This is to certify that the

dissertation entitled

THE EMPIRICAL STUDY OF THE VALIDITY OF THE SECOND GRADE PRIMARY MATHEMATICS PERFORMANCE ASSESSMENT presented by

Othelia Washington Pryor

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in ~~Measurement~~ and Quantitative Methods

Major professor

Date 4/3/97

THE EMPIRICAL STUDY OF THE VALIDITY OF THE SECOND GRADE PRIMARY MATHEMATICS PERFORMANCE ASSESSMENT

By

Othelia Washington Pryor

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
of the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology and
Special Education

1997

ABSTRACT

THE EMPIRICAL STUDY OF THE VALIDITY OF THE SECOND GRADE PRIMARY MATHEMATICS PERFORMANCE ASSESSMENT

By

Othelia Washington Pryor

The increasing popularity of performance assessment instruments has not been supported with psychometrically based empirical and theoretical research pertaining to their validity. This study developed the Performance Assessment Validity Baseline Criteria procedure to examine the validity of the Second Grade Primary Mathematics Performance Assessment. The study evaluated the validity of the Second Grade Primary Mathematics Performance Assessment, the ease with which one could apply the Performance Assessment Validity Baseline Criteria, and the feasibility of using statistical measures to examine the validity of a performance assessment.

The conceptual framework incorporated some of the ideas of the educational reformers, many of whom are dissatisfied with using standardized multiple-choice tests for measuring students achievement, into a psychometric framework. The quality of the validity evidence was rated according to the purpose of the test as reported by its developers. The seven validity principle investigated were: (a) domain representation, (b) representative sampling frame, (c) directly measuring constructs, (d) cognitive complexity, (e) adequacy of scoring rubric, (f) fairness of assessment, and (g) test security. The procedure uses an expert panel to determine the quality of the validity evidence according to a three level scale, with a Level 3 being the highest rating.

The results for the validity of the Second Grade Primary Mathematics Performance Assessment were mixed. The validity evidence supported developers' claims at the Level 3 rating for domain definition, and test security; Level 2 rating for cognitive complexity, and adequacy of scoring rubric; and Level 1 rating for representative sampling frame, tasks directly measured, and fairness.

The study also investigated the possibility of using factor analysis, t-tests, and correlation analysis as traditional psychometric methods for examining the validity of the Second Grade Primary Mathematics Performance Assessment. Although it was possible to use the statistical procedures, the quality of the data did not allow for a robust interpretation of the results.

The study demonstrates that those concerned about the validity of performance assessment instruments, can benefit from a blended approach which combines the concerns of educational reformers and psychometricians. The work also illustrates how the Performance Assessment Validity Baseline Criteria can assist educators in evaluating external performance assessments, improve the quality of locally developed performance assessments, and provide a forum for discussion.

DEDICATION

This dissertation is first dedicated to God, who is the head of my life. It is with his strength and power that I have been able to stand.

I also dedicate this dissertation study to Nellie, my ancestor who was captured in West Africa and enslaved in this country, Anna McIver Hilliard, who prayed for the unborn generations of my family -- of which I am one, and the countless other forgotten members of my race, whose sacrifices made it possible for me to arrive at my goal.

ACKNOWLEDGMENTS

Each of the following individuals gave of themselves to assist me in obtaining my goal.

-- Anna Belle Smith, my grandmother, who by example gave me the courage and strength to press on.

-- Earnest and Dolores Washington, my parents, who instilled in me the ability to dream.

-- Tamara and Tiffany Pryor, my daughters, who each inspired me in their own unique ways.

-- Willa Peterson, Daisy Thomas-Quinney, Patsy McGowan, and Dedria Barker who assisted and comforted me in times of discouragement.

-- Pastor and Sister Owens, and my ITCC family who kept me true to the plan God has for my life.

-- Robert Floden, Ph.D. my chairperson, for his guidance, support, thoughtfulness, and patience.

-- William Mehrens, Ph.D. for posing issues that caused me to examine other perspectives of thought.

-- Sandi Wilcox, Ph.D. for exposing me to a different perspective of testing.

-- Ralph Putnam, Ph.D. for exposing me to the ideas of what it really means to understand mathematics.

--- My many friends and colleagues who encouraged and believed in me.

TABLE OF CONTENTS

LIST OF TABLES xi

CHAPTER 1
INTRODUCTION 1
Purpose of the Study 3
Research Questions 5
Research Question Number One 5
Research Question Number Two 6
Research Question Number Three 7
Research Question Number Four 8
Importance of Study 8
Contribution of the Research 9

CHAPTER 2
LITERATURE REVIEW 13
Standardized Tests Criticized 13
Conceptual Teaching and Learning 15
Mathematics Reform 16
The Assessment Controversy 19
Instrument or Process of Choice 19
What Determines Quality? 20
Validity and Performance Assessments 22
Conclusion 22

CHAPTER 3
HISTORY OF THE SECOND GRADE PRIMARY
MATHEMATICS PERFORMANCE ASSESSMENT 25
Conceptual Background 26
The Achievement Gap 27
Afro-American Cognitive Style 28
Testwiseness 29
Other Interventions 31
Teachers' Assurances of Academic Growth 31

History of Mathematical Reform 32
Conceptual Teaching and Learning 32
The New Mathematics Curriculum 33
Measuring Mathematical Achievement 34
A New Form of Assessment 34
Development of the Second Grade Primary Mathematics Performance Assessment 34
Conceptual Framework of the Assessment 36
Field Testing the Assessment 38
Staff Development 39
Reliability 39
Sample Selection 40
Summary 40

CHAPTER 4
THE PERFORMANCE ASSESSMENT VALIDITY BASELINE CRITERIA 42
Rationale Supporting Middle of the Road Researchers 43
Brief Presentation of Selected Articles 44
William A. Mehrens' Article 44
Eva L. Baker's Article 45
Linn, Baker, & Dunbar's Article 46
Lorrie Shepard's Article 47
Discussion of the Performance Assessment Validity Baseline Criteria 48
General Issues Concerning Criteria 50
Undefined Terms and Procedures 50
Principle 1: Domain Adequately Defined 51
Principle 2: Domain Adequately Sampled 54
Principle 3: Directness of Tasks 57
Principle 4: Cognitive Complexity 60
Principle 5: Concrete Scoring Rubric 63
Principle 6: Fairness 66
Principle 7: Test Security 69
Summary 73

CHAPTER 5
APPLICATION OF THE PERFORMANCE ASSESSMENT VALIDITY BASELINE CRITERIA 74
Validity Scoring Index Sheet 74
Composition of Expert Panel 75
Levels of Validity Evidence 77

Principle 1: Domain Adequately Defined 79
Validity Evidence for Principle 1 79
Scoring of Validity Principle 1 80
Principle 2: Domain Adequately Sampled 82
Validity Evidence for Principle 2 83
Scoring of Validity Principle 2 84
Principle 3: Directness of Tasks 87
Validity Evidence for Principle 3 88
Scoring of Validity Principle 3 89
Principle 4: Cognitive Complexity 91
Validity Evidence for Principle 4 91
Scoring of Validity Principle 4 92
Principle 5: Concrete Scoring Rubric 94
Validity Evidence for Principle 5 94
Scoring of Validity Principle 5 96
Principle 6: Fairness 98
Validity Evidence for Principle 6 98
Fairness: Test Bias 99
Fairness: Opportunity to Learn 100
Scoring of Validity Principle 6 101
Principle 7: Test Security 103
Validity Evidence for Principle 7 103
Scoring of Validity Principle 7 104
Conclusion 106

CHAPTER 6
APPLICATION OF TRADITIONAL MEASUREMENT TECHNIQUES 108
Validity Discussion 109
Content-Related Validity Evidence 110
Criterion-Related Validity 111
Construct-Related Validity 112
Importance of Purpose to Validity 112
Content Validity of the Assessment 113
Factor Analysis 115
Factor 1 118
Factor 2 118
Factor 3 119
Reliability Analysis 119
Limitations of the Factor Analysis 121
Validity Implications 121

Construct Validity of the Assessment 122
T-Tests Analysis 124
Methodology 124
Limitations of T-Test procedure 126
Discussion of Validity Evidence 127
Correlation Analysis 129
Methodology 129
Discussion of Validity Evidence 131
Limitations of the Correlation Analysis 132
Conclusion 133

