

**SAMPLING VARIABILITY OF SELECTED PERFORMANCE  
ASSESSMENTS MEASURING AN EIGHTH GRADE  
MATHEMATICS DOMAIN**

**By**

**MARIDYTH M. MCBEE**

**Bachelor of Arts  
Oklahoma State University  
Stillwater, Oklahoma  
1975**

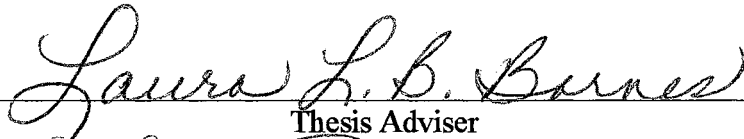
**Master of Science  
Oklahoma State University  
Stillwater, Oklahoma  
1977**

**Submitted to the Faculty of the  
Graduate College of the  
Oklahoma State University  
in partial fulfillment of  
the requirements for  
the Degree of  
DOCTOR OF PHILOSOPHY  
May, 1995**

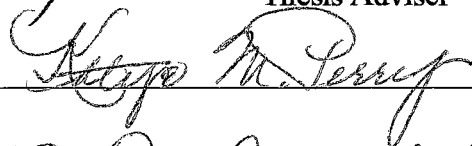
Thesis  
1995D  
M112s

SAMPLING VARIABILITY OF SELECTED PERFORMANCE  
ASSESSMENTS MEASURING AN EIGHTH GRADE  
MATHEMATICS DOMAIN

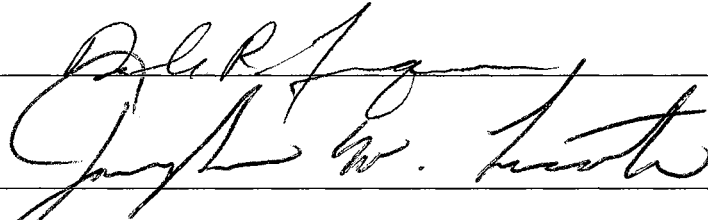
Thesis Approved:

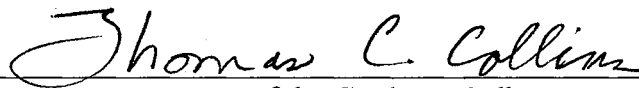
  
Laura L. B. Burnes

Thesis Adviser

  
Keryn M. Perry

  
N. Jo Campbell

  
Julie A. [unclear]

  
Thomas C. Collins

Dean of the Graduate College

## ACKNOWLEDGMENTS

I wish to express heartfelt gratitude to my dissertation advisor, Dr. Laura Barnes, and to my doctoral committee chair, Dr. Kayte Perry. Their teaching and guidance throughout my degree program were invaluable and truly appreciated. Thanks to Dr. Jo Campbell, longtime mentor and friend, for her constant support. Thanks to Dr. Dale Fuqua for his direction and for being such an exemplary role model of an administrator. Thanks to Dr. Joe Licata for showing me a new way of looking at the educational research process.

The support of such wonderful friends as Pat Lowther, Ann Cantrell, Susie Parker, and Kim Apel-Morrow were of inestimable value. Thanks for proofreading, typing and rating the student responses. Thanks also for your encouragement and good will throughout the dissertation process.

I also wish to express thanks to my children, Brandon McBee and Ashley McBee, for the sacrifices they make on behalf of their mother's career. Deepest appreciation and regard go to my husband, Dr. John McBee, for his constant love and support throughout all my endeavors.

## TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION. ....	1
Description of Performance Assessments .....	2
Present Role of Performance Assessments .....	4
Issues in Performance Assessments .....	5
Statement of the Problem .....	9
Purpose of the Study .....	10
Significance of the Study .....	11
Definition of Terms .....	12
Research Questions .....	13
Assumptions .....	14
Limitations .....	15
II. REVIEW OF THE LITERATURE. ....	16
Rationale for Mathematics Performance Assessments .....	16
Characteristics of Multiple Choice Assessments .....	21
Characteristics of Performance Assessments .....	22
Technical Qualities of Performance Assessments .....	24
Technical Qualities of Performance Assessments in Mathematics .....	34
Impact of Performance Assessment Results On Gender and Ethnic Groups .....	38
Summary .....	40
III. METHOD AND PROCEDURES. ....	42
Subjects .....	42
Instrumentation .....	44
Guidelines for Instrument Development .....	44
Development of the Mathematics Problem-Solving .....	47
Data Collection Procedures .....	50

Chapter	Page
Scoring the Performance Assessment Tasks . . . . .	51
Data Analysis . . . . .	52
<b>IV. RESULTS. . . . .</b>	<b>56</b>
Descriptive Data . . . . .	56
Classical Reliability . . . . .	56
Generalizability Studies . . . . .	60
Decision Studies . . . . .	65
Impact of the Order of Task Presentation . . . . .	68
Performance by Gender and Ethnicity . . . . .	72
Summary . . . . .	76
<b>V. SUMMARY AND CONCLUSIONS. . . . .</b>	<b>78</b>
Summary of the Investigation . . . . .	78
Discussion . . . . .	79
Sources of Score Variation. . . . .	79
Number of Raters, Occasions, and Tasks Necessary to Generalize. . . . .	82
Performance by Gender and Ethnicity. . . . .	84
Conclusions . . . . .	85
Recommendations . . . . .	88
<b>REFERENCES. . . . .</b>	<b>89</b>
<b>APPENDICES. . . . .</b>	<b>102</b>
<b>A-PERFORMANCE TASKS ON THE MATHEMATICS     PROBLEM-SOLVING ASSESSMENT. . . . .</b>	<b>103</b>
<b>B- SCORING RUBRICS FOR THE MATHEMATICS     PROBLEM-SOVING ASSESSMENT. . . . .</b>	<b>107</b>

## LIST OF TABLES

Table	Page
I. MPSA Scores by Rater, Task Occasion.....	57
II. MPSA Scores by Task and Occasion.....	57
III. Frequency of Scores Received on the Basketball Camp Task for Each Testing Occasion.....	58
IV. Frequency of Scores Received on the Space Camp Task for Each Testing Occasion.....	58
V. Frequency of Scores Received on the Olympics Task for Each Testing Occasion.....	59
VI. Frequency of Scores Received on the Tug of War Task for Each Testing Occasion.....	59
VII. Variance Components for the Person by Rater by Task Model for Each Testing Occasion.....	61
VIII. Variance Components Using the Full Model.....	63
IX. Variance Components for the Full Model Using Only the Two Parallel Tasks of Basketball Camp and Space Camp.....	64
X. Variance Components for the Two Highly Similar Tasks of Basketball Camp and Space Camp on the First Testing Occasion Only.....	66
XI. Variance Components for the Two Highly Similar Tasks of Basketball Camp and Space Camp on the Second Testing Occasion Only.....	66
XII. Generalizability Coefficients for the Person by Task by Rater by Occasion Decision Study When All Four Tasks Are Included.....	67

Table	Page
XIII. Dependability Coefficients for the Person by Task by Rater by Occasion Decision Study When All Four Tasks Are Included.....	67
XIV. Generalizability Coefficients for the Person by Task by Rater by Occasion Decision Study based on the Two Highly Similar Tasks.....	69
XV. Dependability Coefficients for the Person by Task by Rater by Occasion Decision Study based on the Two Highly Similar Tasks.....	69
XVI. Order of Presentation by Task Split Plot Design for the First Testing Occasion.....	71
XVII. Order of Presentation by Task Split Plot Design for the Second Testing Occasion.....	71
XVII MPSA Scores by Ethnicity .....	73
XIX. Task by Occasion by Ethnicity Analysis.....	73
XX. Results of Dependent T-Test as Follow-up to Significant Task Main Effect .....	74
XXI. MPSA Scores by Gender .....	75
XXII. Task by Occasion by Gender Analysis.....	75
XXII. Independent T-Test As Follow Up to the Gender by Task Interaction.....	76



## CHAPTER 1

### INTRODUCTION

Dissatisfaction with current large-scale methods of assessing student mathematics achievement is evident in the literature. Criticism comes from mathematics educators (Working Group of the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics, 1989), educational researchers (Putnam, Lampert, & Peterson, 1990; Shepard; 1989, Smith 1991;), psychometricians (Bennett, 1993; Braswell & Kupin, 1993; Ebel, 1984; Gitomer, 1993; Masters & Mislevy, 1993; Tatsuoka, 1993), and cognitive psychologists (Royer, Ciscero, & Carlo, 1993; Snow & Lohman, 1993). Much of the complaint centers around the use of multiple-choice tests as the sole measure of mathematics achievement. Critics suggest that the format of multiple-choice tests does not best reflect the most current knowledge concerning the way students learn mathematics. According to modern theorists (e.g., Putnam et al.), learning mathematics is no longer seen as occurring in a linear fashion where students acquire more and more facts. Instead, learning is viewed as a constructive process in which new knowledge is not simply added, but is integrated into existing structures, or causes structures to be reconfigured (Bennett). In contrast to multiple-choice testing which is primarily product/output oriented, the optimal testing format presents an ill-structured problem in a somewhat novel situation, and measures not only the final outcome, but also the process used by the examinee to arrive at the answer (Baxter, 1992). As mathematics educators

seek to reform mathematics curriculum to more closely align with research on cognition, the format of tests used to measure the success of the students and teachers needs to reflect what is being taught. The interpretation of test results becomes extremely tenuous when students are taught via one method such as "pursuing open-ended problems and extended problems-solving projects" (Working Group of the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics, 1989, p. 70), and tested by a measure which requires simple recall of mathematics facts or recognition of the correct solution. The introduction of assessment methods that attempt to address these concerns has been a fairly recent development in large-scale testing programs of mathematics. Typically called performance assessments, these methods differ from traditional multiple-choice tests in the nature of the problems presented and that they require students to demonstrate the processes used in determining their answers.

#### Description of Performance Assessments

There is general agreement on the psychometric definition of a performance assessment as a constructed response to a stimulus which is evaluated by professional judgment (Archibald & Newmann, 1988; Aschbacher, 1991; Bennett, 1993; Bock, 1991; Lane, 1989; Mehrens, 1991; Phillips, 1993; Stiggins, 1987). The performance assessments referred to in this study include complex constructed response items. The process used to obtain the answer is evaluated as well as the final answer. The items require the examinee to perform multiple steps to obtain the answer. In addition, the scoring decisions cannot typically be made immediately and ambiguously using a scoring key but require some degree of expert judgment. An example of a performance

assessment task is "Lee has 36 feet of fence to use to make a backyard pen for his dog. If Lee wants the perimeter of the dog pen to be 36 feet, what are all possible dimensions for the pen if the dimensions are expressed in whole feet? Show how you arrived at your answer." The answer to the task would reveal whether or not the student knew the appropriate mathematical procedure to solve the problem, the process used to determine the answer, as well as whether or not the student could state the correct answer. If the student was asked a question such as: "Select the dimensions below that would result in a perimeter of 36 feet." in a multiple-choice format, and was given four options from which to choose, the selection of an option would only provide information regarding whether or not the student could recognize the right answer. All information concerning the processes used to determine the answer need to be inferred. Thus, the multiple-choice question is an indirect measure of the student's ability to solve the problem presented as the performance assessment.

The technology of test development has primarily followed an agenda independent of cognitive psychology and curriculum reform (Baxter, 1992). Multiple-choice tests have been the format used most frequently for large-scale testing occasions due to their cost efficiency, ease of administration, and ease in determining reliability and validity. Sophisticated techniques for developing multiple-choice test items have been developed, based on the specification of well-defined learning outcomes (Roid & Haladyna, 1982). Currently, some mathematics educators (Working Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics, 1989) assert that the specification of discrete learning outcomes and the assessment of

those outcomes has resulted in a fragmented mathematics curriculum. Students are not being taught to integrate mathematics knowledge for problem-solving applications because integration of mathematics topics is not being tested.

### Present Role of Performance Assessments

Performance assessments are making their way into the accountability measures used by state departments of education. Aschbacher (1991) surveyed state testing directors to determine the use of performance assessments on a state by state basis. Nearly three-fourths of the states used a form of direct writing assessment, in which students are asked to write an actual essay pertaining to a given topic. This contrasts with indirect measures of writing, in which students are asked questions about writing, such as finding errors in sentences or identifying misspelled words. In addition, about half the states were conducting, or considering conducting, alternative assessments beyond direct writing assessment. Most interest was directed toward developing performance assessments in science and mathematics. The Office of Technology Assessment (1992) found that seven states (Arizona, California, Connecticut, Kentucky, Maryland, New York, and Vermont) were moving toward performance assessments as the primary statewide achievement measure. Response of the states to the survey indicated that the change was precipitated by the discovery that standardized, norm-referenced tests do not test a sufficient percentage of the essential skills identified for public school students. For example, Arizona's State Department of Education staff discovered that the Iowa Test of Basic Skills and the Tests of Academic Proficiency tested only 20 to 40% of the essential skills in reading, writing, and mathematics for their state (Office of Technical Assessment,

1992). If the trend continues, performance assessments outcomes will be used routinely as accountability measures of educational outcomes.

### Issues in Performance Assessments

Much that has been written about performance assessments has been in the context of high stakes testing. High stakes testing situations occur when the outcomes of the assessment will result in certification or selection decisions about teachers or students, or used in the determination of school system accountability (Baker, O'Neil, & Linn, 1993). Examples of high stakes testing situations include those in which students must pass the assessment before they are allowed to attend kindergarten, are promoted to the next grade, or receive their high school diploma. Other high stakes tests identify which schools or school districts are achieving at a specified standard. If the students score below the standard, schools may be put under pressure to raise scores, experience withdrawal of support, receive a new principal, and/or experience a large turnover of teaching staff. For example, one elementary school which failed for two years to test at the standard specified by the Oklahoma State Department of Education, experienced over 50% turnover of faculty for each of the two years (W. Edwards, personal communication, October 15, 1992).

As tests are increasingly used to make high stakes decisions about individuals or educational institutions, the importance of using the most appropriate assessment measures is underscored. Teachers find their instruction must emphasize what is being tested. If the target of assessment is discrete pieces of knowledge, the instruction will be towards discrete information training (Hartle & Battaglia, 1993; McNeil, 1988; Smith,

1991). Frederiksen and Collins (1989) argue for developing tests that foster the kinds of teaching and learning which most directly further the goals of education. Tests that influence teachers to instruct students in the ways deemed by experts as most appropriate are labeled by Frederiksen and Collins as systemically valid. Assessment activities that are systemically valid are worthwhile activities in and of themselves. Performance assessments are purported to have higher systemic validity than multiple-choice tests because the former is a direct assessment of an intended outcome as opposed to an indirect or proxy measure. Some educators write that as teachers teach the concepts measured by systemically valid tests, appropriate instruction will occur and student achievement scores will truly reflect their progress toward reaching the most appropriate educational goals (Baker & Herman, 1983; McNeil, 1988; Resnick & Resnick, 1992; Shavelson, Baxter, & Pine, 1991; Wiggins, 1989). Others such as Shepard (1989) and Mehrens (1991) concur that assessment tasks need to be redesigned to more closely resemble real work applications, but warn that even performance assessment tests are susceptible to distortion if scores are used for accountability or other high stakes decisions. To date, empirical evidence is lacking regarding whether or not systemically valid tests improve instruction.

Although performance assessments are promoted by some as the solution to problems with multiple-choice tests, they are not a panacea. One problem associated with performance assessments is the high cost of the procedure. Mehrens furthermore points to the problems of test security in high stakes environments. As there are limited numbers of questions or prompts to which to respond, those taking the test can easily remember the

test content. Those taking the exam on subsequent occasions will not have identical testing conditions if they have been told the content of the test. Unfortunately, if the stakes of testing are high, the temptation of teachers or students to share the content of the test will also be high. To date, evidence is not available to demonstrate that the use of performance assessment eliminates the problem of teaching to the test. To address this problem, frequent changes of the test content must be made. The combined cost of constant test development and expensive scoring procedures cause performance assessments to be quite costly as compared to multiple choice tests.

Phillips (1993) writes that performance assessments frequently do not have sufficient evidence of reliability or validity to withstand legal challenges in high stakes decision arenas. In most cases, high inter-rater reliability is attainable with extensive rater training and specific scoring rubrics (Office of Technology Assessment, 1992). While inter-rater reliability can be adequately high, evidence for other forms of reliability and validity are more difficult to establish (Aschbacher, 1991; Dunbar, Koretz & Hoover, 1991; Linn, Baker & Dunbar, 1991; Quellmalz & Capell, 1979; Shavelson et al., 1991; Shavelson, Mayberry, Li, & Webb, 1990; Swanson, Norcini, & Grasso, 1987).

Performance assessment tasks take longer to complete than multiple choice test items. Therefore, fewer tasks can be included in each assessment and each task has relatively greater weight in influencing students' scores. If individual student performance differs from task to task or if the overall performance differs from task to task, then inferences from task performance to performance in the domain of interest is threatened. Studies involving performance assessments in writing (Cantor & Hoover, 1986; Dunbar et al.,

1991; Hieronymus & Hoover, 1987), science (Shavelson, Baxter & Gao, 1993), bar exams (Baker, O'Neil, & Linn, 1992), as well as in mathematics (Baxter, Shavelson, Herman, Brown, & Valadex, 1993; Lane, Stone, Ankenmann, & Liu, 1992; Shavelson et al., 1993) indicate that performance differs substantially by task. Evidence of the technical qualities of performance assessments in mathematics relies extensively on content validity, and inter-rater reliability as opposed to inter-task reliability (Johnson, Mazzeo, & Kline, 1993). More information is needed to determine how many tasks need to be included on mathematics performance assessments for the results to be generalized to the mathematics construct assessed.

