significant treatment/gender interaction was indicated by a significant Wilks Lambda. This significant interaction effect suggests that males and females varied in their success on the three National Assessment of Educational Progress (NAEP) measures according to their treatment assignment. Univariate analysis revealed that the interaction effects were mainly found in the Procedural scale. Graphs of the interactions revealed that females from the ATOSU group have more Procedure knowledge when compared to females from the Test-
only and Tutored-only groups; on the other hand, males from the ATOSU group have less procedure knowledge when compared to males in the Test-only and Tutored-only groups.

Since MANOVA analyses generally did not support one type of assessment feedback as superior for both males and females in a particular treatment group, conclusions made as a result of the experiment are less definitive. A general statement about the treatment effect is misleading because the effects come and go depending on the NAEP scale or whether the groups are male or female. This interaction was mainly the result of the superior performance of the Test-only males on the NAEP scales -- particularly, on the procedural scale. It is not entirely clear why Test-only males outscored all the other students since the research did not follow the Test-only group prior to NAEP testing. However, different from the previous school year when the ATOSU and Tutored-only groups were tested, the Test-only group received a mathematics program called Connections and guided instruction in the Title 1 program. It may be that the Test-only males greatly benefitted from this program as compared to other students in this study; or it may be that the males benefitted from guided forms of instruction implemented by the Title 1 teachers during the school year; or it may be that the Test-only males had an advantage due to the more traditional gender differences summarized in previous research. These and other possible explanations cloud the analysis and interpretations of the multivariate interaction effect. Likewise, the treatment and gender main effects could not be examined due to the multivariate significant interaction. Therefore, interpretations involving the multivariate test effects remained unclear and inconclusive.
MANOVA results may have been deleteriously influenced by the learning loss attributed to length of time between treatments and the relatively low reliabilities of the Conceptual and Problem-solving scales used as the dependent variables. The timing of the treatments was problematic because of the number of students that needed to be tutored. Students seemed to need more opportunity for varied practice at closer points in time. A better test of the model might occur when students are trained in groups rather than as individuals, because there would be more opportunities for practice, instruction, and assessment. The low reliabilities of the dependent measures may have reduced the power of the statistical tests and this malady may have been responsible for these inconclusive results. More reliable measures of the dependent variables should be utilized in any future study.

5.1.3 What are some major themes within the qualitative data when categorized by the ATOSU attributes? Do the qualitative data support the use of the ATOSU protocol?

The first part of the qualitative analysis portrays eight cases that illustrate the strengths and weaknesses of the individual students. An attempt was made to characterize the students patterns of activity and responses in relation to the quality of, or reasoning behind, a student’s responses. An additional attempt was made to use this specific information as a guide for planning instructional activities and environmental conditions that address the interests, needs, and learning styles of the students. These interpretive endeavors were performed in an effort to illustrate potential assessment practices and procedures. Judges needed to develop hypotheses related to the meaning of the students’ responses from
task to task. The difficulty with such a procedure was that it required an in-depth knowledge of the student’s abilities, familiarity with the connections between the different tasks, and an intuitive sense to notice and tease out patterns in both the qualitative and quantitative data.

The ATOSU framework served as a predetermined set of categories where qualitative snippets of externally represented communication could illustrate each construct. Specific instances relate problem-solving activity to different levels of the ATOSU framework. Each instance illustrated a progressively more complex level of communicative thinking that defined a set of standards. These qualitative instances allow a judge to experience the connections between verbal, visual, and mathematical representations, allowing problem-solving phenomena to be departmentalized before, during, and after an assessment session. These exemplars of communication may be reviewed and judged on their own merit by the reader. However, potential interpretations are reinforced by tracking patterns of qualitative and quantitative data through time. Reproducibility of the patterns shown within the frequency tables needs to be triangulated, monitored, and interpreted with qualitative exemplars over occasions and tasks.

5.2 Implications

In ATOSU assessment, students are rewarded for discovering relationships or connections between verbal, visual, and mathematical forms of communication. The measurement results summarized in this dissertation indicate that the ATOSU assessment monitors hierarchical conceptual and procedural forms of evidence that seems increasingly
more complex. Although the measurement results are still considered preliminary, the initial evidence suggests that the ATOSU model does provide some valid and reliable evidence for analyzing some problem-solving conditions. Facets analysis demonstrated that the predicted problem-solving behaviors can be scored with a high level of agreement even on a limited number of tasks. However, Guilford’s method for inter-judge reliability did suggest the same promising results. Guilford’s method used the Spearman-Brown Prophecy formula to determine the reliability for more than one judge. Results indicated moderate reliability for the [A]ttention and [T]actic attributes, but low reliability coefficients for the [O]perations and [S]olution attributes. Generally speaking, there were lower reliability coefficients for attributes whose ranked values seldom varied among judges. Measurement results do not merit adoption of the model into classroom practice until more is known about what teachers, parents, and students might think about this assessment approach.