CHAPTER 7
WHAT WAS LEARNED 135
Research Question Number One 137
Research Question Number Two 138
Research Question Number Three 140
Extracting Practical Applications from Theory 140
Research Question Number Four 143
Advice for Future Developers of Performance Assessment Instruments 144
Conclusion 145

APPENDIX A: Second Grade Primary Mathematic Performance Assessment . . 148
APPENDIX B: Scoring Rubric 151
APPENDIX C: Factor Analysis 153
APPENDIX D: T-tests Analysis 157

REFERENCES 160

LIST OF TABLES

Table 1-Performance Assessment Validity Baseline Criteria 49

Table 2-Principle 1: Validity Criteria Matrix 53

Table 3-Principle 2: Validity Criteria Matrix 56

Table 4-Principle 3: Validity Criteria Matrix 59

Table 5-Principle 4: Validity Criteria Matrix 62

Table 6-Principle 5: Validity Criteria Matrix 65

Table 7-Principle 6: Validity Criteria Matrix 68

Table 8-Principle 7: Validity Criteria Matrix 72

Table 9-Principle 1: Validity Scoring Matrix 81

Table 10-Performance Assessment Test Blueprint 84

Table 11-Principle 2: Validity Scoring Matrix 86

Table 12-Principle 3: Validity Scoring Matrix 90

Table 13-Principle 4: Validity Scoring Matrix 93

Table 14-Principle 5: Validity Scoring Matrix 97

Table 15-Principle 6: Validity Scoring Matrix 102

Table 16-Principle 7: Validity Scoring Matrix 105

Table 17-Performance Area Description of Task Items 114

Table 18-Second Grade Primary Mathematics Performance Factor Analysis 116

Table 19-Tasks in Each Factor 117

Table 20-Reliability of Factors 120

Table 21-T-test Performance Assessment Data 125

Table 22-T-test CAT Data 126

Table 23-Correlation of Performance Assessment to the CAT Subtests 130

CHAPTER 1

INTRODUCTION

In recent years, the public's dissatisfaction with the nation's school system has thrust the measurement community into the midst of an educational controversy. This controversy stems from a concern surrounding the education students are receiving. This discontentment has lead to at least six major cycles of educational reform in the last 50 years. These reform movements in the past, relied heavily on standardized testing as the barometer for measuring systemic change and assigning accountability (Madaus & Tan, 1993). The 1990s reformers differ dramatically from these historical educational movements. Some of these educators attribute many of the problems in education to the very assessment systems that have been traditionally used to measure achievement.

The public has been concerned with the level of achievement of American students for several decades. In 1957, the Russian Sputnik launching caused the United States to question America's technical capabilities. A *call for reform* in the 1960s emerged when performance differences between African American and Caucasian students on standardized multiple-choice tests was publicized. The 1970s reform movement surrounded the public's dissatisfaction over declining Stanford Achievement Test (SAT) scores (Madaus & Tan, 1993). In the late 1980s, John Cannell alarmed educators and the general public when he reported that *all* 50 states were testing above the national average (Cannell, 1988; Madaus & Tan 1993).

The 1990's call for educational reform partially centers around the public's concern for the future economic competitiveness of the United States (Madaus & Tan, 1993). Students performance in mathematics and science and the declining productivity of American industry have caused educators, politicians, and the general public to turn an accusing eye toward education.

This new generation of educational reformers, like those in the past, is interested in changing the curriculum and educational delivery systems. However, many of these 1990s reformers are adamant about replacing traditional multiple-choice tests with alternative forms of assessment. They believe these new assessment instruments, which reportedly differ substantially from traditional assessments, will improve student achievement (Hipps, 1993; Madaus & Tan, 1993; Moss, 1992; Wiggins, 1989). Some proponents of performance assessment instruments view them as a means to revamp the entire educational system.

Many advocates of performance assessment instruments believe adoption of these new assessment systems will help solve the Nation's educational problems. Support for this idea is increasing at an astonishing rate (Cizek, 1991; Moss, 1992). Almost half of the states are considering or planning to implement some form of performance assessment as part of their standard testing policy (Rothman, 1990). Unfortunately, this rise in popularity of performance assessments has not been supported by the technical research base necessary to assure their worthiness.

This reform movement has caused an assessment controversy which has fractured the educational community. At one end of the continuum stands traditional psychometricians, who believe that any test, regardless of format, can be evaluated with

standard measurement theory and psychometric principles. At the opposite end of the continuum are the educational reformers, who advocate for the complete dismantling of the validity concept as it now exists (Moss, 1992). Positioned somewhere between these two extremes are the *middle of the road* psychometricians, advocates for a thoughtful, albeit blended approach between the two radical positions. This empirical study is an attempt to evaluate how appropriate the middle of the road concept of validity is for determining the worthiness of performance assessment instruments. This research is also designed to determine whether the traditional psychometric techniques can evaluate the validity of a performance assessment instrument that has retained much of the standardization of a multiple-choice test instrument.

Purpose of the Study

Most of the discussions surrounding the validity of performance assessments have been based on a hypothetical instrument. In the literature, this instrument is often described as a creative, self-reflective, multi-faceted tool capable of assessing multiple skills (Moss, 1992). Proponents of performance assessments claim that traditional validity measurement theories and procedures are incapable of assessing these individualized free-form instruments. Admittedly, current psychometric techniques would have difficulty evaluating the validity of these performance assessment instruments. However, it is possible that, within the sphere of performance assessments, a more standardized instrument exists. It might be possible to evaluate the integrity of this type of performance assessment with traditional psychometric techniques. This type of performance assessment instrument might lend itself to a validity analysis which

incorporates some of the ideas of middle of the road psychometricians.

In 1990, the Oak Hill School District developed such an instrument; the Second Grade Primary Mathematics Performance Assessment. This instrument, designed to measure higher-order-thinking skills, incorporates some of the standardization of multiple-choice tests. The validity analysis of this performance assessment instrument is the focus of this empirical dissertation study.

This empirical validity study of the Second Grade Primary Mathematics Performance Assessment is an analysis of a performance assessment instrument that has a standardized administration format. The assessment was designed to measure students' knowledge of the district's second grade mathematics' curriculum. The instrument was developed by a local committee of content specialists, classroom teachers, and psychometric experts.

The Second Grade Primary Mathematics Performance Assessment represented Oak Hill's first step toward developing a district-wide performance assessment. The eleven task assessment required students to formulate a response or explanation with the aid of manipulatives they routinely worked with in mathematics classes. There was only one right answer to each task, but students had the option of presenting their response in numerous ways.

The Second Grade Primary Mathematics Performance Assessment, like many other performance assessments, grew out of educators' discontentment with the California Achievement Test (CAT) (CTB/McGraw-Hill, 1985). The developers designed this assessment to (a) assess individual second graders' mathematics capabilities or skills as defined by the Oak Hill Curriculum, (b) provide information on individual

students' mathematical abilities which were not tested on the CAT, and (c) provide district level accountability standards on the successful implementation of the district's mathematics (Shiffler, Beyer, & Sperling, 1992). The Second Grade Primary Mathematics Performance Assessment was developed as an attempt to evaluate these concerns. A series of research questions designed to evaluate the validity of the performance assessment will be the focal point of the research.

Research Questions

A major component of the concept of validity involves the degree to which the instrument serves the purpose for which it was designed (Messick, 1989). The issue of validity, which addresses the inferences and interpretations from the test scores, should be examined from both a logical and empirical perspective to determine if the assessment satisfies its stated purposes. The research questions guide the research of the Second Grade Primary Mathematics Performance Assessment.

Research Question Number One

Research Question Number One aids in understanding what the Oak Hill School District was trying to do when they developed the performance assessment instrument. This question was, "What was the purpose of the Second Grade Primary Mathematics Performance Assessment?" This question is important in examining the validity of any test instruments.

Messick (1989) believes that the purpose of a test aids in determining the appropriateness of the inferences made from the test scores. Therefore, the validity of a test must consider the purpose the developers had when the test was designed. A portion

of this study investigates the contextual backdrop which preceded the development of the Second Grade Primary Mathematics Performance Assessment.

Research Question Number Two

The second question addresses the logical framework of the Oak Hill School District in developing the performance assessment. This question asks, "What procedure or process did the Oak Hill School District use to develop and implement the Second Grade Primary Mathematics Performance Assessment?" An old adage says, "It is impossible to understand where you are going until you first understand where you have been". The historical context surrounding the development of the Oak Hill School District assisted in the validity analysis. The answer to this question was important because it formed the historical context of the test's development.