Another issue which must be addressed is the fairness of test results to all ethnic and gender groups. Linn et al. (1991) and Miller and Legg (1993) report that studies addressing differential performance by racial groups lead to the conclusion that performance assessments will not likely eliminate the achievement gaps present in multiple-choice achievement tests. In fact, Shepard (1989) points out that any group for whom instruction has emphasized memorization of basic facts will be expected to perform poorly on performance assessments in comparison to students who have been instructed in higher-order thinking skills. A common response of teachers and administrators whose students have failed to perform well on multiple-choice, standardized achievement tests is to increase the drill and practice of basic skills. Schools being pressured to raise standardized test scores often serve lower socioeconomic students and are frequently overrepresented by minority students. Students are dependent on the extent to which their teachers can embrace the teaching methods recommended by groups such as the Working



Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics (1989) without fear of retribution due to low test scores on accountability measures. Linn et al., (1991) assert that teachers must be provided with training and other support to move in the new direction of instruction. However, this training is not being supplied uniformly. Given the push for accountability as demonstrated by standardized achievement test scores, school districts who serve disadvantaged students may invest teacher training funds in other ways than toward workshops related to the instruction best measured by performance assessment (Baker et al., 1993).

#### Statement of the Problem

Performance assessments are proposed as a means of obtaining more valid information about students' mathematical skills than traditional assessments by requiring students to demonstrate and describe the problem-solving processes they use to apply mathematical concepts in authentic situations. However, relatively little is known about the impact that the limited number of tasks found on performance assessments has on the ability to make valid and reliable inferences from task performance to performance in the mathematics domain of interest. Research on the use of performance assessments in mathematics is limited, but tends to show that these tests have low inter-task reliability (Baxter, et al., 1993, Lane, et al., 1992; Shavelson, et al., 1993). As yet, it is unknown whether the low inter-task reliability is caused by the fact that with typically few tasks, each task must measure a broad domain composed of many non-overlapping objectives, or whether the nature of this type of assessment simply introduces more construct irrelevant

variability. It may be anticipated that tasks designed to be highly similar would yield lower inter-task variability than less similar tasks. To date no published research has examined this issue. A second issue is the stability of mathematics performance assessments over time and how task similarity may affect stability. Even though students are typically tested on one occasion, the test results are generalized across occasions. Research conducted by Ruiz-Primo, Baxter, and Shavelson (1993) on the stability of performance assessments measuring a sixth grade science domain indicate scores across time were only moderately stable. To date, research is lacking regarding the stability of mathematics performance assessments. Another concern is the impact of the use of performance assessment tests on females and ethnic minorities. Studies conducted thus far do not indicate that performance assessments result in improved performance for ethnic minorities (Baxter, et al., 1993; Linn et al., 1991; Miller & Legg, 1993; Office of Technology Assessment, 1992). Research on the effects of performance assessment on gender is limited, but seems to indicate that females do somewhat better on written essays than on multiple-choice tests (Office of Technology Assessment). More research is needed to determine if this trend is found in mathematics performance assessment tasks.

#### Purpose of the Study

The purpose of this study is to evaluate the psychometric properties of an eighth grade mathematics performance assessment. Specifically, this study examines the temporal stability and inter-task reliability of these tasks, and seeks to address the issue of how task similarity affects the ability to generalize results of the assessments. The study also examines differences in task performance by gender and ethnicity. Four performance

tasks, two being highly similar, were developed to measure a complex problem-solving domain from the Oklahoma state learner outcomes for eighth grade mathematics and were administered to 101 eighth graders in one Oklahoma middle school.

### Significance of the Study

Across the nation, mathematics performance assessment tasks are included in a number of statewide assessment programs (Office of Technology Assessment, 1992). In addition, the National Assessment of Educational Progress includes performance assessments to measure mathematics concepts at grades four and eight. Innovative programs to teach mathematics such as QUASAR (Quantitative Understanding: Amplifying Student Achievement and Reasoning) are being evaluated with performance assessment tasks (Lane et al., 1992). Inter-rater reliability is frequently reported to show the stability of scores between raters. However, limited information is published regarding the effects of limited task sampling and the stability of measures over time.

Three studies have addressed the effect of limited task sampling of mathematics performance assessments given to sixth and seventh graders (Baxter et al., 1993; Lane et al., 1992; Shavelson et al., 1993). In each of these studies, a large person by task interaction was found indicating differential ranking of student performance depending on the task given. The present study will investigate the issue with performance assessment tasks given to eighth grade students. Information will also be provided about the stability of performance over time. The results of the study can be used to evaluate the number of tasks, raters, and occasions that are likely necessary to obtain scores from mathematics performance assessments which can be generalized to the domain of interest. As the tasks

used for this study are selected to measure the mandated curriculum for Oklahoma eighth graders, the study will have direct application for the use of performance assessments for the State of Oklahoma. However, as the Oklahoma mandated curriculum closely follows the guidelines outlined by the Working Groups for the Commission on Standards for School Mathematics of the National Council of the Teachers of Mathematics (1989), the results of this study will easily transfer to other settings.

### Definition of Terms

#### Generalizability Theory

Generalizability (G) theory (Brennen, 1982; Shavelson & Webb, 1991) allows expansion of the classical reliability, in that multiple sources of error can be estimated simultaneously in the same analysis. Instead of the observed score being decomposed into only true score variance and error variance, the error variance can be further broken down into potential sources such as task error, occasion error, rater error, as well as random error component in examinee scores. By providing a statistical theory that estimates how much each component of the sample measurement (e.g., tasks, rater, etc.) contributes to measurement error (including both systematic and random), G theory establishes a basis for evaluating how well each assessment can be substituted for other assessments and thus reliably and validly represent the domain of interest.

#### Performance Assessment

The term performance assessments used in this study refers to complex constructed response items (Archibald & Newmann, 1988; Aschbacher, 1991; Bennett, 1993; Bock, 1991; Mehrens, 1991; Phillips, 1993; Stiggins, 1987). Simple constructed

response items such as sentence completion items are not included. Further stipulations are that the process used to obtain the answer is evaluated as well as the final answer; the items require the examinee to perform multiple steps to obtain the answer; and the scoring decisions cannot be made immediately and unambiguously using a scoring key but require some degree of expert judgment.

### Raters

Raters determine examinee's scores on performance assessments. They must have some professional expertise in the subject matter domain being assessed and be trained to reliably use the scoring criteria.

### Tasks

Mathematics performance assessment items are referred to as tasks in this study. Each task presents a situation in which the examinee demonstrates process knowledge as well as the ability to reach a solution.

### Research Questions

The following research questions will be answered from the analysis of the four performance assessment tasks which represent a complex problem-solving domain from the Oklahoma state learner outcomes for eighth grade mathematics. G theory (Brennen, 1982; Kane, 1982; Shavelson & Webb, 1991) will be used for the analysis of the first four questions. Split plot ANOVA's will be used to answer the fifth question.

Question One: How much of the variance in students' scores on the mathematics performance assessment tasks can be attributed to universe score (true score) variance and how much can be attributed to error variance?

Question Two: What are the relative impacts of the following potential sources of error variance on the mathematics performance assessment: raters, occasions, and tasks?

Question Three: What effect does task similarity have on inter-task variability and other potential sources of measurement error?

Question Four: How many raters, occasions, and tasks are necessary to achieve acceptable levels of generalizability (reliability) for the mathematics performance assessment?

Question Five: Do students scores on mathematics performance assessments differ by gender or ethnic groups?

#### Assumptions

The following assumptions are required to accomplish this study. First, the performance of students in the sample is similar to the performance of students of similar background. Second, the students serving as subjects in the study were taught by a method that is appropriately assessed through performance assessments. The mathematics curriculum consultant for the district from which the sample is taken, indicated that the teachers of the students in the study are trying out ideas which reflect the recommendations for mathematics curriculum reform (L. Bailey, personal communication, December 1, 1993). However, they have not as yet completely revised their curriculum. Therefore, the students in the sample have experienced some of the new approaches to mathematics instruction, as well as some of the more traditional types of mathematics classroom activities. While it is hoped by both the mathematics curriculum consultant and the students' teachers that the classroom instruction would allow all students to solve

mathematics problems, instruction aimed at this end alone was not the sole focus of the curriculum.

### Limitations

The results of this study will address only eighth graders assessed in a domain taken from the subject area of mathematics. Information regarding other age groups or subject area domains will not be obtained. In addition, the students in the sample come from an affluent suburban middle school. As a result, no information will be obtained to determine whether or not the information obtained by this study can be generalized to other types of students.

A high stakes testing condition was absent for the students participating in the study. The students' teacher and the researcher stressed the importance of the assessment, and students were encouraged to do their best work. However, students' levels of motivation to perform to the best of their ability is unknown.

Although care has been taken to follow all recommendations regarding the development of performance assessments, these results can not generalize to all performance assessments in the mathematics domain. Also, there may be uncontrolled sources of variability which are not specified in the model used for this study (e.g., difference in administration conditions) that may limit generalization to other settings.

## CHAPTER II

### REVIEW OF THE LITERATURE

The review of the literature that follows provides a rationale for the use of mathematics performance assessments and discusses research that directly examines the psychometric properties of performance assessments. Because much of the research on performance assessments comes from the subject matter domain of direct writing assessment, evidence regarding the reliability and validity of direct writing assessments is provided. Next, evidence of the reliability and validity of mathematics performance assessments is reviewed. Finally, the effect of performance assessment results on gender and ethnic groups is discussed.

#### Rationale for Mathematics Performance Assessments

The impetus for using performance assessments comes from the call to reform mathematics curriculum. The reform efforts utilize the knowledge acquired through cognitive psychology in order to better meet the goals for a mathematically literate workforce. Robinson (1993) reports that simple mathematics literacy is not a concern of high priority for employers. New employees are able to read and perform simple computation tasks. Instead, the most debilitating weakness of beginning employees is the inability to process the information required to accomplish moderately complex mathematical tasks encountered in the work setting. The Information Age, with the accompanying technological advances, has changed what workers need to know to function productively. In the past, the goal of mathematics educators was to ensure that



all students graduated from high school with basic computation skills. Complex mathematical reasoning and problem-solving was reserved for the brightest students. The challenge for the present is to educate all students to be able to use mathematics to solve complex problems (Putnam et al., 1990). The current economic context requires that workers in all job categories be able to recognize problems amenable to mathematical solutions, analyze the potential options for solutions, and communicate findings (Robinson, 1993; Working Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics, 1989).

In response, the Working Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics recommends that all students need to learn more mathematics. In order for this to happen, the Working Group recommends that instruction be revised to meet the requirements of the kinds of mathematics that students need to know today. In the past, minimum competencies in mathematics were provided universally, and only the most talented were instructed in advanced mathematics applications. However, the shift from the Industrial Age to the Information Age changed what the typical student needs to know about mathematics. For example, technological advances such as low-cost calculators and computers have changed how mathematics is used in the business setting.

The curriculum addressed for reform by the Working Group (1989) includes mathematical problem-solving, communication, reasoning, connections, number/operations/computation, patterns and functions, algebra, statistics, probability, geometry, and measurement. In the problem-solving area, increased attention is proposed

for "pursuing open-ended problems and extended problems-solving projects, investigating and formulating questions from problem situations, and representing situations verbally, numerically, graphically, geometrically or symbolically" (p. 70). Decreased attention is recommended for "practicing routine, one-step problems, and practicing problems categorized by types (e.g., coin problems, age problems)" (p. 71).

The reform measure responding to the need for universal mathematics competencies follow the most up to date view of learning. During the last two decades, the field of psychology has shifted from an emphasis on behaviorism to the study of cognition (Putnam et al., 1990). The product of cognitive research is a more complex view of the way people learn than is the stimulus/response model of behaviorism. Current models of learning propose that learners gain understanding when they construct their own knowledge and determine their own interconnections among concepts and facts. As new information is connected to existing information, knowledge is restructured with new and more complex knowledge organizations replacing old concepts. The updated knowledge structures are adapted further as they are tried in new settings (Mislevy, 1993; Shepard, 1989; Snow & Lohman, 1989).

Learning is no longer viewed as something which occurs in isolated pieces, with the accompanying need to teach one skill at a time. Instead, learners increase competence not by simply accumulating new facts, but by reconfiguring their knowledge structures. According to Wolf, Bixby, Glenn, and Gardner (1991), the differences between experts and novices are not as much in the knowledge they possess, as in their ability to organize that information to create a product or accomplish a task. The difference between

beginning and advanced learners is not only one of discrepancies in factual knowledge, but also variations in the types of conceptions and understanding that students bring to a problem situation, and in the strategies they use to find the answer to the problem. Rather than being wrong, novice learners frequently indicate a partial understanding (Masters & Mislevy, 1993). A study conducted by Carpenter and Moser (1984) illustrates this point. They found that most children in grades 1-3 could provide the correct answer to an addition question with single digits such as  $6 + 8 = ?$ . However, the children used a variety of strategies to come up with the answer. Some children counted 6 and 8 objects. Others began counting 8 more objects from 6. Others began counting 6 more objects from 8. Still others knew the correct answer because they had memorized the math fact. So, to be most useful, a test of addition would need to record the strategy used to determine the answer, as well as whether or not the student arrived at the correct answer.

The recent developments in cognitive theory suggest that achievement tests need to measure several important aspects of performance, such as the processes underlying problem-solving and the strategies students use. Snow and Lohman (1993), as well as others (Bennett, 1993; Gitomer, 1993; Glaser, 1988; Resnick & Klopfer, 1989; Tatsuoka, 1993), call for educational assessments that measure cognitive processes in addition to terminal performance. As educators respond to these societal requirements by reforming curriculum to more closely reflect both the changing needs of society and modern learning theory, measures of achievement need to reflect what is taught. Many agree that tests composed of multiple-choice questions alone are inadequate measures of the most

significant outcomes of learning (Bennett, 1993; Bennet, Rock, Braun, Frye & Spohrer, 1990; Gitomer, 1993; Robinson, 1993).

In order to prepare students to be mathematically literate, the Working Group (1989) calls for students to be presented with problem situations in which mathematics can be utilized. They emphasize that learning mathematics is applying mathematics, as opposed to merely mastering mathematical concepts and procedures. Equating knowing mathematics skills to knowing mathematics, is compared to equating proficiency in music scales to proficiency in playing a song, or similarly that knowing the moves of individual chess pieces is equivalent to competency in playing the game of chess (Bennett, 1993). The Working Group (1989) recommends that classroom activities be based on the solution of problems. For example, a true-to-life problem situation that requires certain mathematical skills should be presented to the class before possible solutions are taught. The skills then have a context for application. The usual pre-reform strategy is to drill students in various computational skills before application problem situations are, if ever, presented.

Attainment of mathematical competencies as reflected in the recommendations of the Working Group (1989) cannot be comprehensively measured by conventional multiple-choice format tests alone (Bennett, 1993). If complex mathematical problem-solving is the instructional target, then direct measures need to be used for assessment. Otherwise, inappropriate conclusions regarding the results of instruction may be made, such as concluding that an examinee is proficient in playing chess if the results of assessment only indicated that the examinee knows the moves associated with each chess

piece. Performance assessments are one vehicle for providing a direct measure of the mathematical competencies identified by the Working Group (1989) in their Curriculum and Evaluation Standards for School Mathematics.

### Characteristics of Multiple Choice Assessments

The optimal multiple-choice items are unambiguous, tightly structured, and require limited knowledge beyond that needed to solve the problem (Gitomer, 1993). However, most problem-solving outside of the school setting (e.g., the workplace) does not fit the conditions for good multiple choice items. In the "real world", the critical skill is to be able to identify a structure for solving the problem. Chase and Simon (1973) assert that the majority of the problem-solving task is determining the structure of the problem, and only a small part of the task is solving the problem once it is structured. Outside of the school setting, people are introduced to situations that often are ambiguous, ill-defined, loosely structured, and require integration of knowledge. Success in situations such as these require more than choosing the best response from several options. Therefore, multiple-choice items only approximate the application of knowledge required outside of the school setting.

Many experts in the teaching of mathematics point to a need to assess elements beyond what can best be measured by multiple-choice tests (Romberg, 1989; Shavelson, McDonnell, Oakes & Carey, 1987; Stenmark, 1991; Working Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics, 1989). The widespread use of multiple-choice tests to monitor student progress and performance as been criticized by educators as failing to assess the most

significant outcomes of learning. Miller and Legg (1993) and Aschbacher (1991) assert that standardized achievement tests are too narrow in scope as they rarely address critical thinking and problem-solving skills. Romberg, Wilson, and Khaketla (1989) analyzed six commonly used commercial achievement tests. They found that at grade 8, an average of one percent of the items in the mathematics sections measured problem solving. Over three-fourths of the items measured lower level skills such as computation or estimation. The National Council of Teachers of Mathematics in their Curriculum Standards, suggest that the intent of mathematics instruction should be to promote making connections among mathematical concepts, understanding mathematical concepts and procedures, developing reasoning skills in mathematics, and enhancing mathematical problem solving skills. The Council also suggested that students' strategies and thought processes be assessed as well as answers. Performance assessment is one vehicle by which students can show their mathematical problem-solving strategies and reasoning skills (Crehan, 1991; Office of Technology Assessment, 1992; Tatsuoka, 1993).