Guided assessment is learning about students and developing their thinking using the lessons learned from these experiences. Guided assessment, utilizing a problem-solving approach, encourages teachers to understand how students analyze problem situations, question hasty generalizations, and express themselves clearly, precisely, and with insight. Tasks and scoring applications are used to introduce new mathematical content, to help students develop both an understanding of concepts and a facility to apply procedures, and to apply and modify their concepts and procedures given changes in problem conditions.

The assessment process in classrooms can embody an emphasis on procedural and conceptual outcomes with higher expectations for all. Although experimental results did not reveal general advantages in using guided-assessment approaches like the ATOSU model over other forms of assessment feedback, it provided some meaningful lessons about what
was working. For instance, the experimental results demonstrated that there was plenty of room for improvement even though a multifaceted assessment approach was employed. The ATOSU guided-assessment was most effective for ATOSU females given how the ATOSU information was utilized in the experiment. One year latter, Test-only males benefitted more than all but the ATOSU females given their educational experiences. Although not definitive, the results suggest that the two genders may respond differently to alternative methods of classroom assessment. Females may need different forms of feedback, or learn from different kinds of tasks as compared to males. More research needs to be done to note whether these differences replicate in other samples and, if reproducible, clarify the nature of these differences.

Eight cases were presented that illustrate how a variety students reasoned through a problem during an assessment session. Each of the cases represent proposed learning themes in an attempt to describe two kinds of phenomena: (1) how students best learn from their learning environment and (2) how teachers might use assessment information to make better decisions about the learning environment, thereby, improving problem-solving. Classroom assessment is learning about students as they interact with the teacher and other learners in the learning environment. Each student enters the learning experience with different prior knowledge and abilities. What learning conditions work for one student may not be so appropriate for another student. Classroom assessment approaches work when they ultimately focus on the individual learning styles consistently and routinely. The learning experience and tasks planned by the teacher must be adapted to the interests, needs, abilities, and strengths of the individual students.
The qualitative results helped to illustrate how the model may be employed to define multiple levels of performance or standards and how classroom assessment might better inform how expectations may be flexibly employed to motivate. When setting adequate and appropriate goals for students, teacher’s expectations should be clearly delimited and understood. It appears that the ATOSU criteria potentially outline what students should come to know and be able to do to develop multiple levels of competency. A student’s response is not simply classified as right or wrong. Instead, judgments are made about the “quality and appropriateness” of a response as students apply their conceptual and procedural skills. Classroom standards may be viewed as “level(s) of excellence or attainment regarded as a measure(s) of adequacy” (Madaus, Raczek, & Clarke, 1997, p. 6). As the qualitative analysis illustrates, these levels of quality and appropriateness potentially operationalize and sanction a “common framework of meaning among judges — shared standards for recognizing what is important in performance and mapping it into a summarizing structure” (Myford & Mislevey, 1994, p. 1).

Classroom assessment that is a source of useful information to the teacher encourages the possibility for future growth by encouraging reflective practice about the student vis a vis the learning environment. Teachers control and alter the learning environment by designating whether classroom activities are performed in groups or as individuals, defining the scope and sequence of the curriculum, altering the tasks conditions, and designing the instructional interventions. Assessment information that improves instruction optimizes the decisions teachers make about the learning environment as deemed necessary to the teacher’s purposes.
Overall, if classroom assessment potentially contributes to student success by increasing conceptual and procedural learning, it must furnish information that guides not only the kinds of tasks students are asked to engage in, but also the kinds of environmental conditions teachers control as they teach the learner. Guided assessment could then be naturally employed as a complementary activity to mathematical instruction to promote problem-solving and reflective decisions about the student vis a via his/her environment. Monitoring such measures of mathematical problem-solving seems increasingly plausible and inevitable given the advancements in computer and scanning technology.

5.3 Suggestions for Future Research

Since guided assessment demands increased staffing, more teacher time, intensive observation and a more developed knowledge of classroom assessment, research is necessary before implementing such an approach on a larger scale. Several concerns present themselves before the ATOSU program is implemented.

Guided assessment has potential for improving the teachers’ understanding of their students when employed over time; but because of the time involved, teachers need to be committed to such methods and form a theory of action (Patton, 1986; McCutcheon, 1995). Formulating a programmatic plan of action is a step beyond implementation or outcome evaluation. For this reason, teachers’ opinions regarding the utility of guided assessment methods need to be understood so that the materials and activities employed are related to actual outcomes and impacts that can constitute a more quality program. The focus should be on intended use by intended users. Clarifying a theory of action means specifying how
program staff believe that what they do will lead to the desired outcomes, step by step. The researcher then compares the intended program with how it might be actually used by teachers to make appropriate adjustments in teacher training or the ATOSU program itself. The intent is to improve the program by monitoring its dynamics so that it better meets its current and future goals.