This question is also important because Oak Hill's experience probably mirrors the logical approach other districts might take in attempting to develop a performance assessment instrument. This school district had a clearly defined curriculum which was related to their classroom practices. Their development committee included local mathematics consultants, psychometricians, and classroom teachers. In addition, the district also had the support for moving in this direction. The Oak Hill School District appeared to have all of the theoretical building blocks in place to develop a valid performance assessment instrument. An insiders' look into the process will help one determine how important these building blocks were to produce a valid performance assessment instrument.

If valid inferences can be made from the Second Grade Primary Mathematics Performance Assessment, this model will aid other school districts searching for

guidelines and procedures to assist them in developing their own quality performance assessment instruments.

If the analysis reveals the Second Grade Primary Mathematics Performance Assessment is invalid for the purposes designed, that knowledge would also be of benefit to the educational community. Those results might indicate processes a school district might use to improve procedures they had designed to develop a performance assessment. The study also might give a school district a preview of the Herculean task they are considering.

Research Question Number Three

The third research question is designed to promote a common junction whereby educational researchers, regardless of their perspective, can actively contribute to the discussion of the validity of performance assessment instruments. This research question is, "Does the Second Grade Primary Mathematics Performance Assessment address some of the validity concerns posed by middle of the road psychometricians?" This analysis strives to understand how the theories of validity can be applied to an actual test instrument. The investigation of the Second Grade Primary Mathematics Performance Assessment might form the centerpiece of a debate whereby all those interested in this topic can participate.

This empirical study will analyze the position espoused by the middle of the road psychometricians who support the possible move toward performance assessment instruments. These researchers have combined their traditional psychometric training with some assessment ideals proposed by educational reformers. Their middle of the road stance attempts to merge two opposing forces, educational reformers and traditional

psychometrics, into a blended approach which gleans the best thinking from both.

Research Question Number Four

The final research question investigates the appropriateness of evaluating the validity of a performance assessment instrument with psychometric principles. This question asks, "Are traditional psychometric techniques able to yield useful information for evaluating the validity of the Second Grade Primary Mathematics Performance Assessment?" Proponents of performance assessments often claim that traditional measurement theories and psychometric techniques of validity are unable to assess the creative, multi-faceted skills required by performance assessments (Moss, 1992). However, these techniques might be valuable if applied to a more conservative performance assessment instrument. The application of these methods in evaluating the validity of the Second Grade Primary Mathematics Performance Assessment, will help determine if these techniques can offer additional support concerning their validity.

Importance of Study

This empirical analysis of the Second Grade Primary Mathematics Performance Assessment is important from both a theoretical and a practical perspective. Theoretically, the study provides a basis whereby several assessment ideas can be tested. This instrument conformed to many of the standards researchers of performance assessment advocate (Shavelson, Baxter, & Pine, 1991a). The tasks on the Second Grade Primary Mathematics Performance Assessment examined areas valued by mathematical reformers and were directly linked to the district's curriculum. The assessment required students to develop an explanation by actively responding with manipulatives instead of

selecting a response from a predetermined list. In addition, the Second Grade Primary Mathematics Performance Assessment used a standardized structured interview format for administration purposes.

From a practical perspective, the empirical research of the Second Grade Primary Mathematics Performance Assessment is important because it provides a procedural study for school districts interested in developing their own performance assessments. The study of the process Oak Hill used to convert abstract theoretical principles into an actual test instrument would be important for a school district considering a similar undertaking.

Contribution of Research

This empirical study of the validity of the Second Grade Primary Mathematics Performance Assessment will contribute to the educational community. It is one of the few empirical studies which discusses the validity of performance assessment instruments. Most of these theoretical discussions of validity are aimed at hypothetical instruments; this study is the analysis of an actual instrument.

This research will also assist in the examination of the theories that are currently being proposed by middle of the road psychometricians. The inclusion of performance assessments in the quest to reform education seems like a laudable idea; but in reality, the data and knowledge known about performance assessments are not supported by an expansive research base (Baker, 1991; Brandt, 1992; Mehrens, 1992; Rothman, 1990; Stiggins, 1991; Williams, Phillips, & Yen, 1991).

The empirical study of the validity of the Second Grade Primary Mathematics

Performance Assessment is also an important contribution to the educational community because it discusses one of the current hot topics in education in an objective manner. Cizek (1991), in utilizing Slavin's categorization of the 12 stages of the swinging education pendulum, placed performance assessment somewhere between the *gee whiz* and the *hot topic* stage.

In 1991, President Bush advocated using performance assessments as a way of determining if the Nation had arrived at *world class educational standards* (cited in Shavelson, Baxter & Pine, 1991b). Some states are currently designing performance assessment tasks as part of their high school competency exit exams (Schiller, 1992). These assessment instruments are so popular that they have even been discussed in popular journals and have been featured numerous times on national network news broadcasts (Frechtling, 1991). However, objective studies of their validity are lacking. This study will help give practitioners and researchers an actual view of the ramifications of starting down the performance assessment path as a means of documenting student achievement.

This empirical study of the Second Grade Primary Mathematics Performance Assessment is organized as follows:

1. Chapter 2 is the review of the literature which framed the theoretical framework that fueled the interest in the validity of performance assessment instruments.

2. Chapter 3 presents the environment of the Oak Hill School District and processes used to develop the Second Grade Primary Mathematics Performance Assessment.

3. Chapter 4 describes the development of the Performance Assessment Baseline Validity Criteria that will be used to analyze the validity of the Second Grade Primary Mathematics Performance Assessment.

4. Chapter 5 applies the Performance Assessment Baseline Validity Criteria to the Second Grade Primary Mathematics Performance Assessment and contains a more extensive literature review of the documents pertinent to its development.

5. Chapter 6 investigates the issue of validity from the use of psychometric techniques that rely on correlation statistics.

6. Chapter 7 discusses the lessons learned from the empirical validity study of the Second Grade Primary Mathematics Performance Assessment.

Educators must know that scarce resources and limited student and teacher classroom time are not being spent on inefficient, untried, and unproven assessment measures (Williams, Phillips, & Yen, 1991). This study is designed to contribute to the empirical research base concerning performance assessments so educators can make more informed assessment decisions.

This project is also important because it moves the assessment community more towards a collaborative research model instead of the combative stance that has surrounded this issue. The controversy over the validity of performance assessment instruments has escalated to the point where it is difficult for the two sides to communicate with each other. It is not the intention of this researcher to stand on the measurement side of this issue and shout insults at the educational reformers. This study is planned as a merger of the two ideologies.

The measurement community must realize that school districts are developing and implementing performance assessment instruments. State Boards of Education are incorporating performance assessment tasks on competency and exit exams. It seems far more reasonable for psychometricians to participate in the revolution and give the educational community the benefit of their considerable knowledge and experience concerning assessment models, then to stand on the sidelines and allow students to become a casualty in the reform war. This project is designed to approach the Second Grade Primary Mathematics Performance Assessment with an objective lens and perhaps provide a context whereby both sides can start the discussion.

CHAPTER 2

LITERATURE REVIEW

This chapter describes the progression of the theoretical discussions which preceded the current interest in performance assessment instruments. This review describes the historical groundwork from which the theoretical discussions emanated. It details some of the major research developments that helped fuel this assessment controversy.

This controversy started decades ago when some educators became disenchanted with popular standardized test instruments. Their original discontentment from the 1960s has grown into the educational revolution that is brewing today.

Since the 1960s, a growing number of educators have voiced a distrust of measuring student achievement with multiple-choice tests. These educators found it inconceivable to think that these objective instruments which predominate school districts could yield information which accurately assessed student learning.

Standardized Tests Criticized

The public and professional ire concerning the predominant use of standardized multiple-choice testing modes for accountability in the United States school systems has been brewing for decades (Madaus & Tan, 1993; Peterson & Knapp, 1993). However,

amid all of the controversy, the use of standardized multiple-choice tests in schools has been escalating geometrically (Madaus & Tan; Rudman, 1987). The testing of American youngsters prompted one research team to call these students "the most tested but least examined students in the world" (Resnick & Resnick, 1992).