#### Characteristics of Performance Assessments

Performance assessments are constructed responses to a stimulus which is evaluated by professional judgment (Aschbacher, 1991; Archibald & Newmann, 1988; Bennett, 1993; Bock, 1991; Lane, 1993; Mehrens, 1991; Phillips, 1993; Stiggins, 1987). Educators such as Aschbacher (1991), Archibald and Newmann (1988), and Wiggins (1989) put further stipulations on the characteristics of optimal performance assessments beyond merely a constructed response test item. Aschbacher (1991) proposes that performance assessments must have the following common elements. The assessment

must require students to utilize higher order thinking skills. The assessment tasks must consist of valuable instructional activities. The tasks must be a close simulation of a real world activity. Process behavior must be assessed as well as the end product. Also, the standards for performance must be known in advance.

Wiggins (1989) agrees that the tasks contained in a performance assessment must be representative of a real world situation. He adds that those individuals assessed need to be taught the criteria used to judge the performance, as self-assessment plays a greater part in performance assessments than in other types of tests. The self-assessment occurs as examinees plan to address the problem, monitor their progress, question and check their progress, and devise strategies to use when things go wrong (Snow & Lohman, 1993).

Archibald and Newmann (1988) formulated three criteria for the design of performance assessments. First, the tasks must depend on disciplined inquiry. Both outcomes and processes must be assessed. The student must have the opportunity to show an in-depth understanding of a problem which does not depend on merely repeating the knowledge of others. Second, the student must demonstrate an integration of knowledge. Third, the value of the test must go beyond that of providing a mechanism for evaluation.

The components of a performance assessment suggested by Stiggins (1987) are less complex. He proposes that there must be a reason for assessment. A clear and specific definition of the domain being assessed must be made. Exercises to elicit the performance must be developed, and systematic rating procedures must be specified.

Performance assessment items usually take longer to answer than typical multiple-choice items. Therefore, fewer performance assessment items are included per exam. Mathematics tests in which performance assessment items as well as multiple-choice items are included usually contain one or two performance assessment tasks. For example, the National Assessment of Educational Progress includes only one complex constructed response mathematics item on each test form (Johnson, Mazzeo, & Kline, 1993). Tests which are composed solely of mathematics performance assessment items typically contain fewer than ten items. For example, the Quantitative Understanding: Amplifying Student Achievement and Reasoning (QUASAR) requires each examinee to answer nine items (Lane, et al., 1992). In contrast, multiple-choice tests usually contain many more items.

#### Technical Qualities of Performance Assessments

Determining the reliability and validity of performance assessments presents unique challenges to psychometricians. The traditional distinctions between reliability and validity become blurred as one moves from narrowly defined domains with large samples of items drawn from the domain to more broadly defined and complex domains that must be represented by relatively few tasks. Further, within the context of tests that are systemically valid, i.e., the tasks have inherent value (Fredereksen & Collins, 1989), the traditional concept of generalizing from performance on a sample of items to performance on the item domain (reliability) is inextricably linked to the need to demonstrate that the tasks are adequate representations of the target domain (validity). Within a traditional testing framework, the concepts are complementary. Reliability is a necessary, but



insufficient condition for validity. However, reliability and validity do not always increase together. For example, when item selection is based on relation to an external criterion, validity is maximized at the expense of internal consistency (Crocker & Algina, 1986).

The distinction between reliability and validity is difficult to maintain and may be ultimately less useful in assessing the technical properties of performance assessments due to the complexity and ill-defined character of the domains, as well as the limited sampling of the domains which performance assessment represent. Thus, for example, relationships among tasks on performance assessment provide evidence for internal consistency reliability and construct validity.

Nevertheless, the presentation that follows treats information about reliability and validity somewhat separately. In addition, the technical qualities of performance assessments in the domain of writing will be presented because most of the research on performance assessment has been in this area.

The reliability of any assessment refers to the degree to which the scores are free of random error. Baron, Forgione, Findone, Kruglanski, and Duey (1989) define three sources of error typically found in performance assessments. Error can be inferred from differences between two people rating the same task, differences between scores on tasks designed to measure the same construct, or differences in scores from one occasion to the next not attributed to maturation or learning.

Huot (1990a) and Baron et al. (1989) argue for the need to assess all sources of error so that generalizability to the construct measured can be made with confidence. The source of error that has received the most attention in the literature is that of different

raters. Although some studies have reported low correlations among raters in writing assessments (Diederich, French & Carlton, 1961; Dunbar et al., 1991), more recent evidence indicates that with carefully developed criteria and proper rater training, inter-rater reliability can be quite satisfactory (Cantor & Hoover, 1986; Hieronymus & Hoover, 1987; Office of Technology Assessment, 1992). While inter-rater reliability is most frequently reported in performance assessment studies, few studies report evidence of the error due to different tasks or occasions. Those studies addressing error due to sources other than raters, frequently show a large person by task interaction (Baxter, 1992; Shavelson et al., 1991; Shavelson et al., 1990). It appears that the criteria for determining comparable performance assessment tasks is not as well defined as the criteria for obtaining satisfactory inter-rater reliability.

Like reliability, evidence of the validity of performance assessments may require modifying the approach usually taken with multiple-choice tests. Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13, emphasis in the original). Evidence regarding the validity of a test for a particular application has traditionally been divided into three types: content, criterion-related, and construct. Content validity refers to the degree to which the test items reflect the subject matter domain the test is proposed to measure. Subject matter experts judge the alignment between the test content and the domain measured. Criterion-related validity is generally determined by correlating the results of the test with another measure of the domain being

assessed. Construct validity is determined by evidence that the test measures the intended construct or domain assessed (Crocker & Algina, 1986). Messick (1989) points out that almost any kind of information about a test regarding the interpretation or meaning of test scores contributes to the understanding of the construct validity. Therefore, construct validity subsumes both content and criterion-related validity. However, validity evidence which supports the theoretical construct from which the assessed domain is taken produces the most convincing evidence of construct validity.

Traditionally, measures of student achievement in subject matter domains have been shown to be valid through evidence of content validity (Lane, 1993). Performance assessments often appear to be measuring important aspects of the subject matter domain and so have inherent face validity. This has led some to accept the validity of performance assessments without further examination because they seem to assess the goals of curriculum experts such as the Working Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics (1989). For example, Crehan (1991) asserts that "since the performance exercise is the natural goal of instruction, its instructional and content validity should be unassailable" (p. 6). However, content validity based on expert judgment can be fallible, and evidence of content validity (much less face validity) does not replace evidence of construct validity (Huot, 1990b; Lane, 1993; Linn et al., 1991; Mehrens, 1991; Messick, 1989).

Messick identifies two threats to construct validity: construct irrelevance and construct underrepresentation. Construct irrelevance can occur when the test items are too hard or too easy for the subjects being tested. For example, tests are irrelevantly

more difficult for some students if reading comprehension, writing ability, or familiarity with the context of the item interferes with students' opportunity to demonstrate their mathematics ability (Lane, 1993). Construct underrepresentation occurs when the test fails to include all aspects of the construct among the items. Performance assessments are especially vulnerable to the latter, since typically fewer items are included than in multiple-choice tests. Lane warns that even when performance assessment tasks have high content validity, they may underrepresent the construct domain measured, and/or measure the target domain along with other irrelevant domains.

Evidence of construct representation, and thus construct validity, is obtained in part when systematic differences are not evident among the responses to test items or tasks. Such systematic differences can occur when the context, or idiosyncrasies, of a task influence an examinee's score to a large degree in relation to the domain being assessed. Messick (1989) recommends that systematic appraisal of the context effects in score interpretation be conducted. Context effects refer to the characteristics of a particular test item or testing situation. If the context effects are found not to contribute significantly to test scores, then the generalizability of the test results across different tasks is appropriate.

The issue of the validity of performance assessments has been debated throughout the literature. Performance assessments differ from multiple-choice tests in that the former measures process as well as product, and usually addresses higher order thinking skills (Miller & Legg, 1993). As a result, some have asserted that the methods of evaluating the technical qualities of performance assessments should not be limited to those measures of validity and reliability usually associated with multiple-choice tests

(Baker et al., 1993; Linn et al., 1991; Miller & Crocker, 1990; Wolf et al., 1991). Baker et al, and Quallmalz (1991) have identified some criteria to be used as evidence of the validity of performance assessments. They recommend that performance assessments be analyzed to determine if they are measuring competencies amenable to instruction, if the items and scoring are fair to all students, if the scores on the particular assessment tasks are similar to scores on other tasks within the same domain, and if the consequences of the test scores support the most appropriate instruction. These criteria differ from the traditionally reported content validity for achievement tests. However, models for determining evidence for the validity of performance assessments seems to be at the brainstorming stage, as standards for employing the criteria have yet to be established. In fact, Baker et al. point out that their criteria must be operationally defined in order to be utilized as evidence of validity. It appears that psychometricians have yet to establish definitive standards for assessing the validity of performance assessments.

However, both Baker et al. and Quallmaltz assert that a relationship needs to be established among performance assessment tasks from the same subject matter domain, before a performance assessment can be judged valid. Likewise, Messick (1989) states that evidence of construct validity includes information concerning the relationship among tasks which purport to measure the same domain. Therefore, as performance assessment measures often include only a few tasks, it is imperative that empirical evidence supports the generalization of the results of the performance assessment to the broader domain of interest.

A variety of techniques have been used to determine the evidence of the reliability and validity of performance assessment measures. Since performance assessment measures have been most commonly used to assess writing ability, much of the research has been done with direct writing assessments. Information from the studies on direct writing assessment can certainly inform those interested in other content domains such as mathematics. Therefore, a summary of this evidence is presented below.

The inter-rater reliability of direct writing assessments has been reported to be in the .90's (Cantor & Hoover, 1986; Hieronymus & Hoover, 1987). However, care must be taken to create the conditions for high inter-rater reliability. For example, Diederich et al. (1961) gave 300 essays to a number of judges to rate. The judges were given no scoring criteria, but were told to grade the essays using their professional judgment. Diederich et al. found that 94% of the essays received at least seven different scores on a nine point scale, and no paper received less than five scores. Inter-rater reliability was reported as .31.

Similar disparity between inter-rater reliability is reported by Dunbar et al. (1991). In a description of eight studies of direct writing assessment, the inter-rater reliabilities ranged from .33 to .91. Higher reliability coefficients were found when the scoring criteria were specific and the standardization of the testing conditions was high. These studies indicate that inter-rater reliability can be as high as the .90's when raters are thoroughly trained and scoring rubrics are specific. Therefore, at least with respect to writing performance assessments, inter-rater reliability is probably of less concern in performance assessments than other sources of error.

It appears to be easier to obtain inter-rater reliability, than to identify multiple tasks which measure the same domain. A number of studies have examined the correlation between two or more tasks from the same domain and found low to moderate correlations. Inter-task reliabilities were obtained by Dunbar et. al. (1991), by correlating two samples of performance from students in which each pair of samples had been scored by the same raters. The tasks were samples from the same content domain. Correlations ranged from .26 to .60. In all cases, the inter-task reliabilities were lower than the inter-rater reliabilities.

Similarly, Breland and Gaynor (1979) asked freshmen to write three essays. One was written at the beginning of school; the second was written at the end of the first semester; and the last was written at the end of the second semester. Even though there were long time periods between testing occasions and though intervening instruction had occurred, correlations between scores were calculated as alternate form, test-retest reliability coefficients. The estimated coefficients were .50 to .51.

With regard to inter-task consistency, Cantor and Hoover (1986), and Hieronymus and Hoover (1987) report the results of a standardization study of the direct writing assessment of the Iowa Test of Basic Skills. Students in grades 3-8 were asked to write essays from two prompts. Each prompt represented one of several modes of writing, including narrative, explanatory, descriptive, informative, and persuasive. The correlations of scores of two essays from the same mode ranged from .37 to .69. This study, as well as those described above, demonstrate that the correlations between two tasks from the same domain may vary dramatically.

Meredith and Williams (1984) also concluded that the results of direct writing assessment vary by particular task. They report that in the large-scale writing assessment given to all ninth graders in Texas, the pass rate increased by 13% from one year to the next (81%-94%). A different prompt was given each year. A difference of 13% in the pass rate was also found in Maryland when two assessments were administered simultaneously to ninth graders with each student receiving both prompts. Ninety-three percent passed with one prompt, whereas only 80% passed with the second prompt.

The correlations between essays written in different modes of discourse have been found to be quite low. Cantor and Hoover (1986) correlated the scores of students who were asked to write one essay from a prompt in one mode, and another essay from a prompt in a second mode. The correlations between modes ranged from .32 to .40. Similarly, Quellmalz, Capell, and Chou (1982) asked eleventh and twelfth graders to write one essay in the expository mode and one essay in the narrative mode. They concluded that levels of performance vary substantially on tasks presenting different writing purposes. Kegley (1986) also found large differences in performance between writing assignments in the narrative mode and in the persuasive mode.

Breland et al. (1979) used confirmatory factor analysis to see if a single writing ability could account for the performance on six essays. The subjects each wrote two essays in the narrative mode, two essays in the expository mode, and two essays in the persuasive mode. The results showed that the best model contained six factors, one for each topic. So even the two essays written from the same mode did not load together to



create one factor by writing mode. Thus, students could have been rated highly on one topic and mediocre on another topic.

Boodoo and Garlinghouse (1983) and Meredith and Saunders (1984) used analysis of variance to test the hypothesis that there was a topic by person interaction. The interaction effect was found to be significant in both studies, indicating that performance varies by topic.

None of these studies of direct writing assessment support the generalization of the results of a single writing prompt to a single construct of writing ability. Reliability appears to be enhanced when the domain assessed is specific and the conditions of the test are standardized. Huot (1990a) and others (Meredith & Williams, 1984; Purves, 1992) agree that more information is needed regarding the specification of the writing domain and tasks to measure the writing domain in order to generalize to the writing ability of those assessed.

The lack of exchangeability of the scores between two direct writing performance assessment tasks does not support their use in high stakes situations. However, the subject matter domains of writing and mathematics differ. To the extent that mathematics domains can be more precisely defined, it is possible that differential performance by task may be less of a factor for mathematics than for writing. In any event, the studies on direct writing assessment point to a need to seek to establish the conditions of assessment that maximize the likelihood of generalizability. The evidence of success in devising comparable performance tasks for other subject matter domains such as mathematics should not be taken for granted without sufficient evidence.

## Technical Qualities of Performance

### Assessments in Mathematics

Seven studies on the reliability and validity of performance assessments in mathematics have been reported. The type of evidence reported varies by study, although inter-rater reliability is the most common topic.

The magnitude of the inter-rater reliability differed by study. Vermont recently implemented a statewide portfolio assessment program which included mathematics (Koretz, 1993). The inter-rater reliability was reported to range from .33 to .43, depending on the grade level assessed. The low reliability coefficients were attributed by the author to a lack of scorer training. As the coefficients were not high enough to have indicated stable scores, widespread use of the results was precluded. On the other hand, Stevenson, Averett, and Vickers' (1990) report of the North Carolina Department of Public Instruction's field test of geometry proof performance assessments, described inter-rater reliability in the range of .82 to .95. The authors report extensive training of the teachers used as raters in the study. After training, the teachers were required to reach a designated level of accuracy in using the scoring rubric, before they began actual rating of exams.

Another study (Johnson et al., 1993) reporting inter-rater reliability describes the 1992 National Assessment of Educational Progress (NAEP). This exam contains 183 items in the eighth grade mathematics portion of the test. Of those items, 59 were short constructed response, and six were extended constructed response questions. Twenty percent of all the constructed response items were scored by a second rater. The

inter-rater reliability correlations ranged from .69 to .91 with an average of .81 for the extended constructed response items. For the short constructed response items, the inter-rater reliability ranged from .82 to .99 with an average of .96.

With respect to reliability, these studies find similar conclusions to those of direct writing assessment. Inter-rater reliability can be as high as the .90's, when rater training is well done.

Validity information reported for mathematics performance assessments included content and criterion-related evidence. Extensive evidence of content validity was reported for the items included in the NAEP (Johnson et al., 1993). The objectives for items to be included in the assessment were developed through a national consensus process involving educators, experts in the field of mathematics, and interested citizens. They decided to assess mathematical conceptual understanding, procedural knowledge, and problem-solving. The content areas assessed were numbers and operations; measurement; geometry; data analysis, statistics and probability; and algebra functions. The items written to reflect the objectives were reviewed by mathematicians, measurement specialists, experts in bias/sensitivity, and representatives from state education agencies. Correlations between the performance assessment involving two geometry proof tasks and a multiple-choice test, course grades, and instructor ratings ranged from .48 to .63 (Stevenson et. al., 1990). Thus, some evidence of predictive validity of the performance assessment was established.