Even though the experimental results did not fully support the assumption that guided assessment can contribute and lead to assessment gains, qualitative follow-up did illustrate how teachers might make assessment more pedagogically meaningful and the learning more challenging. However, there are problems with the ATOSU framework that must be resolved before committing time and money to future experiments. A valuable aspect of learning why students succeed or fail involves observing the interactions between students and teachers. The assessment act is but one way teachers and students interact. If the purpose of assessment is to improve teaching and learning, the teacher’s style of assessing and responding to a student can be as much a part of the problem as any learning difficulties experienced by the student. If one is to understand the teacher’s style of assessment and feedback, one needs information about how teachers interact with students. Therefore, any assessment designed to assist in reform would need to produce assessment information about these interactions. To produce information about teacher/students interactions using the ATOSU model, further development of the ATOSU assessment approach seems necessary to capture the nature of the student/teacher interactions. For this information to have value, the assessment act needs to improve the timing of guided feedback and needs to broaden the array of task conditions administered to students.
Assessment information gleaned from the ATOSU protocols did not provide timely information without increased investments of teacher time. To meet these demands, ATOSU information needs to be more quickly processed and analyzed so that teachers can adjust tasks and instruction as necessary. Once this information is better analyzed and regularly applied, it may be that this comprehensive approach will promote mathematical problem-solving by enhancing verbal, visual, and mathematical communication.

[A], [T], [O], and [S] attributes need to be examined over varying [U] conditions of the task so that students are required to generate thinking, and to generalize over tasks or occasions. The current study employed card tasks as a simulating device to teach concepts and procedures in probability and statistics under familiar conditions. Card tasks simulate a probability or statistics process utilized in regularly occurring and familiar conditions. If these and other mathematical concepts and procedures are to have meaning in problem-solving, future instruction needs to employ them in authentic ways (e.g., in science topics, sewing, carpentry and building, or the arts) and under unfamiliar conditions. This extension of meaning is important because it better challenges students to generate thinking and generalize concepts and procedures.

The validity of the ATOSU constructs needs to be extended with more judges. More judges should be asked to use the scales, without prior knowledge of the predicted [A], [T], [O], [S], and [U] hierarchy, to discover whether this hierarchy reproduces itself empirically. Generalizability of the simplex pattern may be confirmed using a combination of covariance structural modeling, Facets analysis, and scalogram techniques. Reproducibility of these scales is important when establishing the predicted integrity of the ordering objects along
a continuum called their “scale order” (Guttman, 1948). This order needs to reproduce itself in a generalizable fashion so that scores can be compared over occasions and tasks.

More development is necessary to clarify how any student profile organizes information in understandable and beneficial ways. Flatbed scanners could be used to record both open-ended responses and judged ratings. Software needs to be developed that facilitates the analysis of these coded ratings in a time-efficient fashion. Derived scores make it possible to compare results from one task with results on another, using quality methods to generate ideas, collect information, analyze and display data, reach consensus, and plan actions. These profiles could display scores on quality control charts or Pareto charts more as a profile rather than as single score in time. Each conceptual and procedural vector for individual tasks could be monitored at longitudinal points. However, profile interpretation must remain circumspect until methods of equating the profiled scores could be developed so that scores can be used interchangeably. Equating procedures can adjust for differences in person ability and task difficulty so that scores may be compared (Kolen & Brennan, 1995).

There are additional equity concerns that demand the examination of possible differences in class and race performance. Such differences, when produced by the assessment methods, create intentional bias that has an adverse impact on some protected groups of individuals. The data should be compared across groups that traditionally belong to such protected categories in an effort to prevent bias. For example, in cases like Goss v. Board of Education, courts have nullified tests that produced imbalanced classes of underachievers that perpetuated segregation. It is incorrect to suggest that the ATOSU
assessments are independent of culture, class, gender, and so forth. Whether these scales are suitable for use with another group depends on the group and on the test content. If the ATOSU assessment procedures are to be considered valid, the scales need to treat all groups in a fair and uniform fashion.

While alternative assessment does offer some advantages over traditional approaches, its greatest advantage may be its potential for promoting openness by publicly sharing the results with the parents and, in general, with the community at large. Using such an approach, assessment potentially acts as a bridge between the school and home provided it is frequent and timely. However, to promote openness, parents’ thinking and needs would also need to be understood and addressed. As the primary teachers and supporters of students, parents deserve to be fully involved. If one could increase parent awareness and involvement in the educational endeavor, schools are more likely to improve. Although the privacy rights of students always need to be protected, summarized assessment information needs to be distributed to those people who are responsible and regularly involved in the educational process.

Effects of parental involvement could be monitored to evaluate how such feedback contributes or detracts from any program’s goals. The on-going needs of parents might be assessed to improve the assessment and complement the efforts of teachers and researchers. Problems viewed in many of today’s schools cannot be solved without parents’ cooperation and assistance.