Critics of multiple-choice tests have accused these tests of narrowing the curriculum, stressing routine and repetitive tasks, changing the instructional focus in classrooms, measuring only what is easily tested, and furnishing information which teachers regard as essentially useless for instructional purposes (Fairtest, 1988; LeMahieu, 1984; LeMahieu & Leinhardt, 1986; Linn, 1993; National Council for Teachers of Mathematics (NCTM), 1989; Rudman, 1987; Wolf, Bixby, & Glenn III). Psychometricians, on the other hand, have supported the use of standardized multiple-choice tests for providing valuable comparative, selection, and sorting information for individual students, schools, and school districts. Advocates of performance assessments believe an incompatibility exists between the attributes standardized multiple-choice tests are able to evaluate and the conceptual processing skills educational reformers want to measure (NCTM, 1989). This controversy appears to have the measurement community on a collision course with many individuals promoting change in education (Webb & Romberg, 1992; Williams, Phillips, & Yen, 1991).

The discontentment with standardized testing formats was further propelled by the advent of the theories of cognitive psychologists who supported the notion of conceptual teaching and learning. These cognitive models of learning represented a paradigm shift that further eroded the confidence many educators held for the traditional assessment instruments.

Conceptual Teaching and Learning

The changing emphasis in educational psychology from behaviorist to cognitive models of learning caused a paradigm shift in the way educators viewed learning, learners, and assessment (Wittrock, 1991). During the last 20 years, psychological and educational research has focused on understanding student knowledge and thought processes from a different perspective (Wittrock, 1991). Cognitive psychologists' research tried to delve into the processes of learning instead of solely the resultant behaviors.

Cognitive psychologists believe students assimilate and process new knowledge into pre-existing schemata structures. They believe these structures are shaped and influenced by prior knowledge and experiences (Wittrock, 1991).

Cognitive psychologists also espouse a unique and highly personalized view of student learning, in which each student constructs customized schemata structures which determine how new information will be interpreted, conceptualized, absorbed, and processed (Peterson & Knapp, 1993). Thus, this concept of an individually constructed learning schema caused some educators and educational measurement specialists to reconsider the appropriateness of the traditional multiple-choice instruments (Baxter, Shavelson, Goldman, & Pine, 1992; NCTM, 1989).

These learning models caused some cognitive psychologists to shun traditional assessments and work toward the development of assessment instruments they believed capable of measuring student conceptions, learning strategies, metacognitive, and affective thought processes (Wittrock, 1991). This belief caused them to support the redesign and restructuring of education assessment instruments (Baxter, Shavelson,

Goldman, & Pine, 1992; Lesh & Lamon, 1992; NCTM, 1989; Wittrock, 1991). This new emphasis on conceptual learning and teaching supplied additional ammunition to the reformers who were trying to change the landscape of school testing.

Educational reformers and cognitive psychologists started demanding a more expansive view of assessment which incorporated a broad-based view of school achievement and knowledge. Some psychometricians viewed those leading the assault against classical measurement theories and test development methodology as "radical subversives with hidden political agendas" (Williams, Phillips, & Yen, 1991). Rothman (1990) even called them technically naive researchers in search of the education holy grail. Even though these new theories were starting to divide the educational community, this impetus for change supported by educational reformers was being felt throughout the Nations' school systems.

Some professionals revised their instructional and assessment programs to reflect cognitive psychologists' theories of learning. Mathematics was one of the areas where this change was most strongly felt. For years, the public's dissatisfaction with students' knowledge in mathematics has made this area a prime target for reform. Mathematics educators embraced the ideas of cognitive psychology and combined it with their own radical thinking, which changed the way mathematics was to be taught, learned, and assessed. These changes further advanced the mathematics community toward supporting performance assessment instruments.

Mathematics Reform

The national news services had widely reported the sad state of affairs in

mathematics education in America (Dossey, 1989). The *Second International Mathematics Study* (SIMS) reported that the average Japanese student demonstrated higher achievement levels in mathematics than the top five percent of American students enrolled in college preparatory mathematics courses (McKnight et al., 1987). Experts claimed that the study of mathematics is so important that comparative studies of world economies have used mathematics achievement as an indicator of a nation's future economic potential and productivity (Robitaille et al., 1991). This pronouncement further eroded the public's confidence in the way mathematics was being taught and assessed in the United States.

The National Council for Teachers of Mathematics (NCTM) emerged as a national leader in the educational reform movement that swept the country. This organization actively supported the development of new assessment instruments that assessed those traits, skills, knowledge, and processes valued by mathematicians. Members of this organization believed students should be examined by instruments that measured traits worth learning and teaching; not those attributes that were convenient and easily measured by multiple-choice tests (Resnick, 1987). This prompted mathematics educators to start investigating the methods that might improve the teaching, learning, and assessing of their discipline.

Educators began this investigation by questioning the traditional curriculum which relies heavily on computational drills, rote memorization, and algorithmic formulas (NCTM, 1989). Mathematics professionals believed that the traditional curriculum emphasized skills that were conducive to computational proficiency but not mathematical power and problem solving (NCTM, 1989). Analysis of the current classroom conditions

in mathematics caused the mathematics community to advocate for renovations and revisions in curriculum and assessment if the United States was going to remain competitive in the world today and in the next century (NCTM, 1989; Webb & Romberg, 1992).

In 1986, the NCTM Commission on Standards for School Mathematics developed standards for curriculum, instruction, and assessment necessary for students to learn mathematics at the primary, middle, and high school levels (NCTM, 1989; Webb & Romberg, 1992). This commission determined that students must demonstrate the following five goals in order to be considered mathematically literate:

1. Learn to value mathematics,
2. become confident in one's ability,
3. become a mathematical problem solver,
4. learn to communicate mathematically, and
5. learn to reason mathematically.

These characteristics are some of the traits that NCTM would like students to exhibit (NCTM 1989; Webb & Romberg,1992).

Much of the reform movement in mathematics, revolved around the philosophy of teaching for understanding. This concept focused on students ability to understand mathematical products and processes (Peterson & Knapp, 1993; NCTM, 1989). Therefore, advocates of teaching for understanding believed the increased emphasis on cognitive processes mandated a change in assessment methods which documented student ability (Baxter, Shavelson, Herman, Brown, & Valadez, 1993).

These new assessments were to focus on the processes students used to solve

mathematical problems. The professional mathematicians also wanted the assessment instruments to allow students to communicate mathematically. These assessments were to include the components of multi-step problem solving. Needless to say, traditional multiple-choice instruments were not viewed as being able to fulfill the needs of the mathematics community. Mathematicians viewed performance assessment instruments as the measuring devices that might be able to capture the process they believed relevant in assessing students' achievement. Some researchers believed that the mathematics reform could only be sustained if standardized multiple-choice tests were replaced with performance assessments (Shavelson & Baxter, 1992). These developments in the educational community continued to drive a wedge between the educational reformers and the assessment community.

The Assessment Controversy

Instrument or Process of Choice

One of the major issues in this controversy surrounds selecting the instrument most capable of assessing mathematics achievement. Measurement theorists use the generic term *test* to define any instrument or process which measures student attributes (Crocker & Algina, 1986; Nitko, 1989). What makes a test a test, is the fact that the instrument samples behavioral traits from a universal cognitive domain. Psychometricians believe this simple but broad definition encompasses sampled behavioral traits from either a qualitative or quantitative assessment instrument. Therefore, this test definition allows measurement theorists to use traditional psychometric concepts and techniques to evaluate instruments which range from

structured standardized multiple-choice tests to the more loosely defined expanse of performance assessments. Some proponents of performance assessments strongly object to the use of psychometric principles and concepts to evaluate the validity of performance assessment instruments. They consider the standardized multiple-choice test and performance assessment instruments totally dissimilar because they evaluate different aspects of student knowledge (Baker, 1991; Cizek, 1991).

Supporters of alternative assessments have called their evaluative instruments *direct assessments* (Kirst, 1991), *authentic assessments* (Fairtest, 1988; Williams, Phillips, & Yen, 1991), *performance assessments*, and a host of other names. Portfolio assessments, written essays, performance assessments, oral discourse, opened ended items, and short answer responses are examples of the type of assessment instruments supported by mathematics reformers (Fairtest; Frechtling, 1991). Although the educational community has not yet reached consensus on the proper name for these alternative assessment forms; one thing is clear. Supporters of performance assessment abhor calling their envisioned instruments tests.

While the confusion surrounding the proper term for these assessment devices is not of major concern, it still inhibits communication between the two groups. However, the controversy concerning the appropriate methodology for determining the validity of performance assessments is of major concern.

What Determines Quality?