The most comprehensive reports of the technical qualities of mathematics performance assessments come from Shavelson, et al. (1993), Baxter, et al. (1993), and

Lane, et al. (1992). Shavelson et al. (1993) and Baxter et al. (1993) use a generalizability theory model (Shavelson & Webb, 1991) to provide evidence of the reliability and convergent validity of performance assessment tasks measuring the mathematics domains of measurement, probability, and place-value with sixth grade elementary students. Following this model, a student's achievement on a performance assessment task can be affected by several sources of error: tasks, rater, occasion, and method of assessment. The error due to these conditions of measurement can limit the generalizability of the results, as well as demonstrate a lack of convergent validity. In each of these studies, a large person by task interaction was found, indicating different performance depending on the task given. Generalizability studies can produce two type of coefficients: generalizability coefficients and dependability coefficients. Both are like classical reliability coefficients but may consider several sources of error simultaneously. Generalizability coefficients can be used when the purpose of assessment is to give a normative interpretation, i.e. to rank examinees on the amount of the construct domain assessed. Dependability coefficients can be used when the actual amount of the construct domain possessed by the examinees is of interest (e.g., when a passing score is established). Baxter et al. and Shavelson et al. found that to reach a generalizability coefficient of .80, 15 tasks would need to have been given to each student. A dependability coefficient of .80 would require 20 tasks. Designing a performance assessment with a battery of 15 or more tasks would be quite costly, and if each task took 15 minutes, the testing time would require nearly four hours.

Lane et al. (1992) report evidence of the reliability and validity of a mathematics performance assessment test designed to evaluate the instructional effectiveness of a national project entitled QUASAR. The QUASAR program aims at enhancing the mathematical thinking and reasoning skills for middle school students in economically disadvantaged areas. The QUASAR Cognitive Assessment Instrument (QCAI) was designed to reflect the mathematics domain as it is described in the Curriculum and Evaluation Standards for School Mathematics (Working Groups, 1989). The assessment instrument was separated into four forms, each containing nine tasks. Though the forms were not considered strictly parallel, the tasks were distributed to make the forms as similar as possible with respect to mathematical content, cognitive processes, mode of representation, and task context. To obtain evidence that the tasks on the QCAI could be generalized to the larger domain of mathematics, confirmatory factor analysis was used. A one-factor model fit the data for each form. Thus, support for generalizing scores to the domain of mathematics was obtained. To provide further evidence, a generalizability study was conducted using person, rater, and task facets. The results indicated that the variances due to rater, the interaction of person with rater, and the interaction of rater with task were negligible. Consistent with the results found by Shavelson et al. (1993) and Baxter et al. (1993), the largest percentage of the total variance came from the person by task interaction. The generalizability coefficients ranged from .740 to .836, for the nine tasks on each forms. The dependability coefficients ranged from .709 to .802. Lane et al. (1992) noted that these coefficients were sufficiently high for evaluating program effectiveness, but not high enough to evaluate the performance of individual students.

The conclusions reached by Shavelson et al. (1993), Baxter et al. (1993), and Lane et al. (1992) are consistent with the conclusions reached by those studying direct writing assessment. Their data do not support generalizing from the results of a small number of performance assessments if the domain of interest involves complex mathematics problem-solving.

### Impact of Performance Assessment Results

#### On Gender and Ethnic Groups

Another factor to be considered is the fairness of performance assessment tasks to all groups of students. An important aspect of the development of any type of assessment is eliminating the effect of background factors. The removal of background factors such as gender and ethnicity allows tests to validly measure the learning that has resulted from instruction without confounding the results with construct irrelevant variance (Office of Technology Assessment, 1991). The research addressing the effects of performance assessment on ethnicity and gender is limited. When writing is assessed, the results of studies conducted by Breland and Griswold (1981) and Breland and Gaynor (1979) indicated that compared to males, females perform better on performance assessments such as essays and less well on multiple-choice tests. Likewise, Baker (1992) found significant gender effects on the results of performance assessments measuring a high school history domain. Students were given primary source history texts and asked to write an essay explaining the position of authors, and to draw on students' own background knowledge to explain their reasoning. As with other performance

assessments measuring a writing domain, females outscored males when rated on such dimensions as general content quality, principles, and argumentation.

One study addressed gender differences on the scores of mathematics constructed response tests. Bolger and Kellaghan (1990) found that high school boys in Ireland received relatively lower test scores in mathematics, Irish, and English when they were assessed with constructed response tasks than when they were assessed with multiple choice items. Conversely, high school girls in Ireland performed better on constructed response tasks in the same subject areas. The boys outperformed the girls on both types of measures of mathematics, but the gap was lower on the performance assessment tasks. The results of these studies indicate that females may have an advantage when performance assessment tests are the method of measurement.

The effect of performance assessments on reducing the achievement gap between ethnic groups that is found in standardized, multiple-choice assessments has yet to be established. Linn et al. (1991) found that the results from the 1988 National Assessment of Educational Progress showed similar achievement patterns between white and black students on writing scores and reading scores. Writing was assessed by performance assessment while reading was assessed primarily by multiple-choice items. Therefore, type of assessment did not tend to favor either ethnic group. Likewise, Fenberg (1990) noted that when the California Bar exam included a performance assessment, the passing rates between whites and minorities remained the same as the passing rate with no performance assessment. Contradictory results were found by White (1985). His study showed that minority students performed better on the California State University and

College English Placement Test than on the Test of Standard Written English. The former was a writing performance assessment and the latter was composed of 50 multiple-choice questions. There was no difference in performance between the two measures for white students.

The only study addressing the differential performance by ethnic group on a mathematics performance assessment was conducted by Baxter et al. (1993). This study looked at the impact of mathematics curriculum on performance as well. Mathematics performance assessments were given to sixth graders in both an innovative curriculum designed to reflect the reform standards recommended by the Working Group for the Commission on Standards for School Mathematics (1989), and in a traditional curriculum in which textbook and worksheets predominated. Whites outscored Latino students in all mathematics domains assessed regardless of curriculum. Surprisingly, the gap in achievement level between the two ethnic groups was larger in the reform based curriculum than in the traditional curriculum. These results point to the difficulty of teaching the integrated, problem-solving skills required by the performance assessment instruments in a single school year. Furthermore, the Latino students may have had more of a disadvantage on the performance assessment if they were not native English speakers.

#### Summary

Currently, evidence of the technical qualities of performance assessments in mathematics relies extensively on content validity, and/or inter-rater reliability. However, these do not provide sufficient evidence of reliability and validity. Messick (1989), Baker



et al. (1993), and Quallmaltz (1991) assert that since performance assessment measures often include only a few tasks, it is imperative that empirical evidence be obtained to support the generalization of the results of the performance assessment to the broader domain of interest. Generalizability theory allows evidence of reliability to be extended to include task and occasion facets as well as rater facets. The recommended validation of performance assessments includes examination to determine if the tasks and scoring procedures are fair to all groups of students (Baker et al.; Quallmaltz). The comparison of the performance between gender and ethnic groups as well as the differential performance of all students across tasks and occasions provides some evidence to establish construct validity of the assessment as well.

The present study will evaluate the technical properties of an eighth grade mathematics performance assessment following the recommendations of Baker et al. and Quallmaltz. The inter-task reliability, stability over time and inter-rater reliability is assessed simultaneously through generalizability theory. The impact of highly similar tasks on inter-task reliability is examined. Performance differences by gender and ethnic groups are compared as well.

### CHAPTER III

#### METHOD AND PROCEDURES

This section describes the participants in the study, the development of the Mathematics Problem Solving Assessment (MPSA), and the procedures used to conduct the study. The methods used for analyzing the data are discussed as well.

##### Subjects

The subjects were 101 eighth grade students enrolled in regular mathematics classes in an affluent suburban district located in the Oklahoma. The students were selected based on the willingness of their mathematics teachers to allow class time for participation in the study. Thus, the group can be classified as a convenience sample.

Student participants in the study were classified by their teacher as average and below average students (S. Parker, personal communication, March 1, 1994). No students who were enrolled in science, English or mathematics honors classes on the basis of Iowa Test of Basic Skills subtest scores were used in the study.

The participants were enrolled in regular eighth grade mathematics classes. The curriculum used was revised during the summer preceding the school year to reflect a collaborative, team approach among all eighth grade mathematics teachers at the school. The new curriculum followed the Oklahoma State Department of Education's recommended goals for mathematics instruction. In addition, the district mathematics

curriculum consultant conducted summer workshops for teachers regarding ways to teach mathematics that reflect the most current research on cognitive learning theory.

Students were tested during their regular class period on two days during the spring semester. On the first testing occasion, students taking the performance assessment exam during the third class period did not receive the full testing time because the lunch period was scheduled in the middle of the class. Exams that were not finished before the lunch break were collected but not used. Usable exams were obtained from 73 students on the first testing occasion and from 97 students on the second testing occasion.

Students with usable exams from both testing occasions numbered 73. The total sample of 101 students included 41% female and 59% male. The group consisted of 77% White students and 23% ethnic minorities. The group of the 73 students with usable exams included 38% female and 62% male, and 74% White students and 26% ethnic minorities. Ethnic minorities consisted of Black, Native American, Asian, Hispanic, and Middle Eastern students. It is unknown whether or not the socioeconomic status of the ethnic minority students differed from the White students.

One limitation of this study is the small sample size. Small sample size is common in generalizability studies such as this one in which rater is used as a facet. In order to keep the rater facet from being nested within another facet, it is necessary for each rater to rate all tasks on all occasions. In order to keep the burden of each rater at a manageable level, limitations are placed on the number of participants who can be included in the study. Typical sample sizes for generalizability studies of mathematics performance assessments are close to 100 (Baxter et al., 1993; Lane et al., 1992; Shavelson et al.,

1993) when a person by task by rater design is used. The sample size used by Shavelson and his colleagues (1993) to test a person by task by rater by occasion design for the generalizability of science performance assessment was 29. This experiment used the same design as was used in the present study.

### Instrumentation

#### Guidelines for Development of Performance Assessment Instruments

An abundance of literature exists to provide information regarding the development of items for multiple-choice exams (Lane, Parke & Moskall, 1992; Roid & Haladyna, 1982). Less is available on the selection and development of performance assessment tasks. However, Lane et al. (1992, p. 4-7) have developed principles to follow when performance assessment tasks are constructed. The instrument used in this study was developed by the researcher, and is entitled the Mathematics Problem-Solving Assessment (MPSA). The MPSA contains four performance assessment tasks. Lane's principles guided the identification of the performance assessment tasks for the MPSA as described in the sections that follow.

Principle: Provide a clear description of the purpose of the assessment (Lane et al., 1992, p. 4)

The performance assessment tasks on the MPSA are designed to assess the mathematics problem-solving process skills written by the Oklahoma State Department of Education (1993), as part of the State of Oklahoma's mandated curriculum for eighth graders. The results of the assessment tasks provide a method to demonstrate mastery of

the content domain by students at the individual or school level. Stiggins (1987) classifies this type of purpose as one for certification.

Principle: Provide a clear description of the breadth and depth of the construct domain (Lane et al., 1992, p. 4)

In the Curriculum and Evaluation Standards for School Mathematics (Working Groups, 1989), Resnick describes mathematics as involving problems that are complex, yield multiple solutions, require interpretation and judgment, require finding structure, and require finding a path for a solution that may not immediately be apparent. The problem-solving tasks used for this study were chosen or designed to reflect this view of mathematics.

The content categories in the mathematics domain identified by the National Council of Teachers of Mathematics (Working Groups, 1989) include numbers, operations, and computation; patterns and functions; algebra; statistics; probability; geometry; and measurements. The Priority Academic Student Skills (PASS) that describes the state's mandated curriculum utilizes similar content categories: Number sense and number theory, computation and estimation, patterns and functions, algebraic concepts, statistics, probability, geometry, and measurement (Oklahoma State Department of Education, 1993, pg. 47-48). To maximize the chance of inter-task reliability among the items, all tasks included the similar content areas of number sense and number theory, operations and computation; and patterns and functions. None of the tasks required the content areas of geometry, measurement, statistics, or probability.

Principle: The specified view of the construct domain should be consistent with the curriculum and instruction (Lane et al., 1992, p. 4).

This principle supports the selection of performance assessment tasks which measure the state's mandated curriculum, PASS. These skills are intended to guide the curriculum in Oklahoma schools. The skills reflect the suggestions made for curriculum reform in mathematics (Working Groups, 1989). The recommendations include increasing the presentation of situations requiring open-ended, extended problem solving and decreasing the emphasis on practicing routine, one-step problems. Learning mathematics is to be guided by the search to answer questions. Instruction is to reflect the model of learning where students are seen as approaching a new task with prior knowledge, assimilating the new information, and construction their own meaning. The new ideas are only accepted as the old ideas are found to be cumbersome or not to work (Resnick, 1987). The standards (Working Groups, 1989) direct teachers to follow this model of learning, as opposed to the model where students absorb information, store it, and retrieve pieces as needed. The assessment tasks used in this study are designed to follow the recommended model of learning.

The mathematics curriculum consultant for the district from which the sample is taken, indicated that the teachers of the students in the study are trying out ideas which reflect the recommendations for mathematics curriculum reform (L. Bailey, personal communication, December 1, 1993). However, the curriculum also includes conventional methods of teaching mathematics. Therefore, the students in the sample have experienced

some of the new approaches to mathematics instruction, as well as some of the more traditional types of mathematics classroom activities.

Principle: Specify the processes to be measured by the assessment instrument and each individual task (Lane et al., 1992, p. 5)

The processes tested by the performance tasks in the present study are defined by the following objective from the PASS (Oklahoma State Department of Education, 1993, pg. 47):

"The student will: Develop and test strategies to solve practical, everyday problems which may have single or multiple answers; Evaluate results to determine their reasonableness; Use oral, written, concrete, pictorial, graphical and/or algebraic methods to model mathematical situations; and Apply a variety of strategies (e.g. trial and error, diagrams, making the problem simpler) to solve problems, with emphasis on multi-step and non-routine problems."

#### Development of the Mathematics Problem-Solving Assessment

A number of mathematics performance assessment tasks are found in the literature. Several examples come from items used by national assessment programs such as the National Assessment of Educational Progress (Johnson et al., 1993). Other examples are found in reports of state assessment program (Baron et al., 1989; Far West Laboratory for Educational Research and Development, 1992). Still other examples come from assessment instruments designed to evaluate the effectiveness of particular mathematics programs (Lane, Stone, Ankenmann, & Liu, 1992). In addition, many instructional activities designed to reflect the mathematics curriculum reform standards (Working

Groups, 1989) are easily adapted into performance assessment tasks. Shavelson et al. (1993) and Baxter et al. (1993) used instructional activities to develop the performance assessment tasks used in their studies.

The first steps in the development of the MPSA began by identifying performance assessment tasks published in the professional literature that reflect the problem-solving skills outlined in the PASS (Oklahoma State Department of Education, 1993). Next, instructional activities were obtained from three mathematics curriculum consultants (one at the state level, and two from the school district level). The result was ten written performance tasks developed by mathematics experts for use with eighth graders.

While all the tasks required the desired problem-solving process skills, a number of content skills were also required. It was decided to use only those tasks emphasizing the use of number sense and number theory, computation and estimation, and patterns and functions. Therefore, tasks requiring the content areas of geometry, measurement, statistics, and probability were eliminated.

After elimination based on content categories, three performance tasks remained from the original pool. Two were developed for use in the National Assessment of Educational Progress (Blumberg, Epstein, MacDonald & Mullis, 1986; Johnson et al., 1993), and one was developed by Burns (1992) as a teaching activity. Four items were desired for the MPSA so that the test could be completed within one class period. Therefore, all three were retained for inclusion in the MPSA. The fourth performance assessment task was written by the researcher to be highly similar to one of the NAEP tasks, using the guidelines given by Roid and Haladyna (1982).



The original MPSA comprised of these four tasks was presented to four teachers currently teaching eighth grade mathematics. A focus group format was used. The teachers were given copies of the PASS for eighth grade mathematics and the MPSA. They were then asked to identify any process skills and content skills that the MPSA tasks represented. All teachers agreed that the four tasks required students to use mathematics as problem-solving skills as described in the PASS objective. All also concurred that the content skills required by the tasks included number sense and number theory, computation and estimation, and patterns and functions, and did not include statistics, probability, geometry or measurement. Consensus was not reached on whether or not eighth graders would use algebraic concepts to solve the problems. The teachers were also asked to review the wording of the problems and suggest any revisions to insure that eighth graders would understand the tasks. Suggestions included the need to specify a timeframe for travel for the Basketball Camp and the Space Camp tasks, and a concern that the students would not be able to follow the direction, "explain your reasoning," for any of the tasks.

Two school district curriculum consultants were individually asked the same questions as the teachers. They both concurred with the teachers that the tasks reflected the problem-solving skills outlined in PASS. They also agreed with the conclusions reached by the teachers regarding the content covered by the items, with the exception that both thought that algebraic concepts were reflected in the tasks. Suggestions for revisions in the wording of the tasks included changing "Explain your reasoning." to "Use words, diagrams or pictures to explain your thoughts. Write in complete sentences how

you arrived at your answer." Also recommended were elimination of the terms "fixed costs" and "mentor."