It could be argued that an assessment program that is continuous and that promotes parent involvement would require large investments of time and money. However, a guided
assessment may be implemented in a more cost-efficient fashion as part of the teacher training program in classroom assessment. Teachers in training would regularly develop tasks, practice skills taught in regular classes, and assess student progress in small groups using the developed scales. A profile can be produced that tracks the students’ progress as observed across ATOSU constructs. This profile can then be regularly shared with students, teachers, and parents in an effort to meet several program goals: 1) to develop a field-based program that trains teachers how to assess; 2) to offer instructional assistance to all students in small groups; 3) to involve parents in schools that have had a history of low involvement; 4) to evaluate the training program’s progress in meeting its goals; and 4) to allow school districts the freedom to evaluate and recruit interns to fill future staffing needs.

Boards of education could encourage parental involvement in such guided assessments. Administrators, teachers, and teachers in training could be offered merit pay for thinking of ways of increasing parental attendance and involvement. A parental involvement program could review assessment information and instructional videotapes. Parental meetings could be designed: 1) to identify problems and strengths in student performances; 2) to present assessment information to parents in easy-to-interpret ways; 3) to devise plans where all educators and parents might work together to promote the program’s success.

In the short run, the purposes of this research study would be fulfilled if the ATOSU model is further developed and utilized by teachers to enhance problem-solving skills, especially in mathematics, for all types of students irrespective of gender, race, or class. The ATOSU model was designed as an assessment tool to enhance teaching and learning,
inform parents, and to reward student performance. Future research could bring together university and school resources for the cherished national goal of preparing students for their prominent roles in the twenty-first century.
REFERENCES


Appendix A
ATOSU Scoring Protocol

175
Place the number of the most appropriate behavior for each of the five ATOSU general traits below the double line. Comment on the child's progress and mistake patterns.

<table>
<thead>
<tr>
<th>ATTENTION</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>The child resists trying (no effort).</td>
<td>0</td>
</tr>
<tr>
<td>The child impulsively communicates an alternative by guessing wildly (minimal effort).</td>
<td>1</td>
</tr>
<tr>
<td>The child communicates an alternative but continues down a single path without regard to progress (effort).</td>
<td>2</td>
</tr>
<tr>
<td>Choosing a valid alternative or between alternatives, the child verbally describes a viable path that becomes the center of attention until progress is thwarted, moving to another alternative if necessary until a solution is reached. (sustained effort).</td>
<td>3</td>
</tr>
<tr>
<td>Child verbally generates valid insightful and unique alternatives, attempting viable alternatives (deliberative effort).</td>
<td>4</td>
</tr>
</tbody>
</table>

Mistake Patterns

<table>
<thead>
<tr>
<th>ATTENTION</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>The child resists trying (no effort).</td>
<td>0</td>
</tr>
<tr>
<td>The child impulsively communicates an alternative by guessing wildly (minimal effort).</td>
<td>1</td>
</tr>
<tr>
<td>The child communicates an alternative but continues down a single path without regard to progress (effort).</td>
<td>2</td>
</tr>
<tr>
<td>Choosing a valid alternative or between alternatives, the child verbally describes a viable path that becomes the center of attention until progress is thwarted, moving to another alternative if necessary until a solution is reached. (sustained effort).</td>
<td>3</td>
</tr>
<tr>
<td>Child verbally generates valid insightful and unique alternatives, attempting viable alternatives (deliberative effort).</td>
<td>4</td>
</tr>
</tbody>
</table>

Mistake Patterns

<table>
<thead>
<tr>
<th>ACTICS</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>No visual patterns are expressed (no visualized tactics).</td>
<td>0</td>
</tr>
<tr>
<td>Visualized patterns are expressed with the help of the teacher (cued visual tactic).</td>
<td>1</td>
</tr>
<tr>
<td>Visualized patterns are displayed as an appropriate tactical representation that was independently derived (visual tactic).</td>
<td>2</td>
</tr>
<tr>
<td>The child validly extends a previously taught tactic(s) (modified visual tactic).</td>
<td>3</td>
</tr>
<tr>
<td>The child validly generates their own tactical representations (generative visual tactic).</td>
<td>4</td>
</tr>
</tbody>
</table>

ACTICS score

176
**[O]PERATIONS**

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>No visible operations (no operations).</td>
<td>0</td>
</tr>
<tr>
<td>The child mechanically writes a valid equation or algorithm that was prompted by the teacher (cued operations).</td>
<td>1</td>
</tr>
<tr>
<td>The child discriminates between operations to independently choose a valid set of equations (valid operations).</td>
<td>2</td>
</tr>
<tr>
<td>The child validly extends his/her operations or combines them with other operations, independently changing purposive activity (modified operations).</td>
<td>3</td>
</tr>
<tr>
<td>The child validly generates tactical and operational moves not previously taught and can verbally defend them (generative operations).</td>
<td>4</td>
</tr>
</tbody>
</table>

---

**[O]perations score**

---

**[S]OLUTION**

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>The outcome is judged incorrect (incorrect or no solution).</td>
<td>0</td>
</tr>
<tr>
<td>The student understands and constructively uses feedback to correctly solve (cued solutions).</td>
<td>1</td>
</tr>
<tr>
<td>The student independently calculates the correct answer (valid solutions).</td>
<td>2</td>
</tr>
<tr>
<td>The child checks the accuracy of a modified or generated operation by validly and independently working backwards (reversed thinking).</td>
<td>3</td>
</tr>
<tr>
<td>The child check the accuracy of modified or generated operations and comes up with an improved approach for solving the problem (critical thinking).</td>
<td>4</td>
</tr>
</tbody>
</table>

---

**[S]olution score**

---

177
No valid effort or missing response (missing conditions).

The student correctly applies operational and tactual rules in familiar contexts (near discriminating conditions).