The issue of the validity of performance assessment instruments is a topic that has split the educational community. Some measurement theorists criticize performance assessments for lacking the technical sophistication and research methodology necessary

to support validity (Baker, 1991; Linn, Baker, & Dunbar, 1991; Shavelson, Baxter, & Pine, 1992; Williams, Phillips, & Yen, 1991). Others psychometricians suspect performance assessments of measuring nothing more than the good will and intentions of well meaning, albeit unsound, test developers (Cizek, 1991). Some members in the measurement community point out that many of the supposedly innovative ideas and theories concerning these new assessment forms have existed in educational institutions for decades (Cizek, 1991; Mehrens, 1992). Others accuse performance assessments of depending on face validity, their perceived appearance of measuring higher order thinking and reasoning processes as proof of their validity (Baker, 1991; Lane, Stone, Ankenmann, & Liu, 1992; Linn, Baker, & Dunbar, 1991; Williams, Phillips & Yen, 1991).

Measurement theory considers validity to be the most important characteristic for determining the worthiness of any test. Validity as a concept refers to the appropriateness, usefulness, and meaningfulness of the specific inferences one can make from test scores (Messick, 1989). Certifying a test valid for a particular use means that the theoretical and psychological underpinnings of the test support its intended purpose. More simply stated, a test is considered valid when the test measures what it was designed to measure (Allen & Yen 1979; American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1985; Crocker & Algina, 1986).

It seems reasonable that this minimum--although critical--element of purpose would be valued by developers of any test, regardless of its format (Williams, Phillips, & Yen, 1991). Yet, as straight-forward as this statement of validity seems, the situation becomes highly charged and controversial when the concept of traditional validity is

applied to performance assessment instruments.

Validity and Performance Assessments

On one side of this educational controversy stand radicals who favor dismantling the traditional measurement concepts of validity. These researchers criticize the current concept of validity as being incapable of validating performance assessments, which they purport measures multi-dimensional problems and learning processes (Moss, 1992; Reported in Williams, Phillips, & Yen, 1991).

On the other side of the controversy stand the supporters of traditional psychometric principles and theory. These professionals believe measurement concepts of validity are suitable for any assessment instruments. They resist dismantling any of the traditional techniques and procedures currently used to assess validity. These psychometricians believe that if performance assessment instruments cannot measure up to these standards, the integrity of the test might be compromised.

Somewhere in the middle stand psychometricians and a more moderate group of educational reformers who believe that performance assessment instruments are worth investigating. They support the concept of test validity and are willing to entertain the possibility of expanding the concept and methodologies which analyze the concept (LeMahieu, 1992; Shepard, 1992). These middle of the road researchers are interested in determining the possible validity of performance assessment instruments and the procedures that might be useful in determining it.

Conclusion

This literature review revealed a range of views on assessment models in general

and the various positions concerning the validity of performance assessment instruments. However, even though sharp disagreements exist in the field, this controversy might benefit from a more reasoned approach.

Although it might interest some scholars to consider how radical revisions of validity apply to local assessments, it seems likely that many school practitioners will be more interested in whether their test development procedures design valid instruments. A more extensive literature review which details the thinking and theories which were considered to develop a criteria for evaluating the validity of the Second Grade Primary Mathematics Performance Assessment, is presented in Chapter 4. This analysis of the validity of the Second Grade Primary Mathematics Performance Assessment uses the moderate or middle of the road approach which supports an expanded version of validity within a psychometric framework.

This study selected issues educational reformers believe might be useful in examining the validity of a standardized performance assessment instrument like the Second Grade Primary Mathematics Performance Assessment. It also combined those theories with the psychometric techniques which might be applicable to analyze the validity of a non-traditional testing instrument. This research added the thinking of psychometric moderates in an attempt to develop a validity criteria which incorporates the theories of cognitive psychology and conceptual mathematics within a more moderate psychometric framework.

Before the validity of the Second Grade Primary Mathematics Performance Assessment is examined, an investigation into the purpose and intent of the test must first be explored. This information was gathered from investigating the environment which

preceded the test's development. The developmental history of the Second Grade Primary Mathematics Performance Assessment is discussed in Chapter 3.

CHAPTER 3

HISTORY OF THE SECOND GRADE PRIMARY MATHEMATICS PERFORMANCE ASSESSMENT

Basic to validity is the principle that a valid test should coincide with the purpose for which it was designed. Therefore, researchers interested in evaluating the validity of a test instrument should consider the historical context which preceded the designing of the instrument.

The history of the development of the Second Grade Primary Mathematics Performance Assessment provided the answer to the first two research questions posed by this study. Those questions were (a) what was the purpose of the Second Grade Primary Mathematics Performance Assessment and (b) what procedure or process did the Oak Hill School District use to develop and implement the Second Grade Primary Mathematics Performance Assessment? The answers to these questions assisted in determining the purpose of the instrument. That purpose in turn, was used to frame the rationale necessary for developing the validity criteria used to analyze the Second Grade Primary Mathematics Performance Assessment. This history was culled from interviewing developers of the test, Oak Hill staff, reviewing documents, and examining the notes and impressions made by this researcher, who was employed in the Research and Development Department during a portion of the development stage of the test.

For decades, some educators expressed dissatisfaction with using standardized multiple-choice test for measuring student achievement (Madaus & Tan, 1993). In most cases, school districts attempted to placate the critics by extolling the virtues of testing in general or hiding behind state ordered testing mandates. In this case however, the Oak Hill School District made the unusual move to consider replacing the district's multiple-choice testing instrument with a locally developed performance assessment.

Conceptual Background

The Oak Hill School District is located in a city that has a major research university. The university has considerable influence in the political and educational decisions made in the city (Pryor, 1993). The relationship between the university and the school district produced many collaborative research projects over the years (Pryor, 1989*b*). It was not unusual for district personnel to take additional course work at the university. Consequently, many of the districts' teachers were implementing new educational ideas and procedures into their courses from research they were exposed to at the university (Pryor, 1994*a*). These actions often placed the Oak Hill School District on the cutting edge of educational reform and controversy.

The education provided in the Oak Hill School District was highly regarded by the local community. The district had always prided itself on the level of educational opportunities offered to students (Pryor, 1994h). However, in the mid-1980s, the school district became aware of two problems that forced them to start looking for alternative methods for educating, teaching, and testing students. These problems were (a) the discovery of an academic achievement gap between African American and Caucasian

students and (b) the stagnation of the district's scores on the CAT, the district mandated accountability test.

The Achievement Gap

In the early 1980s, Oak Hill's Research and Development Department discovered that achievement levels for African American and Caucasian students differed significantly from each other on the CAT. The mean group score for Caucasian students was in the 70-75th percentile while the corresponding score for African American students was in the 40-45th percentile (Pryor, 1994h).

The Executive Members of the Oak Hill School Board realized that African American students were failing to achieve satisfactory academic progress. A decision was made to investigate the latest educational research on achievement in general, and the achievement of African American children in particular, in order to develop methods of successfully educating this population. Reducing and ultimately closing the achievement gap between African American and Caucasian students became one of the district's primary goals (Pryor, 1989a). The Research and Development Department was charged with tracking the achievement gap and discover methods of solving the problem (Pryor, 1994a; 1994b; 1994h).

The Research and Development Department identified various district measures which might be used to measure academic accomplishment (Pryor, 1994a; 1994b; 1994h). Staff members primarily charted mean group performance differences between African American and Caucasian students on the CAT. This test was administered to all students in grades one through twelve. Special education students or students not

proficient in English were exempted from taking the exams. The Oak Hill School District charged district administrators, content specialists, and researchers with investigating the educational literature to determine if educational theory might contain solutions to their problems.

Afro-American Cognitive Style

Several educators in the district supported the theory of *Afro-American Cognitive Style*. Supporters of the theory believe that this trait was responsible for the mean group performance differences between African American and Caucasian students on the CAT (Pryor, 1989a;1994b). This theory asserted that students whose ancestors were indigenous to Africa have cognitive processes which function differently than those students who were of European ancestry (Shade, 1982).

Proponents of this theory also believe that schools in the United States, which are established to support the cognitive functioning processes of the majority culture, incorporate a cognitive process and instructional style that advantages Caucasian students and disadvantages African American students. They believe that since many African Americans live in predominately black neighborhoods and socialize in organizations that mostly serve African Americans, this trait, African American Cognitive Style, still persists in African Americans. These researchers also believe that the larger society which limits complete integration of African American's into the larger society also helps perpetuate this trait (Shade, 1982).