One teacher and one school district official, both experienced at working in multicultural settings, examined the tasks used in this study to determine that each was free from cultural, linguistic, and gender bias. No bias was detected.

The original items were pretested with 40 eighth graders during the last week of January, 1994. These students attended the same affluent middle school as the study participants but the pretest sample included both honors (25%) and regular (75%) mathematics students. The MPSA was administered during a time when the students were heterogeneously grouped for instruction in study skills. The results indicated that eighth graders were able to complete the items within one class period. Even though most of the students appeared to understand the problems, the wording of the tasks was revised to reflect the suggestions made by the mathematics teachers and curriculum consultants. The revised performance assessment tasks on the MPSA are presented in Appendix A.

#### Data Collection Procedures

The three performance tasks on the MPSA were administered to the students in the sample during their mathematics class period. The initial assessment took place during the first week of April. The second assessment was administered to the same students three weeks later. The order of the tasks was distributed randomly to students.

The students put their name on the cover sheet of the MPSA. To protect the anonymity of the students, the students' teacher coded each test with a student identification number. She also indicated the ethnicity and gender of the student. The

cover sheet with the students' names was removed before the completed MPSA's were returned to the researcher.

### Scoring the Performance Assessment Tasks

The holistic scoring method used assessed the completeness of the entire response to each task in comparison to a specific scoring rubric (Archbald & Newmann, 1988).

The initial scoring rubrics for the NAEP tasks were modeled from those used by the developers of each task (Blumberg et al., 1986; Johnson et al., 1993). No scoring rubric had been developed to accompany the performance assessment task by Burns (1992).

Therefore, the initial scoring rubric for use with this task was written by the researcher in collaboration with two mathematics teachers and a school district mathematics consultant.

The scoring procedure was piloted using the preliminary scoring rubrics and the performance assessment responses from the pilot study. A high school mathematics teacher and a university professor specializing in educational measurement were the raters.

Each rater scored a sample of ten students' responses for each task. Scores given by each rater were discussed. Ambiguity in the scoring rubrics was rectified, and recommendations for modifications in the scoring rubrics were made. The resulting rubrics allowed for consensus in the appropriate score for each student response. The final scoring rubrics for each task are presented in Appendix B.

The MPSA results from the actual study were scored by two teachers with experience teaching mathematics at the middle school grades. Before scoring began, the raters were instructed about performance assessments in general. Then, the raters were given the tasks to complete without the benefit of the scoring rubrics. Afterward, the

rubric used for each task along with sample student answers for each scoring level were presented. The raters then practiced scoring sample student responses. Following the practice scoring periods for each task, the raters discussed their scores for each response. Discrepancies among scores were analyzed in light of the scoring criteria. Consensus was reached regarding the interpretation of each scoring rubric. Once the raters were rating consistently with the intent of the scoring rubric, they began scoring the actual student responses. All tasks from both testing occasions for all subjects were rated by each rater. The raters spent approximately 24 hours each to score the MPSA responses. The average time to rate each task was 2.67 minutes.

#### Data Analysis

In classical test theory, a person's observed score on an assessment is composed of a true score component and an error component. Reliability is concerned with the effect of random error on the consistency of scores, while the validity is affected by systematic error (Ary, Jacobs, & Razavieh, 1990). According to classical test theory, the proportion of the observed score that is true score variance is estimated by a reliability coefficient. A classical reliability coefficient can be estimated in several ways. Stability reliability coefficients are calculated by administering a test to the same individuals on separate occasions and correlating the scores. A reliability coefficient calculated by correlating individuals' scores on two alternate forms of the same test on the same occasion is called "coefficient of equivalence". When people are tested with one form on one occasion and with another form on a second occasion and their scores on the two forms are correlated, the resulting coefficient is called a "coefficient of stability and equivalence".

Other measures of reliability require only one administration of a test. These coefficients estimate the degree of consistency among items sampled from a domain by splitting the test into two parts and correlating scores on each part (split-half reliability) or by assessing the inter-item consistency (Kuder-Richardson procedures and coefficient alpha). Still another way to estimate classical reliability when test performance must be rated is to correlate the scores obtained by two or more raters. The result is an "inter-rater reliability coefficient".

As classical test theory only defines a single error component at one time, there is only one source of error estimated in each reliability coefficient. For example, test-retest reliability coefficients consider testing occasion as the source of measurement error. Inter-rater reliability coefficients consider only rater effects as the source of error. Reliability coefficients of stability and equivalence consider testing occasion and equivalent forms simultaneously but do not provide separate estimates regarding the amount of error introduced by occasion and form.

An advantage of generalizability theory over classical test theory, is that multiple factors that contribute to the error component can be assessed simultaneously. Each factor is called a facet. In the present study, the observed score is divided into a universe score (analogous to true score), a task facet, a rater facet, and an occasion facet. The amount of variation associated with each facet will be measured as well as the variance associated with problem-solving ability of the person assessed. Generalizability theory can be used to determine how much of the variance found in an observation can be attributed

to each facet. Analysis of variance is used to estimate variance components due to each facet specified in the model.

The degree to which an observed score represents the universe, or true score, can be estimated with generalizability coefficients. Each coefficient is interpreted as an approximation of the correlation between observed scores for two independent random samples of observations from the universe of observations, or as approximately equal to the expected value of the squared correlation between the observed score and the universe score (Shavelson & Webb, 1991).

Generalizability theory provides both relative and absolute coefficients. They differ in the definition of the error term. The choice of coefficients to use depends on the proposed interpretation of the measurement. Generalizability coefficients give relative interpretations that focus on the rank-ordering of individuals. Only the variance components which affect the individual's relative standing are included in the error term. Dependability coefficients give absolute interpretations that provide information about the amount of the domain assessed which the individual possesses. The latter is analogous to a criterion-referenced interpretation, and so includes all sources of variance which affect the individual's level of performance.

Generalizability theory distinguishes between generalizability studies (G-studies) and decision studies (D-studies). The former uses ANOVA to partition variance into universe score and error facets. The latter uses the information from a generalizability study to determine the number of levels of the facets necessary to obtain a test score that will generalize to the domain of interest. For example, once the variance components

have been estimated in a G-study, the D-study analysis can estimate the number of raters, tasks, and/or occasions needed to achieve a desired level of generalizability (reliability). Decision studies estimate the generalizability and/or the dependability coefficients when various numbers of facets such as tasks, raters or occasions are included in a hypothetical assessment. The estimates are made in much the same way as the Spearman-Brown prophesy formula is used to estimate the reliability of an assessment with differing numbers of parallel items (Shavelson & Webb, 1991).

The variation in the scores of the three performance assessment tasks on the MPSA were examined through generalizability and decision studies utilizing GENOVA (Crick & Brennan, 1983). The G-study examined a person, task, rater, and occasion facet ( $p \times t \times r \times o$  design). All facets were assumed to be random. Estimated variance components were used to demonstrate the contribution of each facet to the error component of the MPSA score. The percent of total variability attributed to the universe score, each facet, and the interaction of facets was computed. The standard errors of the variance components were reported as well. The D-study showed the generalizability and dependability coefficients for a performance assessment using varying numbers of tasks, testing occasions, and raters. An estimate was made of the number of tasks, occasions, and raters needed to reach a G-coefficient of .80 or higher.

Split-plot ANOVA was used to compare group differences in performance between males and females and between majority and minority ethnic groups.

## CHAPTER IV

### RESULTS

#### Descriptive Data

Table 1 provides the average MPSA scores by rater, task, and occasion. Potential scores on each task range from zero to four. The raters rated each task similarly, though rater 2 tended to give slightly higher scores than rater 1. Table 2 summarizes the MPSA scores by task and occasion averaged across raters. Highest average scores are seen for the Olympics task and lowest scores for the Space Camp task. As expected, students' scores increased on average from the first testing occasion to the second testing occasion, although very little for the Olympics task.

Tables 3-6 show the frequencies of scores obtained by the students on each task by testing occasion. The distributions for the Basketball Camp, Space Camp, and the Tug of War tasks are positively skewed on the first testing occasion. A higher percentage of students received scores in the high range on all tasks during the second testing occasion.

#### Classical Reliability

Classical reliability coefficients are frequently presented as evidence of the reliability of assessment instruments. These coefficients are contrasted with the generalizability coefficients presented later in the section. The inter-rater reliability coefficients calculated for the first and second testing occasions respectively are .82 and .84 for the Basketball Camp task, .83 and .86 for the Space Camp task, .87 and .86 for the Tug of War task, and .77 and .88 for the Olympics task. Inter-rater coefficients are the



Table 1. MPSA scores by rater ,task, and occasion

Task	Occasion 1 n=68				Occasion 2 n=92			
	Rater 1		Rater 2		Rater 1		Rater 2	
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
Basketball Camp	1.37	0.92	1.44	0.85	1.70	1.18	1.89	1.20
Space Camp	1.26	1.23	1.31	1.12	1.71	1.35	1.59	1.34
Olympics	2.09	1.17	2.53	1.17	2.30	1.19	2.48	1.27
Tug of War	1.57	1.06	1.62	1.05	1.90	1.09	1.93	0.96

Table 2. MPSA scores by task and occasion

Task	Occasion 1 n=68		Occasion 2 n=92	
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
Basketball Camp	1.41	0.84	1.80	1.14
Space Camp	1.29	1.12	1.65	1.30
Olympics	2.31	1.09	2.39	1.19
Tug of War	1.60	1.01	1.92	0.98

**Table 3: Frequency of scores received on the Basketball Camp task for each testing occasion**

Score	Occasion 1		Occasion 2	
	Frequency	Percentage	Frequency	Percentage
0.0	9	13.2	8	8.7
0.5	8	11.8	11	12.0
1.0	10	14.7	11	12.0
1.5	9	13.2	12	13.0
2.0	28	41.2	29	31.5
2.5	2	2.9	5	5.4
3.0	1	1.5	2	2.2
3.5	0	0.0	4	4.3
4.0	1	1.5	10	10.9

**Table 4. Frequency of scores received on the Space Camp task for each testing occasion**

Score	Occasion 1		Occasion 2	
	Frequency	Percentage	Frequency	Percentage
0.0	15	22.1	16	17.4
0.5	14	20.6	12	13.0
1.0	5	7.4	9	9.8
1.5	8	11.8	14	15.2
2.0	18	26.5	20	21.7
2.5	3	4.4	4	4.3
3.0	0	0.0	1	1.1
3.5	0	0.0	3	3.3
4.0	5	7.4	13	14.1

Table 5: Frequency of scores received on the Olympics task for each testing occasion

Score	Occasion 1		Occasion 2	
	Frequency	Percentage	Frequency	Percentage
0.0	3	4.4	6	6.5
0.5	1	1.5	1	1.1
1.0	7	10.3	6	6.5
1.5	6	8.8	12	13.0
2.0	20	29.4	24	26.1
2.5	14	20.6	14	15.2
3.0	2	2.9	3	3.3
3.5	1	1.5	1	1.1
4.0	14	20.6	25	27.2

Table 6. Frequency of scores received on the Tug of War task for each testing occasion

Score	Occasion 1		Occasion 2	
	Frequency	Percentage	Frequency	Percentage
0.0	3	4.4	0	0.0
0.5	2	2.9	0	0.0
1.0	33	48.5	32	34.8
1.5	9	13.2	16	17.4
2.0	5	7.4	19	20.2
2.5	6	8.8	8	8.7
3.0	3	4.4	4	4.3
3.5	2	2.9	3	3.3
4.0	5	7.4	10	10.9

reliability measure given most often for performance assessments. The coefficient of stability is calculated by correlating the scores on each task for the two testing occasions. These coefficients are .28 for the Basketball Camp, .53 for the Space Camp, .40 for the Tug of War task, and .54 for the Olympics task.

### Generalizability Studies

A series of generalizability studies was conducted to determine the amount of variation in the MPSA scores attributable to the object of measurement (students' true score variance) and the amount attributable to the various sources of error. Before the generalizability analyzes were conducted, the pattern of responses of each student in the sample was scrutinized. Students who performed substantially worse during the second administration were identified. There were five students whose difference score on at least one task from the first to the second test administration was greater than or equal to three. These students received either a four on the first test administration and a one or a zero on the second administration, or a three on the first test administration and a zero on the second test administration. It seemed unlikely that these students had lost the ability to perform the tasks, but more likely reflected a loss of motivation. In order to maximize the ability to find the true task effect, these students were deleted from the generalizability studies. Without these students, the sample size was reduced from 73 to 68 on the first occasion, and occasion from 97 to 92 on the second testing.

The first G-study replicates the designs used by Shavelson et al. (1992), Baxter et al. (1992), and Lane et al. (1992). A person by task by rater design was tested for both testing occasion one and testing occasion two. The results are reported in Table 7. An

examination of the estimated variance components indicate that the person by task interaction accounts for nearly two-thirds of the variance on the first testing occasion of the MPSA and over half of the variance on the second testing occasion. This interaction reveals that students' relative performance differed by task and thus performance is quite dependent on the tasks sampled. The error due to the rater main effect as well as the rater by task and the rater by person interaction effects was negligible. These results are consistent with those found in previous studies of both mathematics (Baxter et al., 1992; Lane et al., 1992; Shavelson et al., 1992) and writing performance assessments (Dunbar et al., 1991; Hieronymus & Hoover, 1987). Also noteworthy is the small person effect (i. e., universe score variance) found for the first testing occasion (less than 5%). The person effect increases to nearly one-fourth of the total variance during the second testing occasion.

Table 7. Variance components for the person by rater by task model for each testing occasion

Source	Occasion 1			Occasion 2		
	Estimated Variance Component	Percent Total Variability	Standard Error	Estimated Variance Component	Percent Total Variability	Standard Error
Person (p)	0.059	4.33	0.059	0.372	23.54	0.092
Task (t)	0.190	13.86	0.135	0.101	6.39	0.073
Rater (r)	0.007	0.48	0.010	0.000	0.00	0.003
p x t	0.889	64.89	0.098	0.894	56.58	0.083
p x r	0.019	1.40	0.012	0.007	0.44	0.009
t x r	0.018	1.28	0.013	0.008	0.50	0.006
p x t x r, e	0.188	13.76	0.019	0.198	12.53	0.016

The results of the G-study of the full model (person by task by rater by occasion) using all four tasks are presented in Table 8. Somewhat over one-third of the test score variance is attributed to the person by occasion by task interaction meaning that students performed differentially on the tasks and the pattern of task performance differed by testing occasion. The next highest source of variance was the person by task interaction which accounted for slightly over one-fifth of the total variance. The error due to rater main effects as well as the rater by person, rater by occasion, rater by task, rater by occasion by task, rater by occasion by person and rater by occasion by task interactions was again negligible. The variance attributable to the object of measurement, the student, represented only slightly over twelve percent of the total variance.

Because student performance by task differed substantially, a G-study was conducted using the full model as above but including only the two highly similar tasks of Basketball Camp and Space Camp. The results are shown in Table 9. The percent of variance attributable to the person by occasion by task interaction was very similar to that in Table 8 and remained the single largest source of variance in the ratings. However, the main effect for task decreased from over 10% to less than 1% of the total variance, and the person by task interaction decreased even more dramatically from 22% to less than 1%. Universe score variance (person effect) increased nearly three-fold (from about 12% to over 35%) when only the two highly similar tasks were included. Thus, although differential student ranking by task and occasion remained a large source of variation with two highly similar tasks, other sources of error involving the task effect were minimized and universe score variance increased.

Table 8. Variance components using the full model.

Source	Estimated Variance Component <sup>a</sup>	Percent Total Variance	Standard Error
Person (p)	0.183	12.24	0.068
Occasion (o)	0.031	2.07	0.029
Task (t)	0.158	10.57	0.109
Rater (r)	0.003	0.20	0.005
p x o	0.032	2.14	0.036
p x t	0.332	22.21	0.072
p x r	0.019	1.27	0.009
o x t	0.000	0.00	0.008
o x r	0.000	0.00	0.003
t x r	0.004	0.27	0.007
p x o x t	0.537	35.92	0.062
p x o x r	0.000	0.00	0.008
p x t x r	0.016	1.07	0.013
o x t x r	0.009	0.60	0.007
p x o x t x r,e	0.171	11.44	0.017

<sup>a</sup>Negative variance components replaced with zeros.

Table 9. Variance components for the full model using only the two highly similar tasks of Basketball Camp and Space Camp

Source	Estimated Variance Component <sup>a</sup>	Percent Total Variance	Standard Error
Person (p)	0.465	35.31	0.124
Occasion (o)	0.033	2.51	0.034
Task (t)	0.010	0.76	0.016
Rater (r)	0.000	0.00	0.005
p x o	0.062	4.71	0.081
p x t	0.008	0.61	0.074
p x r	0.018	1.37	0.018
o x t	0.000	0.00	0.007
o x r	0.000	0.00	0.005
t x r	0.002	0.15	0.009
p x o x t	0.520	39.48	0.100
p x o x r	0.025	1.90	0.019
p x t x r	0.031	2.35	0.020
o x t x r	0.011	0.84	0.011
p x o x t x r,e	0.132	10.02	0.023

<sup>a</sup>Negative variance components replaced with zeros.