The student correctly applies current operational and tactual rules in familiar contexts but in unfamiliar ways (far discrimination conditions).

The student connects previously learned tactics and operations to current tactics and operations in unfamiliar contexts and ways (near transfer with connections).

The student connects tactics or operations that were not previously taught to current tactics and operations in valid ways within unfamiliar contexts and ways (far transfer with connections).

---

[U]nique thinking score

---

178
### Appendix B
#### ATOSU Tasks and Goals

<table>
<thead>
<tr>
<th>Task</th>
<th>goal</th>
<th>instruction</th>
</tr>
</thead>
</table>
| #1 Probability of a diamond:  
a) using the number cards  
(ace is included)  
b) One draw | Students will understand that the probability of an event is equal to the number of ways an event can occur divided by the total number of possibilities. | [Tutored] Students are shown how to find patterns using a set of playing cards.  
[ATOSU] Students are given the cards and asked to find patterns |
| #2 Probability of a queen or a black card:  
1) using the face cards  
(ace is not included)  
2) One draw | Students can use the rule of addition to calculate the number of the probability of an event. | [Tutored] Students are presented with different sets (e.g. number cards or face cards) and are shown how to use the rule of addition.  
[ATOSU] Students are asked to discover the rule of addition using the playing cards |
| #3 Rule of Counting:  
Determine the possible ways three face cards can be combined with three number cards. | Student can apply the fundamental principle of counting to discover how many ways for an original event to occur. | [Tutored] Students are taught to categorize relevant aspects of the problem to better count the number of possibilities.  
[ATOSU] Students are asked to discover ways of counting the number of possibilities. |
| #4 Calculate the mean:  
Student is given a series of seven numbers and is asked to calculate the mean. | Students can calculate and interpret the mean, median, and mode. | [Tutored] Students are shown how to draw samples of different sets of number cards and calculate the mean, median, and mode.  
[ATOSU] Students are asked to predict the mean of the cards with samples of the cards |
Appendix C
NAEP Achievement Items
1. Dave will choose one sandwich and one drink for lunch. The menu above shows the choices. List below all the possible combinations of a sandwich and a drink that Dave might choose.

2. The nine chips shown above are placed in a sack and then mixed up. Madeline draws one chip from this sack. What is the probability that Madeline draws a chip with an even number?

- \( \frac{1}{9} \)
- \( \frac{2}{9} \)
- \( \frac{4}{9} \)
- \( \frac{1}{2} \)
- \( \frac{4}{5} \)
3. In the graph above, each dot shows the number of sit-ups and the corresponding age for one of 13 people. According to this graph, what is the median number of sit-ups for these 13 people?

☐ 15
☐ 20
☐ 45
☐ 50
☐ 55

4. The entire circle shown above represents a total of 2,675 radios sold. Of the following, which is the best approximation of the number of radios represented by the shaded sector of the circle?

☐ 70
☐ 275
☐ 985
☐ 25,880
☐ 98,420
5. The average weight of 50 prize-winning tomatoes is 2.36 pounds. What is the combined weight, in pounds, of these 50 tomatoes?

- 0.0472
- 11.8
- 52.36
- 59
- 118

6. There are 15 girls and 11 boys in a mathematics class. If a student is selected at random to run an errand, what is the probability that a boy will be selected?

- \( \frac{4}{25} \)
- \( \frac{11}{25} \)
- \( \frac{15}{25} \)
- \( \frac{11}{15} \)
- \( \frac{15}{11} \)

7. Here are the ages of five children:

13, 8, 6, 4, 4

What is the average age of these children?

- 4
- 6
- 7
- 8
- 9
- 13
- I don’t know.
8. There is only one red marble in each of the bags shown below. Without looking, you are to pick a marble out of one of the bags. Which bag would give you the greatest chance of picking the red marble? 

- Bag with 10 marbles
- Bag with 100 marbles
- Bag with 1000 marbles
- It makes no difference.
- I don't know.

9. Steve was asked to pick two marbles from a bag of yellow marbles and blue marbles. One possible result was one yellow marble first and one blue marble second. He wrote this result in the table below. List all of the other possible results that Steve could get.