In an extension of this theory, supporters of Afro-American Cognitive Style also take issue with the standardized multiple-choice testing instruments routinely administered in American schools. One aspect of this theory supports the aversion of

those who possess this trait with operating successfully in environments in which information is presented or analyzed in a decontextualized manner (Shade, 1986). Therefore, the decontextualization format of standardized multiple-choice tests and classroom instruction forces African American students to employ learning strategies not in keeping with their cognitive development or cultural environment (Hilliard, 1989; Shade 1982; 1986). Supporters of this theory believe that this cognitive misalignment disadvantages African Americans as a group and consequently causes their group mean test scores to be lower than Caucasian students on multiple-choice tests (Mimms, 1988). Consequently, multiple-choice tests scores are not true academic indicators for African American students' ability and are actually an artifact of testing formats.

This research led the Oak Hill School District to educate district personnel about this phenomenon. The district invited experts in Afro-American Cognitive Style to conduct workshops and inservices on this unique learning theory (Pryor, 1989c; 1989d). The district believed these seminars would help remove the barriers which impeded learning for African American students (Pryor, 1994h).

Testwiseness

Other educators in the district believed that the achievement gap was the result of African American students being inexperienced with taking standardized multiple-choice tests. Some educational researchers believe students can be taught procedures that maximize their scores on multiple-choice tests by reducing the measurement error caused by their inexperience. These learning strategies are called *testwiseness* (McPhail, 1978; 1981).

Testwiseness is defined as the ability to increase one's score by utilizing the characteristics and formats of the test, independent of the criterion being measured (Millman, Bishop, & Ebel, 1965). There have been several empirical studies in which students have demonstrated that knowledge of testwiseness has increased performance differences (Cole, 1987; McPhail, 1981; Millman, Bishop, & Ebel; Slack & Porter, 1980).

Messick (1980) found that the scores of African American students in two coaching centers improved more substantially than Caucasian students who had been given the same instruction in the same schools. He attributed this result to the fact that African American students benefited more from the instruction because they had fewer test taking skills initially. Therefore, if African American students were given testwiseness training, their scores on multiple-choice tests would improve (Mimms, 1988). Millman, Bishop, and Ebel advocated training all students in the taking of standardized multiple-choice tests to reduce systematic measurement error.

In pursuit of eliminating the bias caused by the varying degrees of experience with standardized multiple-choice testing formats, the Research and Development Department, in conjunction with curriculum experts, generated a series of inservices and curriculum materials which instructed teachers in how to better teach students to take standardized multiple-choice tests (Pryor, 1989b). The district also arranged to have experts in the field of testwiseness develop district inservices to alert teachers to this problem (Mimms, 1988; Pryor, 1989b).

Other interventions

Educators in the district also researched academic journals and investigated methods that might lead to increased achievement for the district's students in general. Under this initiative, the district revised curricula, brought new textbooks in some cases, eliminated textbooks in others, bought new teaching materials, developed new instructional techniques, and invested in computer laboratories (Pryor, 1994h). However, after seven years and a several million dollar capital investment into numerous educational programs, educational interventions, teacher inservices, computer labs, and countless other programs; the mean group performance difference between African American and Caucasian students remained virtually unchanged (Pryor, 1994a; 1994h). The resilience of the achievement gap to the district's efforts, led the Director of the Research and Development Department to calculate how long it would take the achievement gap to disappear given the district's current success rate. The answer was an astounding *200 years* (Pryor, 1989a).

In the midst of these reported mean group differences on standardized tests, teachers were reporting that achievement levels for all students in general and African American students in particular were improving. However, students' scores on the CAT did not indicate growth in achievement for either Caucasian or African American students (Pryor, 1994a; 1994b; 1994h).

Teachers' Assurances of Academic Growth

Teachers had been complaining for sometime that the CAT was misaligned to the district's curriculum and therefore failed to detect academic gains. With administrators

increasing the emphasis being placed on the achievement gap, teachers complaints began to intensify (Pryor, 1989a).

Many district teachers had become supporters of the conceptual teaching and learning movement that was coming into vogue in educational circles in the 1980s (Pryor, 1994c). These theories were having a profound effect upon the district's curriculum, teaching, and instruction. Teachers continued to complain that multiple-choice tests were incapable of capturing the conceptual learning models being taught in their classrooms (Pryor). The discrepancy between teachers' anecdotal stories of success and the resistant achievement gap on the CAT, caused administrators to begin investigating the possible incompatibility of conceptual teaching and learning models and the ability of the CAT to assess students' knowledge.

History of Mathematical Reform

Conceptual Teaching and Learning

A growing number of educators in the Oak Hill School District began to embrace the paradigm shift from behavioral models of learning to cognitive psychological models. These theories espoused a unique and highly personalized view of student learning, in which each student constructed customized schemata structures which determined how new information would be interpreted, conceptualized, absorbed, and processed (Peterson & Knapp, 1993; Resnick & Klopfer, 1989). It is this concept of an individually constructed learning schema that caused some educators of Oak Hill's staff to re-evaluate the appropriateness of the CAT. As the number of district educators who supported the paradigm shift toward conceptual teaching and learning grew, this momentum provided

the impetus for the redesign of the Oak Hill Mathematics Curriculum.

The New Mathematics Curriculum

The NCTM and the influence of cognitive psychologists became important as the district revised its mathematics curriculum. Many of the district's mathematics educators believed the tool by which achievement was measured had to change as the district moved from reliance on traditional rote instructional procedures towards a more conceptualization vision of mathematical ideas.

In the late 1980s, the district convened a group of teachers, content specialists, and other interested members of the educational community to develop a mathematics curriculum that included the most current research in the discipline. They developed a working document that outlined the objectives of the mathematics curriculum from kindergarten to twelfth grade. The district believed this new mathematics curriculum incorporated many of the ideals of the NCTM and the work of cognitive psychologists. As the committee reviewed their revised curriculum, the old question of the appropriate measuring instrument resurfaced.

Many mathematical reformers advocated replacing traditional standardized multiple-choice tests with performance assessments capable of evaluating student processes as being paramount for reshaping the mathematics curriculum. Some even argued that unless performance assessments replaced the multiple-choice assessment yardstick currently used, this new wave of mathematics reform could not be sustained (NCTM, 1989). These concerns caused the mathematics committee to consider the appropriateness of standardized multiple-choice instruments for evaluating student learning.

Measuring Mathematical Achievement

One of the issues in this reform movement involved the selection of a measurement instrument capable of assessing conceptual mathematics achievement. Supporters of mathematics reform proposed an alternative form of assessment they believed would be capable of capturing student thinking and problem solving skills (Pryor, 1994a; 1994b; 1994h).

A New Form of Assessment

Development of the Second Grade Primary Mathematics Performance Assessment

In the midst of these changing educational ideologies and philosophies in the district, the Oak Hill mathematics coordinator invited first grade teachers to a voluntary working dinner to discuss the district's new mathematics curriculum (Pryor, 1994d). During this meeting, several of the teachers expressed the belief that the CAT was developmentally inappropriate for young children. The first grade teachers were especially concerned that the scores on the CAT were invalid given their belief that the test was not aligned to the district's new mathematics curriculum.

The mathematics coordinator advised teachers to express their concerns to the school board and the Research and Development Department (Pryor, 1994d). The teachers drafted a letter which called for an elimination of the CAT at the first grade level because of its inability to measure learning (Pryor, 1994a; 1994c).

District level administrators, who had been discussing and researching the achievement gap for the last several years, thought that the first grade teachers might have happened upon an explanation as to why students were failing to exhibit academic gains

on the CAT. Administrators reasoned that if the CAT was incapable of measuring students' conceptual learning, this misalignment between assessment and curriculum might account for the stagnant test scores. Educators reasoned that if the theories of Testwiseness and Afro-American Cognitive Style were correct, then the problem was compounded for African American students. These developments laid the groundwork for the emergence of a new assessment instrument in mathematics in the Oak Hill School District. The superintendent agreed to eliminate the first grade CAT test with the stipulation the teachers, in conjunction with the Research and Development Department and content specialist, design a new assessment which would match the district's curriculum.

The Primary Assessment Committee was formed to develop a mathematics assessment that reflected the district's goals. The new assessment was to measure students' skills in mathematical reasoning, for problem solving and students' conceptual understanding of mathematics. The purpose of the mathematics assessment was to (a) assess individual second graders' mathematics capabilities or skills as defined by the Oak Hill Curriculum, (b) provide district level accountability standards on the successful implementation of the district's mathematics curriculum, and (c) assess skills not currently being assessed by the CAT (Shiffler, Beyer, & Sperling, 1992). In addition to these goals, some of the district's educators believed that the achievement gap between African American and Caucasian students would be eliminated if the district-wide accountability test was changed to a performance assessment instrument (Pryor, 1989a; 1994h).