To help pinpoint the nature of the person by occasion by task interaction, separate person by rater by task analyses were conducted for each testing occasion. The results are shown in Tables 10 and 11. A similar pattern is seen for both testing occasions. Approximately 40% of the variance is attributable to both the universe score variance (the object of measurement) and to the person by task interaction. Although, the person by task interaction accounts for slightly more variance on the first testing occasion, a distinctly different pattern for each testing occasion did not emerge.

### Decision Studies

Decision studies were conducted based on the generalizability studies using two models. Generalizability (relative) and dependability (absolute) coefficients developed from the G-study using the full model with all four tasks were determined first. The results are presented in Tables 12 and 13. The coefficients represent the generalizability of the MPSA results when one or two raters, one or two occasions, and 4 to 32 tasks are used. To reach a relative G-coefficient of .80, approximately 32 tasks, 2 occasions, and 2 raters would be needed. Typically, testing includes only a single administration. To obtain a relative G-coefficient of .80 during only one testing occasion, approximately 175 tasks and 2 raters would be needed. As the contribution of the rater facet to the error term was small, the improvement of the reliability with the addition of a second rater is negligible.

Relative G-coefficients or generalizability coefficients are appropriate when the purpose of testing is to obtain an individual's relative standing among those taking the test.

Table 10. Variance components for the two highly similar tasks of Basketball Camp and Space Camp on the first testing occasion only

Source	Estimated Variance Component	Percent Total Variance	Standard Error
Person (p)	0.406	37.43	0.129
Task (t)	0.000	0.00	0.006
Rater (r)	0.000	0.00	0.001
p x t	0.486	44.82	0.098
p x r	0.021	1.91	0.023
t x r	0.000	0.00	0.000
p x t x r,e	0.171	15.82	0.029

Table 11. Variance components for the two highly similar tasks of Basketball Camp and Space Camp on the second testing occasion only

Source	Estimated Variance Component	Percent Total Variance	Standard Error
Person (p)	0.650	43.88	0.181
Task (t)	0.011	0.75	0.031
Rater (r)	0.000	0.00	0.013
p x t	0.570	38.48	0.111
p x r	0.066	4.48	0.028
t x r	0.029	1.95	0.025
p x t x r,e	0.154	10.47	0.026

Table 12. Generalizability coefficients for the person by task by rater by occasion decision study when all four tasks are included

$n'_t$	$n'_o=1$		$n'_o=2$	
	$n'_r=1$	$n'_r=2$	$n'_r=1$	$n'_r=2$
4	0.39	0.39	0.47	0.49
8	0.50	0.53	0.60	0.63
12	0.57	0.60	0.66	0.70
16	0.61	0.64	0.70	0.74
20	0.64	0.67	0.72	0.76
24	0.66	0.69	0.74	0.78
28	0.68	0.71	0.75	0.79
32	0.69	0.72	0.76	0.80

Table 13. Dependability coefficients for the person by task by rater by occasion decision study when all four tasks are included

$n'_t$	$n'_o=1$		$n'_o=2$	
	$n'_r=1$	$n'_r=2$	$n'_r=1$	$n'_r=2$
4	0.32	0.34	0.40	0.43
8	0.44	0.46	0.53	0.56
12	0.50	0.52	0.59	0.62
16	0.53	0.56	0.63	0.66
20	0.56	0.59	0.66	0.69
24	0.57	0.60	0.67	0.71
28	0.59	0.62	0.69	0.72
32	0.60	0.63	0.70	0.73

When the purpose of testing is to provide information about how much of the domain assessed an individual possesses, an absolute G-coefficient or dependability coefficient is necessary. To obtain an absolute G-coefficient of .80, 200 tasks, 2 raters, and 2 occasions would be needed. It would be practically impossible to obtain an absolute G-coefficient of .80 with only one testing occasion as the coefficient for 1000 tasks with 1 rater is .68 and 1000 tasks with 2 raters is .71.

The second decision study used the full model but included only the two highly similar tasks. The results shown in Tables 14 and 15 give the generalizability and dependability coefficients with one or two raters, one or two occasions, and 4 to 32 tasks. To reach a .80 relative G-coefficient, only 4 to 8 tasks would be needed if the test was given on two occasions and rated by one rater. To reach a .80 relative G-coefficient with only one testing occasion, 20 highly similar tasks would be needed with two raters.

To achieve an absolute G-coefficient of .80, 6 highly similar tasks would be needed if the exam were given on two testing occasions and rated by two raters. However, 200 highly similar tasks would be needed if the exam were given on only one testing occasion and rated by two raters. Note that the magnitude of improvement due to adding tasks declines after about 20 or so tasks.

#### Impact of the Order of Task Presentation

When students received copies of the MPSA, the order of the tasks varied. The relative position of the tasks was counterbalanced so that one-fourth of the students found the Space Camp task first, one-fourth found the Basketball Camp task first, one-fourth found the Tug of War task first, and one-fourth found the Olympics task first. The order

Table 14. Generalizability coefficients for the person by task by rater by occasion decision study based on the two highly similar tasks

$n'_t$	$n'_o=1$		$n'_o=2$	
	$n'_r=1$	$n'_r=2$	$n'_r=1$	$n'_r=2$
4	0.63	0.66	0.75	0.79
8	0.71	0.74	0.81	0.84
12	0.74	0.78	0.83	0.86
16	0.76	0.79	0.85	0.88
20	0.77	0.80	0.85	0.88
24	0.78	0.81	0.85	0.89
28	0.78	0.81	0.86	0.89
32	0.79	0.82	0.86	0.89

Table 15. Dependability coefficients for the person by task by rater by occasion decision study based on the two highly similar tasks

$n'_t$	$n'_o=1$		$n'_o=2$	
	$n'_r=1$	$n'_r=2$	$n'_r=1$	$n'_r=2$
4	0.59	0.63	0.73	0.76
8	0.67	0.70	0.79	0.82
12	0.70	0.73	0.81	0.84
16	0.72	0.75	0.82	0.85
20	0.73	0.76	0.83	0.85
24	0.73	0.77	0.83	0.86
28	0.74	0.77	0.83	0.86
32	0.74	0.77	0.84	0.86

of the second, third, and fourth tasks was counterbalanced in a similar manner. To determine if the order of the presentation of the tasks impacted task performance, two split plot ANOVA's were conducted. Each ANOVA tested an order by task design, one being for the first and the other for the second testing occasion. The assumptions for the split plot design are 1) independent observations, 2) multivariate normality, 3) sphericity, and 4) homogeneity of covariance matrices for the groups (Stevens, 1992). These assumptions are addressed before the results are presented. The independence of observations was insured by the manner in which the data were collected. The ANOVA is robust to the violation of the multivariate normality assumption. The departure from sphericity was very slight for both tests with a Huynh-Feldt Epsilon of .988 and .954 for the first and second testing occasions respectively. The Huynh-Feldt Epsilons were multiplied by the degrees of freedom to correct the F's. The homogeneity of covariance matrices was tested by Box's M and found to be tenable in both designs ( $F$  with 30, 9626 = .947,  $p$  = .548 and  $F$  with 30, 20983 = 1.306,  $p$  = .121). The results of order of presentation by task design are presented in Tables 16 and 17. There is no statistically significant order of presentation main effect as well as no statistically significant order of presentation by task interaction. Therefore, students did not perform differently on the tasks as a result of the order in which they were presented in the MPSA. The significant main effect for task was expected based on the previous generalizability analyses.

Table 16. Order of presentation by task split plot design for the first testing occasion

	SS	df	MS	F
<b>Between Subjects</b>				
Order	20.38	3	6.79	1.37
Error Between	316.25	64	4.94	
<b>Within Subjects</b>				
Task	161.95	3	53.98	13.32*
Task x Order	12.88	9	1.42	0.35
Error Within	778.28	192	4.05	

\* $p < .001$ 

Table 17. Order of presentation by task split plot design for the second testing occasion

	SS	df	MS	F
<b>Between Subjects</b>				
Order	20.31	3.00	6.77	0.68
Error Between	870.44	88.00	9.89	
<b>Within Subjects</b>				
Task	112.61	3.00	37.54	9.64 *
Task x Order	41.29	9.00	4.59	1.18
Error Within	1,027.55	264.00	3.89	

\* $p < .001$

### Performance by Ethnicity and Gender

The average MPSA scores by ethnicity are found in Table 18. Whites tended to score higher than minorities on all tasks. A different pattern of scores between testing occasions was seen as well. The scores of the white students improved on all four tasks between the first and the second testing occasion. However, minority students' average scores stayed the same on the Space Camp task for both testing occasions and decreased from the first to the second testing occasion for the Olympics task.

To determine if the differences in the pattern of scores were large enough to be statistically significant, a split plot ANOVA was computed using a task by testing occasion by ethnicity design. The assumptions of the statistic were examined. Slight departure from sphericity was found for the tests involving task main effects (Huynh-Feldt Epsilon=.878). Sphericity held for the tests involving task by occasion interactions (Huynh-Feldt Epsilon=1.00). Where appropriate, the degrees of freedom were corrected by using the Huynh-Feldt Epsilon. The homogeneity of covariance matrices was not found to be tenable with Box's M (F with 36,3639= 1.758, p=.003). As the larger variance is found in the group with the largest number of students (whites), the test is conservative. Table 19 gives the results of the split plot analysis. Significant main effects are found between ethnic groups, testing occasions, and among tasks. No significant interactions were detected. Second occasion test scores were higher than first occasion test scores and scores for whites were higher than for minorities. As there are four tasks, follow up tests are needed to see where the task differences occurred. For split-plot designs Stevens (1992) recommends using multiple t-tests and using the Bonferroni inequality to keep overall alpha under control when group sizes are unequal.



Table 18. MPSA scores by ethnicity

Task	Occasion 1				Occasion 2			
	Majority n=50		Minority n=18		Majority n=71		Minority n=21	
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
Basketball Camp	1.54	0.83	1.03	0.80	1.91	1.15	1.41	1.05
Space Camp	1.33	1.07	1.17	1.30	1.79	1.28	1.17	1.28
Olympics	2.43	1.19	1.97	0.70	2.55	1.22	1.86	0.98
Tug of War	1.71	1.07	1.28	0.81	1.99	1.01	1.67	0.84

Table 19. Task by occasion by ethnicity analysis

	SS	df	MS	F
<b>Between Subjects</b>				
Ethnic	92.07	1	92.07	8.57*
Error Between	709.05	66	10.75	
<b>Within Subjects</b>				
Task	192.11	3	64.04	12.20**
Ethnic x Task	6.15	3	2.05	0.39
Task Within Error	1,039.25	198	5.25	
Occasion	21.95	1	21.95	7.22*
Ethnic x Occasion	2.37	1	2.37	0.78
Occasion Within Error	200.66	66	3.04	
Task x Occasion	13.52	3	4.51	1.82
Ethnic x Task x Occasion	5.07	3	1.69	0.68
Task x Occasion Within Error	491.74	198	2.48	

\* $p < .01$ \*\* $p < .001$

Six dependent t-tests were computed. Statistical significance was found when alpha was less than or equal to .05/6 or .008. Students scored significantly higher on the Olympics task than on the other three tasks. No differences were found between the Space Camp, the Basketball Camp, and the Tug of War tasks. Table 20 illustrates statistically significant differences between tasks.

Table 20. Results of dependent t-tests as follow-up to significant task main effect

	Olympic	Tug of War	Basketball Camp
Olympic			
Tug of War	**		
Basketball Camp	**	n.s.	
Space Camp	**	n.s.	n.s.

\*\* $p < .001$

The average score on each MPSA task by gender is found in Table 21. On both testing occasions female scores were higher on three of the four tasks. On the Tug of War task, male scores were higher on both occasions. To see if score differences were large enough to rule out chance occurrence, a task by occasion by gender split plot design was computed. The assumptions of the statistic were again examined. The homogeneity of covariance assumption was tenable as shown by Box's M ( $F(36,9568) = .683, p = .925$ ). Slight departure from sphericity was noted again for the test involving task main effects (Huynh-Feldt Epsilon = .878). Sphericity was found for the tests involving task by occasion interactions (Huynh-Feldt Epsilon = 1.00). Where appropriate, the degrees of freedom were corrected. The results are presented in Table 22. Again, a significant main

Table 21. MPSA scores by gender

Task	Occasion 1				Occasion 2			
	Female n=26		Male n=42		Female n=38		Male n=54	
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>
Basketball Camp	1.46	0.84	1.37	0.86	2.11	1.15	1.58	1.10
Space Camp	1.91	1.17	0.91	0.92	1.99	1.17	1.41	1.34
Olympics	2.46	1.12	2.22	1.08	2.52	1.14	2.31	1.24
Tug of War	1.52	1.12	1.65	0.97	1.81	1.01	2.00	.97

Table 22. Task by occasion by gender analysis

	SS	df	MS	F
<b>Between Subjects</b>				
Gender	49.05	1	49.05	4.30*
Error Between	752.07	66	11.40	
<b>Within Subjects</b>				
Task	231.24	3	77.08	15.60**
Gender x Task	67.24	3	22.41	4.54*
Task Within Error	978.16	198	4.94	
Occasion	35.84	1	35.84	11.65*
Gender x Occasion	0.01	1	0.01	0.00
Occasion Within Error	203.02	66	3.08	
Task x Occasion	12.27	3	4.09	1.69
Gender x Task x Occasion	16.60	3	5.03	2.28
Task x Occasion Within Error	480.21	198	2.43	

\* $p < .01$ \*\* $p < .001$

effect was found for testing occasion. A significant gender by task interaction was found, so the significant main effect for gender was not discussed. No other significant interactions were detected.

Follow up tests were run for the gender by task interaction to see where the differences occurred in scoring patterns among males and females. Four independent t-tests were calculated using gender as the independent variable. Differences were determined to be significantly different when the probability of Type I error was less than .05/4 or .0125. The only difference in scores between males and females was found on the Space Camp task where females scored statistically significantly higher than males. See Table 23 below.

Table 23. Independent t-tests as follow up to the gender by task interaction

Task	Female		Male		t
	<u>M</u>	<u>SD</u>	<u>M</u>	<u>SD</u>	
Olympic	2.45	1.03	2.29	1.06	0.64
Tug of War	1.70	0.87	1.86	0.83	-0.77
Basketball Camp	1.83	0.81	1.43	0.71	2.14
Space Camp	1.90	1.01	1.07	0.92	3.49*

\*p < .05/4 or .0125

### Summary

Evidence of the reliability and validity of the tasks on the MPSA were presented. Classical reliability coefficients indicate inter-rater reliability coefficients averaging in the .80's. However, test-retest reliability coefficients are substantially lower ranging from .28 to .54. G-study results show substantial sources of error due to task sampling, person by

task interaction, and person by occasion by task interaction. True score variance accounts for less than 15% of the total score variance. When only the two highly similar tasks are used in the G-study, a large person by occasion by task interaction is still observed although other task effects decreased. The true score variance increases to approximately one-third of the total variance.

The D-study results indicate that for the MPSA to reach a relative generalizability coefficient of .80, 32 tasks would need to be given on two occasions and rated by two raters if the tasks were not highly similar but measured the same construct. A relative generalizability coefficient of approximately .80 could be obtained on one testing occasion if 20 highly similar tasks were included and rated by two raters.

The performance on the tasks differed somewhat by ethnicity and gender. White students scored higher than minority students. The only statistically significant difference in task performance by gender occurred on the Space Camp task where females obtained higher scores than males.

## CHAPTER V

### SUMMARY AND CONCLUSIONS

#### Summary of the Investigation

This study examined evidence of the reliability and validity of four performance assessment tasks on the MPSA. The tasks were chosen to represent selected process and content skills from the mathematics portion of the PASS, the state-mandated curriculum developed by the Oklahoma State Department of Education (1993). Two tasks were designed to be highly similar. The other two tasks sampled the same content domain but were not highly similar tasks.

The most frequently reported evidence of the validity and reliability of performance assessment instruments are inter-rater reliability and content validity (Johnson et al., 1993; Koretz, 1993; Lane, 1993). Evidence of each was obtained for the tasks on the MPSA. Four eighth grade mathematics teachers and two school district mathematics curriculum consultants found the tasks to represent the process skill of using mathematics as problem-solving as described in the PASS for eighth grade mathematics. The mathematics educators also concurred that the content skills required by the tasks were found to be a subset of those content areas outlined in the PASS for eighth grade mathematics. The tasks included only the content areas of number sense and number theory, computation and estimation, patterns and functions, and possibly algebra concepts but did not include statistics, probability, geometry, or measurement. These mathematics subject matter

experts unanimously certified that the tasks on the MPSA were valid measures of the State of Oklahoma learner outcomes for eighth grade mathematics.

Inter-rater reliability was found to be in the acceptable range of .77 to .87 for the four tasks. If the analysis of the reliability and validity of the MPSA had stopped at this point, it would have been concluded that there was evidence to support the reliability and validity of the MPSA for use with students like those in the sample used in the present study. However, consistent with previous research on performance assessment (Baxter et al., 1993; Dunbar et al., 1991; Huot, 1990a; Shavelson et al., 1993), a large percent of error associated with differential response by task was found. Another identified source of error was differential response by testing occasion. Decision studies were conducted to find out how many raters, occasions, and, tasks would be necessary to generalize the test results to the domain of interest.