<table>
<thead>
<tr>
<th>y</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>stands for one yellow marble</td>
<td>stands for one blue marble</td>
</tr>
</tbody>
</table>

GO ON TO THE NEXT PAGE
### TELEPHONE CALLING RATES

<table>
<thead>
<tr>
<th></th>
<th>Day Rate</th>
<th>Evening Rate</th>
<th>Night Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8 AM-5 PM</td>
<td>5 PM-11 PM</td>
<td>11 PM-4 AM</td>
</tr>
<tr>
<td></td>
<td>Mon-Fri</td>
<td>Mon-Fri</td>
<td>ALL DAYS</td>
</tr>
<tr>
<td></td>
<td>Sat-Sun</td>
<td>Sat-Sun</td>
<td>ALL DAYS</td>
</tr>
<tr>
<td>From Allenville</td>
<td>8 AM-5 PM</td>
<td>5 PM-11 PM</td>
<td>11 PM-4 AM</td>
</tr>
<tr>
<td>To</td>
<td>Mon-Fri</td>
<td>Mon-Fri</td>
<td>ALL DAYS</td>
</tr>
<tr>
<td></td>
<td>Sat-Sun</td>
<td>Sat-Sun</td>
<td>ALL DAYS</td>
</tr>
<tr>
<td>Burneyford</td>
<td>$.09</td>
<td>$.07</td>
<td>$.05</td>
</tr>
<tr>
<td></td>
<td>$.03</td>
<td>$.02</td>
<td>$.02</td>
</tr>
<tr>
<td>Camptown</td>
<td>$.28</td>
<td>$.22</td>
<td>$.17</td>
</tr>
<tr>
<td></td>
<td>$.09</td>
<td>$.07</td>
<td>$.05</td>
</tr>
<tr>
<td>Doming</td>
<td>$.37</td>
<td>$.30</td>
<td>$.22</td>
</tr>
<tr>
<td></td>
<td>$.11</td>
<td>$.09</td>
<td>$.07</td>
</tr>
<tr>
<td>Edgerton</td>
<td>$.42</td>
<td>$.34</td>
<td>$.25</td>
</tr>
<tr>
<td></td>
<td>$.12</td>
<td>$.10</td>
<td>$.07</td>
</tr>
</tbody>
</table>

The table above provides information about the cost of placing phone calls between certain cities at different times during the day. How much more would it cost to place a 10-minute call from Allenville to Edgerton at 3 pm on Friday than at 3 pm on Saturday?

Answer: __________

### HAIR COLOR SURVEY RESULTS

<table>
<thead>
<tr>
<th>Color of Hair</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blond</td>
<td>17</td>
</tr>
<tr>
<td>Brown</td>
<td>50</td>
</tr>
<tr>
<td>Black</td>
<td>33</td>
</tr>
<tr>
<td>Totals</td>
<td>100</td>
</tr>
</tbody>
</table>

The table above shows the results of a survey of hair color. On the circle below, make a circle graph to illustrate the data in the table. Label each part of the circle graph with the correct hair color.
12. Akira read from a book on Monday, Tuesday, and Wednesday. He read an average of 10 pages per day. Indicate in the ovals below whether each of the following is possible or not possible.

<table>
<thead>
<tr>
<th>Possible</th>
<th>Not Possible</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>4 pages</td>
<td>4 pages</td>
<td>2 pages</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>9 pages</td>
<td>10 pages</td>
<td>11 pages</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>5 pages</td>
<td>10 pages</td>
<td>15 pages</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>10 pages</td>
<td>15 pages</td>
<td>20 pages</td>
</tr>
</tbody>
</table>

**POPULATION OF CAMEO**

1900–1990

<table>
<thead>
<tr>
<th>Years</th>
<th>Population in Millions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1890</td>
<td>5</td>
</tr>
<tr>
<td>1900</td>
<td>10</td>
</tr>
<tr>
<td>1910</td>
<td>15</td>
</tr>
<tr>
<td>1920</td>
<td>20</td>
</tr>
<tr>
<td>1930</td>
<td>25</td>
</tr>
<tr>
<td>1940</td>
<td>30</td>
</tr>
<tr>
<td>1950</td>
<td>35</td>
</tr>
<tr>
<td>1960</td>
<td>40</td>
</tr>
<tr>
<td>1970</td>
<td>45</td>
</tr>
<tr>
<td>1980</td>
<td>50</td>
</tr>
<tr>
<td>1990</td>
<td>55</td>
</tr>
<tr>
<td>2000</td>
<td>60</td>
</tr>
</tbody>
</table>

13. If the population increases by the same amount from 1990 to 2000 as from 1980 to 1990, approximately what is the expected population by the year 2000?

- 0 47 million
- 0 50 million
- 0 53 million
/4/. How many boxes of oranges were picked on Thursday?

- 55
- 60
- 70
- 80
- 90
- I don't know.

/5/. On which day were more boxes of lemons picked than either boxes of oranges or boxes of grapefruit?

- Monday
- Tuesday
- Wednesday
- Thursday
- Friday
- No day
- I don't know.
16. The graph above shows the number of traffic tickets that were issued in City X each month of 1988. Approximately how many were issued during the entire year?