As the committee of first grade teachers began to study the issues of alternative assessment in mathematics, it soon became clear that these issues spanned kindergarten and second grade. Therefore the committee was expanded to include teachers from these grades. Actions taken by the Primary Assessment Committee ultimately lead to the development of the Second Grade Primary Mathematics Performance Assessment.

Conceptual Framework of the Assessment

The Director of the Research and Development Department was convinced that there were a few broad concepts students needed to know in order to be conceptually prepared to meet the challenges of the next grade (Pryor, 1994h). He explained his position in the following way:

> ... there are only ten commandments, and not 150 of them, and they're pretty powerful. So I thought, probably there are only a few things kids need to know to be successful, not a whole bunch. Now, a whole bunch is what we teach, but out of that whole bunch you can distill it down to a handful that if kids knew these five things in mathematics eventually they would be adequate mathematicians. And those kind of five things are at each grade level...So I began to look at those things that were few enough and meaningful enough that teachers could measure and we could collect data on and catalogue that would in the long run predict how well the student's going to do.

The Primary Assessment Developmental Committee analyzed the revised second grade curriculum to identify the traits, skills, and abilities they believed critical for mathematical success. This curriculum emphasized more of the conceptual approach to mathematics instead of the rote procedures emphasized in the past. (A detailed version of the curriculum used by the Oak Hill School District during this developmental stage was no longer available when this research study was written. The curriculum documents

were destroyed when the school district administrative offices moved.) Based on the tasks generated for the Second Grade Primary Mathematics Performance Assessment it appeared that the developers thought the *handful of topics* emphasized in the curriculum were:

1. Modeling and explaining the place value of each digit in a two digit number.
2. Arriving at the correct response to a simple two digit addition problem that includes regrouping and modeling and giving an explanation of the answer.
3. Determining the correct response to a simple two digit subtraction problem that includes regrouping and modeling and explaining the answer.
4. Arriving at the correct response to solve a simple two digit subtraction problem that includes regrouping, modeling an answer, and giving an explanation of the answer.
5. Determining the correct response to solve a word problem and explaining the strategy used to arrive at that solution.
6. Sorting items according to attributes and providing a rational explanation.

Many district educators believed that a hands-on performance assessment was the appropriate method to measure students' skills and capabilities. Several teachers had used *Mathematics Their Way*, a manipulatives-based instruction program in their classes. They believed that these curriculum formats and the assessments included with these instructional materials could serve as a model for the performance assessment instrument being developed.

The Primary Assessment Developmental Committee devised a standardized interview script similar to the *Mathematics Their Way* materials. The committee also developed probes to be used when the raters needed clarification of the explanations and

modeling of the tasks they wanted to measure. They also developed a scoring rubric thought to encompass the majority of student responses.

The Primary Assessment Developmental Committee included staff members from the Research and Development Department, content specialists, principals, and grade level teachers to aid the developmental process. The research staff provided psychometric and statistical support for the project. The content specialists had in-depth knowledge of the mathematics curriculum from kindergarten to grade twelve, and helped guide the committee in the selection of the skills that were critical in the learning process. The teachers were familiar with the curriculum and had experience as to what one might expect from student responses. The broad expertise of the committee helped to develop the exercises, tasks, and analysis of the criteria that were important for the assessment (Pryor, 1994a; 1994b; 1994c; 1994d).

Field Testing the Assessment

The assistant coordinator of the Mathematics Department was charged with field testing the Second Grade Primary Mathematics Performance Assessment. The committee selected a subcommittee responsible for going to the schools and field testing the tasks (Pryor, 1994a; 1994b). The student interviews were videotaped so they could be used as training materials.

The procedure required two members of the team to interview each student selected for the field test. One interviewer administered the assessment while the other took extensive field notes. Both raters assigned a score to the student's performance on the tasks. When the test was over, the two assessors conferred as to the appropriate score to be assigned to the student's task. The range of scores started at *0*, (or no response) to

5 (for a completely correct response). The score received was dependent upon the completeness of the task and the child's explanation.

Staff Development

Once the field tests were completed, the Research and Development Department and the mathematics coordinators reviewed and revised the assessment and scoring rubric. The sub-committee developed staff development materials from the videos and then adjusted the written assessment interview format. This material was used to train the interviewers as to the appropriate procedures. The assessors practiced scoring the assessment tasks so that all of the tasks would be graded with a similar standard. A final set of tapes was prepared for checking the reliability of each task.

Reliability

Although this research study focuses on validity, it was still necessary to assure that the Second Grade Primary Mathematics Performance Assessment was a reliable test instrument. The scoring reliability process was based on the simple percent of agreement among the trained interviewers (Shiffler, Beyer, & Sperling, 1992).

The trained interviewers watched videotapes of students performing particular tasks. When the tape was finished, the interviewers were asked to score the performance using the scoring rubric developed for that task. The trainers were shown two different tapes of each task. The average score for the two trials was defined as the reliability for the task. For the Second Grade Primary Mathematics Performance Assessment, the reliability score ranged from a low of 61% for the explanation of the regrouping with addition task to 100% reliability on conservation of a number, the regrouping with addition solution, and the solution to problem solving.

The modeling tasks were the most complex for the interviewers to score. These tasks were complicated because students could choose from several types of materials to build the mathematical models. In addition, interviewers had to make judgments about the appropriateness of the responses under the various modeling conditions (Pryor, 1994; 1994b).

The interviewers discussed the scoring procedure after each reliability task was completed. These discussions after the scoring of the video tape, helped interviewers gain a better understanding of the scoring criteria (Pryor, 1994a; 1994b; 1994h).

Sample Selection

The 1990-91 Second Grade Primary Mathematics Performance Assessment was designed to be a tryout of how well the performance assessment process might work in an educational setting. Thus, only a small sample of second graders were selected to participate in this phase of the assessment process. The samples were drawn up using a random sample that included an over-sampling of African American students. Only students who qualified to take the district-wide administered multiple choice tests were eligible to participate in the Second Grade Primary Mathematics Performance Assessment pilot. The assessment was administered in the 1990-91 academic year (Shiffler, Beyer, & Sperling, 1992). The Second Grade Primary Mathematics Performance Assessment is in Appendix A. The scoring rubric is in Appendix B.

Summary

This analysis examined the various components of validity evidence of the 1990-91 Second Grade Primary Mathematics Performance Assessment. This analysis

however, was grounded in historical context of the knowledge surrounding the development of the test.

The development of the Second Grade Primary Mathematics Performance Assessment put the Oak Hill School District in the middle of an educational testing controversy. They attempted to do what few other districts had attempted at that time: develop a performance assessment instrument from the hypothetical description contained in the literature.

Chapter 4 describes the development of the Performance Assessment Validity Baseline Criteria which was used to analyze the validity of the Second Grade Primary Mathematics Performance Assessment. This chapter also discusses the research that contributed to its theoretical basis.

CHAPTER 4

THE PERFORMANCE ASSESSMENT VALIDITY BASELINE CRITERIA

Educators across the nation, like those in the Oak Hill School District, were moving toward replacing some of the grade level standardized multiple-choice tests with performance assessment instruments. However, as school districts were making this paradigm shift, many in the psychometric community were not yet sure of the criteria one might use to validate performance assessment instruments. This situation becomes serious as school districts consider using performance assessment instruments as accountability tests. The use of performance instruments in this setting necessitates that the quality and integrity of these assessments be determined before they are adopted as measurements of academic achievement.

This research study was designed to evaluate the various validity aspects of the Second Grade Mathematics Performance Assessment. However, defining and assessing the validity of performance assessment instruments is an area where researchers and educators have not yet reached agreement. Therefore, in order to achieve this goal, a procedure for evaluating the validity of performance assessment instruments had to be developed. The Performance Assessment Validity Baseline Criteria, was designed for that purpose.

The first part of this chapter, discusses several research articles written by

psychometricians who have considered issues concerning the validity of performance assessment instruments in large scale testing situations. Some of these research articles advocated merging select psychometric concerns about performance assessment instruments with those of educational reformers. This union generated a middle of the road psychometric approach which formed the conceptual structure of the Performance Assessment Validity Baseline Criteria. The final section of this chapter presents the Performance Assessment Validity Baseline Criteria and discusses the issues imbedded in the conceptual framework.