Another validity criteria to be investigated regarding the performance assessment tasks, is the fairness of each task to all groups of students. To obtain an estimate of this type of evidence for validity, analysis of variance was used to determine if students' scores differed by gender or ethnic groups.

## Discussion

### Sources of Score Variation

The generalizability study using only task and rater as sources of error revealed that universe score (true score) variance accounted for less than 5% of total variance on the first testing and less than one-fourth of the total variance on the second testing occasion. The most potent contributor to error variance was the person by task

interaction which comprised nearly two-thirds of the total variance on the first testing occasion and over half of the total variance on the second testing occasion. The interaction indicates that students' relative performance differed by task. The error associated with the rater main effect and its interactions with the other effects were negligible.

The generalizability study using the full model of all three sources of error (rater, task, and occasion) simultaneously found the person effect to be 12% of the total variance. This leaves 88% of the total to error variance. The most potent error source was the person by occasion by task interaction (over one-third of total variance). Students performed differently depending on the task given, and the pattern of response differed by testing occasion. The next largest source of variation was the person by task interaction (nearly one-fourth of the total variance) meaning that students performed differently on different tasks. The task main effect (ten percent of total variance) indicated that the tasks differed in difficulty. The error due to rater or interactions among rater and other effects was again negligible.

A unique aspect of this study was the inclusion of two highly similar tasks. A generalizability study was conducted to determine the sources of score variation with the full model using the three sources of error but including only the two highly similar tasks. With this model the person effect increased to about one-third of the total variance with error variance accounting for the remaining two thirds of the total. The most potent error component was again the person by occasion by task interaction (about 40% of the total variance). All other sources of explained error were small including the person by task



interaction and the task main effect. Ten percent of the variance was attributed to the person by occasion by task by rater interaction confounded with random error. When this model was run separately for each occasion, a large person by task interaction was found on each occasion. So, persons performed differently by task even when the tasks were highly similar.

These generalizability studies indicate that only a small percent of variance in students' scores on the four tasks of the MPSA can be attributed to universe score (true score). A much larger percent of variance can be attributed to error. Within the error component, the task effect and the occasion effect are the largest contributors due to large task by occasion by person, and task by person interactions as well as a nontrivial task main effect. The rater effect was consistently negligible. The large task effect and small rater effect are consistent with the findings of Baxter et al. (1993) and Shavelson et al. (1993) regarding mathematics performance assessments. The large person by task by occasion interaction is consistent with the results found by Shavelson et al. when examining the contributions of task, rater, and occasion facets to error of science performance assessments.

The magnitude of the effects in this study is similar to those found in other experiments. Shavelson et al. found that the person by task by occasion effect accounted for 59% of the total variance and the person by task effect for 32% of the total variance in science performance assessment scores. Likewise, Baxter et al. (1993) found that the person by task effect accounted for 48% of the total variance in mathematics performance assessment scores.

### Number of Raters, Occasions, and Tasks Necessary to Generalize

In response to the parameters established by the generalizability studies, the question was asked how many raters, occasions, and tasks are necessary to generalize the test results to the domain of interest. The answers vary depending on whether or not highly similar tasks are used.

If using tasks that are not highly similar but do assess the same construct, an estimated 32 tasks given on two testing occasions and rated by two raters or 200 tasks given on one testing occasion and rated by two raters would be needed to obtain a relative G-coefficient of .80. Relative G-coefficients provide the generalizability of tests given for the purpose of ranking students for norm-referenced testing purposes. If the test is used to determine how much of a construct is possessed by a student, as in criterion-referenced testing situations, an absolute G-coefficient would be needed to determine the generalizability of scores. To obtain an absolute G-coefficient of .80 with tasks that are not designed to be highly similar, 200 tasks given on two testing occasions and rated by two raters would be needed. An absolute G-coefficient of .80 could not be obtained on one testing occasion even with 1000 tasks rated by two raters.

Obtaining a relative G-coefficient of .80 with highly similar tasks would require five tasks given on two testing occasions and rated by one rater, or 20 tasks given on one occasion and rated by two raters. An absolute G-coefficient of .80 could be obtained with six highly similar tasks given on two occasions and rated by two raters or 200 highly similar tasks given on one occasion and rated by two raters.

A test composed of many highly similar performance assessment tasks would not be the optimal way to measure how well eighth grade students could perform the PASS. The state learning outcomes are intended to generalize to many different mathematics problem-solving situations. It would be impossible to identify a single set of task specifications that represent the entire domain. Highly similar tasks were used in this study to obtain a baseline task effect. The baseline was compared to the task effect of less similar tasks. When highly similar tasks are given on two testing occasions, the five tasks necessary for a relative G-coefficient of .80 would require approximately one hour and fifteen minutes on each occasion. However, if the highly similar tasks were given on only one testing occasion as is typically done, the 20 tasks needed for a relative G-coefficient of .80 would require five hours of testing. An absolute G-coefficient of .80 could be obtained in one and one-half hours of testing (six tasks) on two occasions or 50 hours of testing (200 tasks) on one testing occasion.

Performance assessment tasks that are not highly similar would most likely be given to assess the degree to which students accomplished the state learner outcomes for mathematics. The testing time to obtain a relative G-coefficient of .80 would require eight hours on two testing occasions (32 tasks) or fifty hours on one testing occasion (200 tasks). To obtain an absolute G-coefficient of .80 would require fifty hours of testing on two testing occasions. The length of these testing periods point to the difficulty of obtaining performance assessment task results that can generalize to the domain of interest.

### Performance by Gender and Ethnicity

The analysis of the differential performance by gender and ethnic groups was conducted to provide preliminary evidence of the fairness of the performance assessment tasks to all students. These results need to be considered merely as indicators of potential threats to the validity of the tasks as students of each subgroup were not matched by ability before differences in obtained scores were analyzed.

When scores were compared between males and females, no statistically significant difference was found between performance on the Basketball Camp task, the Olympics task or the Tug of War task. Females outscored the males only on the Space Camp task. It is unknown what characteristics of the Space Camp task made it easier for females.

Scores were also compared between majority and minority students. Results indicated that majority students scored higher on each performance assessment task. These results concur with those found by Linn et al. (1991) and Baxter et al. (1993).

Part of the test construction process included the examination of each performance assessment task by two experts in multicultural education. These experts found no cultural bias in the MPSA. While difference in scoring patterns may or may not prove that the performance assessment task were unfair to minority or to male students, the results indicate that further study is necessary to ensure that score differences represent true differences in the amount of the construct possessed and are not due to construct irrelevance (Messick, 1989). Construct irrelevance may occur if the performance assessment tasks required abilities outside of the mathematics domain such as written expression. If students differed in scores on the mathematics performance

assessments due to differing abilities to express themselves in writing instead of differing abilities in mathematics problem-solving, then the test would lack construct validity due to construct irrelevance. The tasks on the MPSA required students to write a paragraph to explain how they arrived at their answer. Raters were instructed to disregard errors in writing or sentence construction. Instead, the raters were to focus solely on the content of students' responses. Nevertheless, students who were unable to express their mathematical reasoning in written form were at a disadvantage on the MPSA tasks.

### Conclusions

This study addressed two of the criteria proposed by Baker et al. (1993) and Quallmalz (1991) for determining the validity of performance assessments: tasks and scoring are fair to all students, and scores on each task are similar to the scores on other tasks within the same domain. Evidence was not found to support either the fairness of the task to all students or the generalizability of the results of the performance assessment tasks as a true representation of the universe score composed of all possible tasks measuring the same construct.

The contributions of the field of cognitive psychology change the way we look at how students learn and point to the need to reform instructional practices. The increased awareness of the limitations of the multiple-choice tests to measure learning provide a strong rationale for the need to pursue the use of alternative assessment methods such as performance assessments. Unfortunately the field of educational measurement is only beginning to develop validation criteria and techniques for performance assessments.

The use of performance assessments is expanding in high stakes testing situations. Therefore, evidence of reliability and validity is crucial. Since performance assessment tasks measure a more integrated ability to solve problems, they are being touted as the most appropriate assessment methods. A number of curriculum specialists (Crehen, 1991; Newmann, 1993; Stenmark, 1991) and policy makers (Aschbacher, 1991; Office of Technology Assessment, 1992) rely on the high content validity of an "authentic" assessment and so do not question whether the test results will generalize to other tasks within the same content domain. The results of this study do not support this practice. Messick (1989) states that construct validity may be lacking even when high content validity is found if the items contain construct irrelevant areas or if the particular group of test items underrepresent the construct. When test takers respond differently to tasks of the same content domain, then the likelihood of high construct irrelevance or construct underrepresentation is evident. The lack of generalizability of the MPSA tasks used in this study indicate a similar lack of construct validity.

The advances in cognitive psychology are the impetus for the recommendations to make complex mathematical problem-solving the instructional target. However, Miller and Legg (1993) assert that linking performance assessment to instruction is more difficult than linking multiple-choice test objectives to instruction because of the complexity of teaching higher order thinking skills. The test scores on the the MPSA tasks are low even though the study participants' mathematics teacher indicated that she had begun implementing new teaching techniques in line with the suggestions made by the Working Groups of the Commission on Standards for School Mathematics of the National Council

of Teachers of Mathematics (1989). One explanation for the low test scores is the lack of blueprints for teaching complex mathematical problem-solving. Messick (1989) maintains that learning theory has yet to establish methods for teaching expertise. It is also possible that the reform-based teaching strategies impact different students in different ways. Baxter et al. (1993) found that a reform-based curriculum improved White students' performance assessment scores more than Latino students' performance assessment scores. Linn et al. (1991) and Miller and Legg (1993) indicate that the use of performance assessment in high stakes testing is probably not appropriate until stronger relationships can be established between instruction and success on performance assessment. The results of this study support their conclusion.

The small sample size used in this study could be construed as a limitation. Small sample size is common in generalizability studies such as this one in which rater is used as a facet. Interestingly, large person by task and person by task by occasion effects are found consistently. However, it is possible that studies with a larger number of participants would yield more stable task effects.

The reform initiatives proposed by the Working Group of the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics (1989) seem to call for new ways to assess student achievement. Performance assessment tasks appear to meet the need of measuring how well students can solve complex problems requiring mathematics. However, such measures will be fair and credible only to the degree to which adequate evidence of validity and reliability is obtained. The findings of substantial task and occasion sampling variability and of possible differential sampling

by gender and ethnicity do not support the use of performance assessment tasks like those on the MPSA in high stakes testing situations.

### Recommendations

In light of the findings of this study, the following recommendations are offered.

1. The State of Oklahoma would be well advised to refrain from including performance assessment tasks such as those on the MPSA in any statewide assessment plan until more convincing data are in hand.
2. The source of measurement error contributed by raters is consistently low when rater training is sufficient and including a rater facet in the generalizability design limits the sample size. Therefore, a study should be conducted in which a person by task by occasion design is employed and sample size is increased from the typical 100 participants.
3. Future studies addressing the fairness of mathematics performance assessment tasks to all groups of participants should match students of different groups on a measure of mathematics achievement before comparisons are made on performance assessment scores.
4. As yet it is unknown whether or not similar results to those of this study would be found for older students, younger students, or for students of varying socioeconomic levels. Future studies should use participants in high school and elementary grades as well as students of high, medium, and low socioeconomic status to determine if results are consistent.



## REFERENCES

- Archibald, D. A., & Newmann, F. M. (1988). Beyond standardized testing. Reston, VA: National Association of Secondary School Principals.
- Ary, D., Jacobs, L. C., & Razavieh, A. (1990). Introduction to research in education (4th ed.). Fort Worth: Harcourt Brace Joanovich College Publishers.
- Aschbacher, P. R. (1991). Performance assessment: State activity, interest, and concerns. Applied Measurement in Education, 4 (4), 275-288.
- Baker, E. L. (1992). The role of domain specifications in improving the technical quality of performance assessment. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Baker, E. L., & Herman, J. L. (1983). Task structure design: Beyond linkage. Journal of Educational Measurement, 20 (2), 149-164.
- Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. American Psychologist, 48 (12), 1210-1218.
- Baron, J. B., Forgione, P. D. Jr., Rindone, D. A., Kruglanski, H., & Duey, B. (1989, April). Toward a new generation of student outcome measures: Connecticut's common core of learning assessment. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Baxter, G. P. (1992). Exchangeability of science performance assessments. Dissertation Abstracts International, 53 (2), 474.

- Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. Journal for Research in Mathematics Education, 24 (3), 190-216.
- Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment, (pp. 1-28). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., & Spohrer, J. C. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice & open-ended items. Applied Psychological Measurement, 14 (2), 151-162.
- Blumberg, F., Epstein, M., MacDonald, W., & Mullis, I. (1986). A pilot study of higher-order thinking skills assessment techniques in science and mathematics. Princeton, NJ: National Assessment of Educational Progress.
- Bock, R. D. (1991). The graded mark-point method of scoring performance exercises and open-ended items. University of Chicago: Author.
- Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27 (2), 165-174.
- Boodoo, G. M., & Garlinghous, P. (1983). Use of the essay examination to investigate the writing skills of undergraduate education majors. Educational and Psychological Measurement, 43, 1005-1014.

- Braswell, J., & Kupin, J. (1993). Item formats for assessment in mathematics. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment, (pp. 167-182). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Breland, H. M., & Gaynor, J. L. (1979). A comparison of direct and indirect assessments of writing skill. Assessment of Writing Skill, 16 (2), 119-128.
- Breland, H. M., & Griswold, P. A. (1981). Group comparisons for basic skills measures. New York: College Entrance Examination Board.
- Brennan, R. L. (1982). Elements of generalizability theory. Iowa City: The American College Testing Program.
- Burns, M. (1992). About teaching mathematics. Sausalito, CA: Math Solutions Publications.
- Cantor, N. K., & Hoover, H. D. (1986, April). The reliability and validity of writing assessment: An investigation of rater, prompt within mode, and prompt between mode sources of error. Paper presented at the Annual meeting of the American Educational Research Association, San Francisco.
- Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. Journal for Research in Mathematics Education, 15, 179-202.
- Chase, K. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), Visual information processing (pp. 215-281). New York: Academic Press.

- Crehan, K. (1991, October). Performance assessment: Comparative advantages. Paper presented at the Annual Meeting of the Arizona Educational Research Organization, Flagstaff.
- Crick, G. E., & Brennan, R. L. (1983). Manual for GENOVA: A generalized analysis of variance system (ACT Technical Bulletin No. 43). Iowa City, IA: American College Testing Program.
- Crocker, L., & Algina, J. (1986). Introduction to classical & modern test theory. Fort Worth: Holt, Rinehart, & Winston, Inc.
- Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in Judgment of Writing Ability (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service.
- Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4 (4), 289-303.
- Ebel, R. L. (1984). Achievement test items: Current issues. In B.S. Plake (Ed.), Social & technical issues in testing implications for test construction and usage. (pp. 141-155). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Far West Laboratory for Educational Research and Development (1992). Alternative assessment projects. San Francisco: Author.
- Fenberg, L. (1990). Multiple-choice and its critics. The College Board Review, No. 157.
- Fredereksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18 (9), 27-32.

- Glaser, R. (1988). "Cognitive and environmental perspectives on assessing achievement." In Assessment in the Service of Learning: Proceedings of the 1987 ETS Invitational Conference. Princeton, NJ: Educational Testing Service.
- Gitomer, D. H. (1993), Performance assessment and educational measurement. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing and portfolio assessment (pp. 241-264). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hartle, T. W., & Battaglia, P. A. (1993). The federal role in standardized testing. In R. E. Bennett & W.C. Ward (Eds.), Constructing Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment (pp. 241-264). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hieronymous, A. N., & Hoover, H. D. (1987). Iowa test of basic skills: Writing supplement teachers guide. Chicago: Riverside.
- Houston, W. M., Raymond, M. R., & Svec, J. C. (1991). Adjustments for rater effects in performance assessment. Applied Psychological Measurement, 15 (4), 409-421.
- Huot, B. (1990a). The literature of direct writing assessment: Major concerns & prevailing trends. Review of Educational Research, 60 (2), 237-263.
- Huot, B. (1990b). Reliability, validity, and holistic scoring: What we know and what we need to know. College composition and communication, 41 (2), 201-213.