- 800
- 1,200
- 1,600

17. The pictograph shown above is misleading. Explain why.

Answer: ____________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________
__________________________________________________________________________
18. From a shipment of 500 batteries, a sample of 25 was selected at random and tested. If 2 batteries in the sample were found to be dead, how many dead batteries would be expected in the entire shipment?

- 10
- 20
- 30
- 40
- 50

19. The total distances covered by two runners during the first 28 minutes of a race are shown in the graph above. How long after the start of the race did one runner pass the other?

- 3 minutes
- 8 minutes
- 12 minutes
- 14 minutes
- 28 minutes
Appendix D
Item Analysis of the NAEP Achievement Items
# RESPONSE DISTRIBUTIONS

**ITEM: 1 SUBSCALE: S1**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>3</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>4</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>MID 20%</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>TOTALS</td>
<td>29</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td></td>
<td>39%</td>
<td>61%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ITEM: 2 SUBSCALE: S1**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>2</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>5</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>MID 20%</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>12</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>TOTALS</td>
<td>30</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td></td>
<td>40%</td>
<td>60%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ITEM: 3 SUBSCALE: S2**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>13</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>MID 20%</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>15</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>TOTALS</td>
<td>55</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td></td>
<td>73%</td>
<td>27%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Test and Questionnaire Analysis

**Copyright © 1984 by Small Systems Associates**

**Tinkler 98**

## Response Distributions

### Item: 4 Subscale: S3

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP  20%</td>
<td>12</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I *</td>
</tr>
<tr>
<td>- 20%</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I *</td>
</tr>
<tr>
<td>MID  20%</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I *</td>
</tr>
<tr>
<td>- 20%</td>
<td>15</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I *</td>
</tr>
<tr>
<td>BOT  20%</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I *</td>
</tr>
<tr>
<td>TOTALS</td>
<td>57</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

|    | 76% | 24% | 0% | 0% | 0% | 0% |

### Item: 5 Subscale: S2

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP  20%</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I *</td>
</tr>
<tr>
<td>- 20%</td>
<td>15</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I *</td>
</tr>
<tr>
<td>MID  20%</td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I *</td>
</tr>
<tr>
<td>- 20%</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I *</td>
</tr>
<tr>
<td>BOT  20%</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I *</td>
</tr>
<tr>
<td>TOTALS</td>
<td>63</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

|    | 84% | 16% | 0% | 0% | 0% | 0% |

### Item: 6 Subscale: S1

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP  20%</td>
<td>5</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I *</td>
</tr>
<tr>
<td>- 20%</td>
<td>8</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I *</td>
</tr>
<tr>
<td>MID  20%</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I *</td>
</tr>
<tr>
<td>- 20%</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I *</td>
</tr>
<tr>
<td>BOT  20%</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I *</td>
</tr>
<tr>
<td>TOTALS</td>
<td>44</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

|    | 59% | 41% | 0% | 0% | 0% | 0% |

---

192
# RESPONSE DISTRIBUTIONS

## ITEM: 7  SUBSCALE: S2

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>- 20%</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>MID 20%</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>- 20%</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I*</td>
<td></td>
</tr>
<tr>
<td>TOTALS</td>
<td>55</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0</td>
<td>1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td>73% 27%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## ITEM: 8  SUBSCALE: S3

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>1</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>1</td>
<td>*</td>
</tr>
<tr>
<td>- 20%</td>
<td>6</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>MID 20%</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>- 20%</td>
<td>4</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>TOTALS</td>
<td>22</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0</td>
<td>1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td>29% 71%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## ITEM: 9  SUBSCALE: S3

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>- 20%</td>
<td>8</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>MID 20%</td>
<td>7</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>- 20%</td>
<td>11</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
<td>*</td>
</tr>
<tr>
<td>TOTALS</td>
<td>43</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0</td>
<td>1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td>57% 43%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Response Distributions

**Item: 10  Subscale: S1**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I*</td>
</tr>
<tr>
<td>- 20%</td>
<td>17</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I*</td>
</tr>
<tr>
<td>MID 20%</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I*</td>
</tr>
<tr>
<td>- 20%</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I*</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I*</td>
</tr>
<tr>
<td>TOTALS</td>
<td>74</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

**Item: 11  Subscale: Total**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>7</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>7</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>MID 20%</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>6</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>TOTALS</td>
<td>26</td>
<td>49</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

**Item: 12  Subscale: S3**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>3</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I*</td>
</tr>
<tr>
<td>- 20%</td>
<td>10</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>MID 20%</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I*</td>
</tr>
<tr>
<td>- 20%</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>I*</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>I*</td>
</tr>
<tr>
<td>TOTALS</td>
<td>43</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>
TEST AND QUESTIONNAIRE ANALYSIS
COPYRIGHT (C) 1984 BY SMALL SYSTEMS ASSOCIATES
TINKLER 98
RESPONSE DISTRIBUTIONS

<table>
<thead>
<tr>
<th>ITEM: 13</th>
<th>SUBSCALE: S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMIT</td>
<td>A*</td>
</tr>
<tr>
<td>TOP 20%</td>
<td>2</td>
</tr>
<tr>
<td>- 20%</td>
<td>7</td>
</tr>
<tr>
<td>MID 20%</td>
<td>5</td>
</tr>
<tr>
<td>- 20%</td>
<td>12</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>6</td>
</tr>
<tr>
<td>TOTALS</td>
<td>33</td>
</tr>
<tr>
<td>P</td>
<td>44%</td>
</tr>
</tbody>
</table>

ITEM: 14  SUBSCALE: S3

<table>
<thead>
<tr>
<th>ITEM: 14</th>
<th>SUBSCALE: S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMIT</td>
<td>A*</td>
</tr>
<tr>
<td>TOP 20%</td>
<td>6</td>
</tr>
<tr>
<td>- 20%</td>
<td>7</td>
</tr>
<tr>
<td>MID 20%</td>
<td>8</td>
</tr>
<tr>
<td>- 20%</td>
<td>9</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>11</td>
</tr>
<tr>
<td>TOTALS</td>
<td>41</td>
</tr>
<tr>
<td>P</td>
<td>55%</td>
</tr>
</tbody>
</table>

ITEM: 15  SUBSCALE: TOTAL

<table>
<thead>
<tr>
<th>ITEM: 15</th>
<th>SUBSCALE: TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMIT</td>
<td>A</td>
</tr>
<tr>
<td>TOP 20%</td>
<td>8</td>
</tr>
<tr>
<td>- 20%</td>
<td>10</td>
</tr>
<tr>
<td>MID 20%</td>
<td>3</td>
</tr>
<tr>
<td>- 20%</td>
<td>3</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>3</td>
</tr>
<tr>
<td>TOTALS</td>
<td>27</td>
</tr>
<tr>
<td>P</td>
<td>36%</td>
</tr>
</tbody>
</table>

195
### RESPONSE DISTRIBUTIONS

**ITEM: 16 SUBSCALE: S1**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>5</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>9</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17 I</td>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>MID 20%</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>15</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>TOTALS</td>
<td>44</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

**ITEM: 17 SUBSCALE: S1**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>2</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>6</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17 I</td>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>MID 20%</td>
<td>2</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>7</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>TOTALS</td>
<td>23</td>
<td>52</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>

**ITEM: 18 SUBSCALE: S1**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A*</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>3</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>5</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17 I</td>
<td>17</td>
<td>I</td>
</tr>
<tr>
<td>MID 20%</td>
<td>5</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>- 20%</td>
<td>13</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td>18</td>
<td>I</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td>11</td>
<td>I</td>
</tr>
<tr>
<td>TOTALS</td>
<td>35</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td>75</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
</tbody>
</table>
TEST AND QUESTIONNAIRE ANALYSIS  
COPYRIGHT (C) 1984 BY SMALL SYSTEMS ASSOCIATES  

TINKLER 98  

RESPONSE DISTRIBUTIONS  

<table>
<thead>
<tr>
<th>ITEM: 19</th>
<th>SUBSCALE: S2</th>
<th>GROUP</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OMIT</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>TOP 20%</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- 20%</td>
<td>6</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>MID 20%</td>
<td>6</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>- 20%</td>
<td>14</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BOT 20%</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TOTALS</td>
<td>43</td>
<td>32</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>57%</td>
<td>43%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

ITEM: 20  SUBSCALE: S3  

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>16</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td></td>
</tr>
<tr>
<td>- 20%</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17 I</td>
<td></td>
</tr>
<tr>
<td>MID 20%</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td></td>
</tr>
<tr>
<td>- 20%</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td></td>
</tr>
<tr>
<td>BOT 20%</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td></td>
</tr>
<tr>
<td>TOTALS</td>
<td>72</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75 0 1 2 3 4 5 6 7 8 9</td>
<td></td>
</tr>
<tr>
<td>96%</td>
<td>4%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ITEM: 21  SUBSCALE: TOTAL  

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td></td>
</tr>
<tr>
<td>- 20%</td>
<td>14</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17 I</td>
<td></td>
</tr>
<tr>
<td>MID 20%</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td></td>
</tr>
<tr>
<td>- 20%</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td></td>
</tr>
<tr>
<td>BOT 20%</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td></td>
</tr>
<tr>
<td>TOTALS</td>
<td>67</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75 0 1 2 3 4 5 6 7 8 9</td>
<td></td>
</tr>
<tr>
<td>89%</td>
<td>11%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

197
### RESPONSE DISTRIBUTIONS

**ITEM: 22  SUBSCALE: S3**

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MID 20%</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BGT 20%</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTALS</td>
<td>54</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OMIT</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>SIZE</th>
<th>P - VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOP 20%</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18 I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MID 20%</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BGT 20%</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11 I</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTALS</td>
<td>54</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>75</td>
<td></td>
</tr>
</tbody>
</table>

68% 72% 0% 0% 0% 0%