- Johnson, E. G., Mazzeo, J. & Kline, D. L. (1993). Technical report of the NAEP 1992 trial state assessment program in mathematics. Washington, DC: U.S. Department of Education.
- Kane, M. T. (1982). A sampling model for validity. Applied Psychological Measurement, 6 (2), 125-160.
- Kegley, P. H. (1986). The effect of mode of discourse on student writing performance. Implications for policy. Educational Evaluator and Policy Analysis, 8, 147-154.
- Koretz, D. (1993, January). New report on Vermont portfolio project documents challenges. National Council on Measurement in Education Quarterly Newsletter, 1 (4), 1-2.
- Lane, S. (1993). The conceptual framework for the development of a mathematics performance assessment instrument. Educational Measurement: Issues & Practice, 12 (3), 16-23.
- Lane, S. (1989). Implications of cognitive psychology for measurement and testing: Diagnosis of procedural errors. Educational Measurement: Issues and Practice, 8 (4), 17-20.
- Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1992, April). Empirical evidence for the reliability and validity of performance assessments. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

- Lane, S., Parke, C., & Moskal, B. (1992, April). Principles for developing performance assessments. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.
- Linn, R. L., Baber, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20, 15-21.
- Masters, G. N., & Mislevy, R. J. (1993). New views of student learning: Implications for educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests (pp. 218-242). Hillsdale, NJ: Lawrence Erlbaum Associates.
- McNeil, L. M. (1988). Contradictions of control, Part 3: Contradictions of reform. Phi Delta Kappan, 69, 478-485.
- Mehrens, W. A. (1991, April). Using Performance Assessment for Accountability Purposes: Some Problems. Abridged from a paper presented as part of a symposium: Frechling, J. Performance assessment and accountability programs: Match or mismatch, given at the Annual Meeting of the American Educational Research Association, Chicago.
- Meredith, V., & Saunders, J. (1984). The effects of prompt length, modes of discourse and grade level on student writing performance. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

- Meredith, V. H., & Williams, P. L. (1984). Issues in direct writing assessment: Problem identification & control. Educational measurement: Issues & Practices, 3 (1), 11-15, 35.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed.). (pp. 13-103). New York City: American Council on Education and MacMillan Publishing Company.
- Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. Applied Measurements in Education, 3 (3), 285-296.
- Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stake environment. Educational Measurement: Issues & Practices, 12 (2), 9-15.
- Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I.I. Bejar (Eds.), Test theory for a new generation of tests, (pp. 19-40).
- Newmann, F. M. (1993). Beyond common sense in educational restructuring, the issues of content & linkage. Educational Researcher, 22 (2), 4-13, 22.
- Office of Technology Assessment. (1992). Performance assessment: Methods and characteristics. Washington DC: Congress of the United States. (ERIC Document Reproduction Service N. ED 340 770).
- Oklahoma State Department of Education. (1993). A core curriculum for our children's future: Pass. Oklahoma City: Author.
- Phillips, S.E. (1993). Legal issues in performance assessment. Educator Law Report, 709-738.



- Purves, A. C. (1992). Reflections on research and assessment in written composition. Research in the Teaching of English, 26 (1), 108-123.
- Putnam, R. T., Lampert, M., & Peterson, P. (1990). Alternative perspectives of knowing mathematics in elementary schools. In C. B. Cazden (Ed.), Review of Research in Education. (pp. 57-150). Washington D.C.: American Education Research Association.
- Quallmalz, E. S. (1991). Developing criteria for performance assessments: The missing link. Applied Measurement in Education, 4 (4), 319-331.
- Quallmalz, E. S., & Capell, F. (1979). Defining writing domains: Effects of discourse and response mode. Los Angeles: University of California at Los Angeles, Center for the Study of Evaluation. (ERIC Document Reproduction Service N. ED 212 661).
- Quallmalz, E. S., Capell, F. J., & Chou, C. (1982). Effects of discourse & response mode on the measurement of writing competence. Journal of Educational Measurement, 19 (4), 241-258.
- Resnick, L.B. (1987). Education and learning to think. Washington DC: National Academy Press.
- Resnick, L. B., & Klopfer, L. E. (1989). Toward the thinking curriculum: An overview. In R. B. Resnick & L. E. Klopfer (Eds.), Toward the thinking curriculum: Current cognitive research. 1989 ASCD Yearbook. (pp. 1-18). Alexandria, VA: Association for Supervision and Curriculum Development.

- Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford and M.C. O'Connor (Eds.), Changing assessments: Alternative views of aptitude, achievement and instruction, (pp. 37-75). Boston: Kluwer.
- Robinson, S. P. (1993). The politics of multiple-choice versus free-response assessment. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment, (pp. 313-324). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. New York: Academic Press.
- Romberg, T. A. (1989). Evaluation: A coat of many colors. In D. F. Robitaille (ed.), Evaluation and assessment in mathematics education, (pp. 10-24). Vancouver, Canada: Science and Technology Education, University of British Columbia.
- Romberg, T. A., Wilson, L., & Khaketla, M. (1989). The alignment of six standardized tests with the NCTM standards. Unpublished paper. Madison, WI: University of Wisconsin.
- Royer, J. M., Cisero, C. A., & Carlo, M. S. (1993). Techniques & procedures for assessing cognitive skills. Review of Educational Research, 63 (2), 201-243.
- Ruiz-Primo, M. A., Baxter, G. P., & Shavelson R. J. (1993). On the stability of performance assessments. Journal of Educational Measurements, 30, 41-53.
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessment. Journal of Educational Measurement, 30 (3), 215-232.

- Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. Applied Measurement in Education, 4, 347-362.
- Shavelson, R. J., Mayberry, P., Li, W., & Webb, N. M. (1990). Generalizability of military performance measurements: Marine corp infantrymen. Military Psychology.
- Shavelson, R. J., McDonnell, L., Oakes, J., & Carey, N. (1987, August). Indicator systems for monitoring mathematics and science education. (Report No. R3570-NSF). Washington DC: National Science Foundation.
- Shavelson, R. J., & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publications.
- Shepard, L. (1989). Why we need better assessments. Educational Leadership, 46 (7), 4-9.
- Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. Educational Researcher, 20 (5), 8-11.
- Snow, R. E., & Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), Test theory for a new generation of tests, (pp. 1-18). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational measurement, (3rd ed.), (pp. 263-331). New York: NY. American Council on Education and MacMillan Publishing Company.

- Stenmark, J. K. (Ed.) (1991). Mathematics assessment: Myths, models, good questions, and practical suggestions. Reston, VA: The National Council of Teachers of Mathematics.
- Stevens, J. (1992). Applied multivariate statistics for the social sciences. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Stevenson, Z. Jr., Averett, C. P., & Vickers, D. (1990, April). The reliability of using a focused-holistic scoring approach to measure student performance on a geometry proof. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.
- Stiggins, R. J. (1987). Design & development of performance assessments. Educational Measurement: Issues & Practices, 6 (1), 33-42.
- Swanson, P., Norcini, J., & Grasso, L. (1987). Assessment of clinical competence: Written and computer-based simulations. Assessment and Evaluation in Higher Education, 12, 220-246.
- Tatsuoka, K. K. (1993). Item construction and psychometric models appropriate for constructed responses. In R. E. Bennett & W. C. Ward (Eds.), Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment, (pp. 107-134). Hillsdale, NJ: Lawrence Erlbaum Associates.
- White, E. M. (1985). Developing successful college writing programs. San Francisco: Jossey-Bass Publishers.

Wiggins, G. (1989, April). Teaching to the (authentic) test. Educational Leadership, 46 (7), 41-47.

Wolf, D., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), Review of Research in Education. (pp. 31-74). Washington, DC: American Educational Research Association.

Working Groups for the Commission on Standards for School Mathematics of the National Council of Teachers of Mathematics, (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: National Council of Teachers of Mathematics.

**APPENDICES**

**APPENDIX A**  
**PERFORMANCE TASKS ON THE MATHEMATICS**  
**PROBLEM-SOLVING ASSESSMENT**

**MATHEMATICS PROBLEM-SOLVING ASSESSMENT**

**Directions:** Each task requires you to solve a problem. After you determine your answer, write a paragraph to explain how you got your answer. Your paragraph should be clear enough so that another person can read it and understand the reasoning and steps you used to solve the problem. It is important that you show all your work including calculations used to determine your answer.

**TASK 1.** Treena won a 7-day scholarship worth \$1,000 to the Pro Shot Basketball Camp. Round trip travel expenses to the camp are \$335 by air or \$125 by train. At the camp she must choose between a week of individual instruction at \$60 per day or a week of group instruction at \$40 per day. Treena must spend \$45 per day for food and other expenses. If she does not plan to spend any money other than the scholarship, what are all choices of travel and instruction plans that she could afford to make? Write a paragraph to explain how you arrived at your answer. You may include diagrams or pictures in your paragraph if you like.

**TASK 2.** Damon's friend gave him \$1,500 to attend Space Camp for 14 days. Round trip expenses to the camp are \$435 by air or \$225 by bus. At the camp, he must choose between two weeks of general camp experience at \$30 per day or a two-week simulated space trip experience for \$75 per



day. Damon must spend \$25 per day for food and other expenses. If he does not plan to spend any money other than that given by his friend, what are all choices of travel and camp experiences that he could afford to make? Write a paragraph to explain how you arrived at your answer. You may include diagrams or pictures in your paragraph if you like.

**TASK 3.** Use the information given to figure out who will win the third round in a tug-of-war.

**Round 1:** On one side are four acrobats, each of equal strength. On the other side are five neighborhood grandmas, each of equal strength. The result is dead even.

**Round 2:** On one side is Ivan, a dog. Ivan is pitted against two of the grandmas and one acrobat. Again, it's a draw.

**Round 3:** Ivan and three grandmas are on one side and the four acrobats are on the other.

Who will win the third round? Write a paragraph to explain how you arrived at your answer. You may include diagrams or pictures in your paragraph if you like.

**TASK 4:** Joe, Sarah, Jose, Zabi, and Kim decided to hold their own Olympics after watching the Olympics on TV. They needed to decide what events to have at their Olympics. Joe and Jose wanted a weight lift and a frisbee toss event. Sarah, Zabi, and Kim thought running a race would be fun. The

children decided to have all three events. They also decided to make each event of the same importance.

Name	Frisbee Toss	Weight Lift	50-Yard Dash
Joe	40 yards	205 pounds	9.5 seconds
Jose	30 yards	170 pounds	8.0 seconds
Kim	45 yards	130 pounds	9.0 seconds
Sarah	28 yards	120 pounds	7.6 seconds
Zabi	48 yards	140 pounds	8.3 seconds

Who would be the all-around winner? Write a paragraph to explain how you arrived at your answer. You may include diagrams or pictures in your paragraph if you like.

Source:

Task 1 is taken from the sample extended constructed-response item shown in the Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics (Johnson et al., 1993, p. 112).

Task 2 is written to be a highly similar item to task 1.

Task 3 is taken from About Teaching Mathematics (Burns, 1992, p. 110).

Task 4 is taken from the National Assessment of Educational Progress's Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics (Blumberg et al., 1986, p. 18).

**APPENDIX B**

**SCORING RUBRICS FOR THE**

**MATHEMATICS PROBLEM-SOLVING ASSESSMENT**

SCORING RUBRICS FOR THE  
MATHEMATICS PROBLEM-SOLVING ASSESSMENT

Space Camp and Basketball Camp

Score 4 for the correct solution and an explanation that includes the cost of daily expenses, camp option expenses, and transportation expenses. No prose is required in the explanation.

Score 3 for the correct mathematical evidence that Treena (Damon) has 3 (2) options, but the explanation is unclear or incomplete.

Score 2 for a response that indicates one or more correct conclusions without complete explanation. Additional supporting computations beyond a score of 1 must be present. The work may contain computational flaws. A score of 2 is also given if the student has the correct mathematics for 1 or more options but indicates no conclusions.

Score 1 for a response indicating valid conclusions with no mathematical evidence, or some correct mathematics beyond computing fixed costs ( $7 \times 45 = 315$  or  $14 \times 25 = 350$ ) but no conclusions indicated. A score 1 response may contain major mathematical errors or flaws in reasoning. For example, the student does not consider the fixed expenses or realize that camp daily expenses must be multiplied by the number of days at camp.

Score 0 for a response that is completely incorrect, irrelevant or off task. Just computing  $7 \times 35 = 315$  or  $14 \times 25 = 350$  is a score of 0.

Source:

These rubrics were adapted from those presented in the Technical Report of the NAEP 1992 Trial State Assessment Program in Mathematics (Johnson et al., 1993, p. 112).

## Olympics

Score 4 for a accurate ranking of the children's performance on each event and citing Zabi as the overall winner

Score 3 for using a ranking approach to evaluate the children's performance but either misinterpreting performance on the dash event (i.e., mistaking longer time for better score) or making a calculation error and therefore citing the wrong child as the overall winner.

Score 2 for a response which cites an overall winner or a tie between children with an explanation that demonstrates some recognition that a quantitative means of comparison is needed to choose the winner (i.e., Joe won 2 games out of 3).

Score 1 is the student makes a selection of an overall winner with an illogical or non-quantitative comment or without providing any explanation (i.e., I think that Joe will win because he did pretty good in al the events).

Score 0 for a nonresponse or no winner given based on non-quantitative reasoning (i.e., All the children did well.).

### Source:

These rubrics are adapted from those found in the National Assessment of Educational Progress's Pilot Study of Higher-Order Thinking Skills Assessment Techniques in Science and Mathematics (Blumberg et al., 1986, pp 19-34).

## Tug of War

Score 4 for correct answer and rationale including the strength of each side as determined from rounds 1 and 2. Three examples of rationale are described below. For each G=grandmas, A=acrobats, I=Ivan, and R3=round 3.

$$\begin{aligned} 4A &= 5G \\ I &= 2G + 1A \\ I + 3G &? 3G \\ \text{R3 } 2G + 1A + 3G &? 4A \\ 5G + 1A &> 4A \end{aligned}$$

or

$$\begin{aligned} G &= .8A \\ I &= 1.6A + 1A \\ I &= 2.6A \\ \text{R3 } 2.6A + 3(.8A) &? 4A \\ 5A &> 4A \end{aligned}$$

or

$$\begin{aligned} A &= 1.25G \\ I &= 2G + 1.25G \\ I &= 3.25G \\ \text{R3 } 3.25G + 3G &? 4A \\ 6.25G &> 4A \end{aligned}$$

Score 3 for determining a winner based on the strength of each side as determined by rounds 1 and 2, but making a calculation error or including only partial information from rounds 1 and 2. For example, determining from round 2 that Evan equals 2 grandmas.

Score 2 for a response which cites an overall winner or a tie in round 3 with an explanation that demonstrates some recognition that a quantitative means of establishing a winner is needed such as Ivan equals 3 people so Ivan and the grandmas will win because 6 people are stronger than 4 people.

Score 1 if the student makes a selection of an overall winner with an irrelevant or non-quantitative comment or without providing any explanation (i.e., I think that the acrobats will win because they are stronger.).

Score 0 for nonresponse or irrelevant response such as dogs cannot play tug of war.

Note: No scoring rubrics were published by the author of the tug of war task, so the rubrics on this page were written by the researcher with collaboration from mathematics teachers and school district mathematics curriculum consultants.

VITA



Maridyth Montgomery McBee

Candidate for the Degree of

Doctor of Philosophy

Thesis: SAMPLING VARIABILITY OF SELECTED PERFORMANCE  
ASSESSMENTS MEASURING AN EIGHTH GRADE  
MATHEMATICS DOMAIN

Major Field: Applied Behavioral Studies

Biographical:

Personal Data: Born in Tulsa, Oklahoma, January 20, 1954, the daughter  
of Mr. and Mrs. Thomas Oliver Montgomery.

Education: Graduated from Memorial High School, Tulsa, Oklahoma, in May of  
1972; received Bachelor of Arts degree in Psychology from Oklahoma State  
University in 1975; received Master of Science degree in Student Personnel  
and Guidance Counseling from Oklahoma State University in 1977; entered  
doctorate program at Oklahoma State University, June, 1991;  
completed requirements for the Doctor of Philosophy degree at  
Oklahoma State University in May, 1995.

Professional Experience: Research Associate, Oklahoma City Public Schools,  
August, 1978, to February, 1984; Senior Research Associate, Oklahoma  
City Public Schools, February, 1985, to May, 1991; Research Fellow,  
Oklahoma State University, June, 1991, to June, 1994; Adjunct Faculty,  
Oklahoma City University, January, 1993, to present; Research Associate,  
Oklahoma State University, September, 1994, to present.

Professional Memberships: International Society of Educational Planning,  
American Educational Research Association, Phi Kappa Phi,  
Association for Supervision and Curriculum Development, National  
Council on Measurement.

OKLAHOMA STATE UNIVERSITY  
INSTITUTIONAL REVIEW BOARD  
HUMAN SUBJECTS REVIEW

Date: 02-15-94

IRB#: ED-94-058

Proposal Title: SAMPLING VARIABILITY OF PERFORMANCE ASSESSMENTS  
MEASURING EIGHTH GRADE MATHEMATICS PROBLEM-SOLVING

Principal Investigator(s): Laura L.B. Barnes, Maridyth M. McBee

Reviewed and Processed as: Exempt

Approval Status Recommended by Reviewer(s): Approved

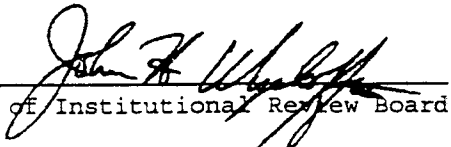
APPROVAL STATUS SUBJECT TO REVIEW BY FULL INSTITUTIONAL REVIEW BOARD AT NEXT MEETING.

APPROVAL STATUS PERIOD VALID FOR ONE CALENDAR YEAR AFTER WHICH A CONTINUATION OR RENEWAL REQUEST IS REQUIRED TO BE SUBMITTED FOR BOARD APPROVAL. ANY MODIFICATIONS TO APPROVED PROJECT MUST ALSO BE SUBMITTED FOR APPROVAL.

---

Comments, Modifications/Conditions for Approval or Reasons for Deferral or Disapproval are as follows:

Signature:

  
Chair of Institutional Review Board

Date: March 4, 1994