Rationale Supporting Middle of the Road Researchers

The theoretical underpinnings of the Performance Assessment Validity Baseline Criteria stem from reviewing and studying the works of the middle of the road psychometricians. These educators either wrote about (a) the psychometric benefits of performance assessments, (b) the validity concerns involving performance assessments, or (c) the general validity of test instruments.

Most of the consternation by the psychometric community comes from the unscientific rhetoric and claims made by some of the proponents of performance assessment instruments. Middle of the road psychometricians are concerned with developing a procedure whereby the validity of performance assessment instruments can be examined. The research articles used to develop the criteria for the Performance Assessment Validity Baseline Criteria were:

1. *Using Performance Assessment for Accountability Purposes* (Mehrens, 1992),
2. *Expectations and Evidence for Alternative Assessment* (Baker 1991),

3. *Complex, Performance Based Assessment: Expectations and Validity Criteria* (Linn, Baker & Dunbar, 1991), and

4. *Evaluating Test Validity* (Shepard, 1992).

These research articles provided original thinking concerning issues relating to the validity of the performance assessment instruments. Most of these authors approached the issue of the validity of performance assessment instruments as a type of specialized assessment. From their perspective, changing the format of the test did not liberate the developers from addressing matters related to validity. Some of these psychometricians also considered some of the assessment concerns voiced by educational reformers. This blend of conceptual ideas allowed for a fusion of theories between the two groups.

Brief Presentation of Selected Articles

William A. Mehrens' Article

Mehrens' (1992) article, *Using Performance Assessment for Accountability Purposes* discussed his beliefs and concerns surrounding the proposed uses of performance assessment instruments. He credits cognitive psychologists research on procedural knowledge with the increased interest into performance assessment instruments. Mehrens believes that cognitive psychologists have much to contribute toward the improvement and increased understanding of measurement theories. However, he believes the adoption of performance assessments for accountability purposes should be delayed until educators have a better understanding of these newly improved measurement theories.

The issues Mehrens raised that were pertinent to the development of the

Performance Assessment Validity Baseline Criteria were (a) an adequately defined domain, (b) a well sampled domain, and (c) the security of performance assessment instruments. In psychometric circles, these issues he raised are ones that test developers routinely consider. However, in the realm of performance assessment instruments, these psychometric principles are often not considered as necessary elements of a good, much less valid, performance assessment instrument by educational reformers (Moss, 1992; Wiggins, 1993).

Mehrens (1992) was also concerned that performance assessment instruments were being implemented without the technical research that normally precedes the administration of any test for accountability. He admitted that the public's rhetoric surrounding performance assessments appeared compelling, but the lack of a significant research base gave him reasons to advocate for caution.

Eva L. Baker's Article

Eva L. Baker, a Co-director at the UCLA Center for Research on Evaluation, Standards, and Student Testing (CRESST), presented a paper at the 1991 American Educational Research Association (AERA) convention entitled, *Expectations and Evidence for Alternative Assessment.* This presentation discussed her concerns over the wide-scale adoption of performance assessment instruments. Although Baker is an avid supporter of using performance assessments in educational settings, she is also concerned about their fragile research base.

Baker's (1991) concerns stem from the unsubstantiated claims often made by educational reformers. Many believed that merely replacing standardized multiple-choice tests with performance assessment instruments will change curricula, promote integrative

tasks, and make learning more complex and meaningful. The issues Baker discussed that are pertinent to the development of the Performance Assessment Validity Baseline Criteria follow: (a) Task specification, (b) task sampling, and the (c) 'til later scoring criteria. Baker also has expressed the need to develop high quality performance assessment which are supported by credible theoretical and empirical research.

Linn, Baker, and Dunbar's Article

Robert Linn, Eva L. Baker and Stephen B. Dunbar (1991) wrote an article entitled, *Complex, Performance Based Assessment: Expectations and Validity Criteria.* Although these researchers strongly support the idea of using performance assessment instruments in educational settings, they do not support the lackadaisical manner with which some proponents of performance assessments prove their superiority over standardized test instruments. Linn, Baker, and Dunbar stressed the need for a rethinking of the traits by which the validity of educational assessments are judged. Their major contribution in this article is the basis of a theoretical standard by which performance assessments might be evaluated.

Linn, Baker, and Dunbar (1991) advise fellow practitioners to exercise caution as they rush to supplant multiple-choice tests with portfolios, essays, hands-on-science tasks, and computer simulations. The issues they raised that were considered in the development of the Performance Assessment Validity Baseline Criteria were the (a) directness of the task being measured, (b) fairness of the assessment, (c) and cognitive complexity of the assessment tasks. Although Linn, Baker, and Dunbar support the move towards performance based assessment instruments, they advocate supplying the journey with credible research.

Lorrie A. Shepard's Chapter

Shepard's (1992) chapter, *Evaluating Test Validity,* discussed many aspects of empirical research relating to test validity. This chapter was intended to add to the general discussion of test validity. Shepard's expanded version of validity included both test score meaning and testing administration effects. Her general discussion of test validity procedures and the use of probing questions in an empirical case study method were the major contributions made toward the development of the Performance Assessment Validity Baseline Criteria.

Shepard's (1992) investigation of validity incorporates Messick's (1989) discussion of an expanded concept of validity, Kane's (1992) extension of practical reasoning and evaluation procedures, and Cronbach's (1989) evaluation arguments, which prioritized the inquiry method. Shepard extended the issues raised by Cronbach and Kane as the basis for specific empirical evidence concerning validity.

Shepard started her list of validity questions with, "What does the testing procedure claim to assess?" She believed this question was critical in the determination of test validity. This first question led her to consider the next logical question, "What is the purpose of the test?" It is from these primary questions that additional issues of test validity evolved.

Shepard's (1992) research generated the questioning framework of the investigatory method for the Performance Assessment Validity Baseline Criteria. Her emphasis on the following issues, although not used to generate specific concerns, did contribute to the underlying principles that were conceptually important. These issues were (a) what is the intended purpose of the test, (b) what is the adequacy of the

psychometric quality of the test, and (c) continually assessing the validity of the instrument.

Discussion of Performance Assessment Validity Baseline Criteria

The Performance Assessment Validity Baseline Criteria is comprised of seven principles designed to evaluate the quality of performance assessment instruments. The underpinnings of the Performance Assessment Validity Baseline Criteria are framed by the research concerns expressed by the middle of the road psychometricians. Table 1 lists the seven principles of the criteria.

TABLE 1 PERFORMANCE ASSESSMENT BASELINE CRITERIA

Principle 1	Is the performance domain for the test adequately defined?
Principle 2	Does the sampling frame for the performance assessment instrument adequately represent the domain of interest?
Principle 3	Are the tasks on the performance assessment measured as directly as the construct allows?
Principle 4	Do the performance assessment tasks effectively represent cognitive complexity?
Principle 5	Does the scoring rubric maximize the likelihood of student responses being accurately and consistently judged?
Principle 6	Did the developers of the performance assessment consider fairness in the procedures they used to design and implement the test instrument?
Principle 7	Is the performance assessment instrument secure?

General Issues Concerning Criteria

Some psychometricians and educational reformers have written theoretically about the validity of performance assessment instruments. Few have tried to apply these abstract ideas to a process one can actually use to evaluate a performance assessment instrument. The Performance Assessment Validity Baseline Criteria attempted to form these theoretical ideas into a procedure that practitioners can use when investigating these assessment devices. As with most first attempts in developing a process, this exercise resulted in the procedure having some missteps and a few unresolved issues. The concerns associated with this procedure are discussed in the next section.

Undefined Terms and Procedures

One of the first problems encountered in the development of the Performance Assessment Validity Baseline Criteria was that many of the researchers used concepts that do not have universal definitions in the educational community. Linn, Baker, and Dunbar (1991) wrote of the need for performance assessment tasks to measure the construct being tested as directly as possible. They also said that the tasks should be *cognitively complex* and appropriately and accurately measure the traits or skills being assessed. Although most test developers would agree that these concepts should be qualities of any assessment used to measure students' capabilities, the actual interpretation and how one might operationalize these ideas for performance assessment instruments is not agreed upon in the educational community.

The Development of the Performance Assessment Validity Baseline Criteria also created another problem. Researchers were in agreement as to a theoretical concept but

did not have a method for putting this theory into practice. For example, researchers point out the importance of having the domain of a test adequately defined (Crocker & Algina, 1986). The method that one might use however, is not well specified for using in a practical application.

In order to proceed with the development of the Performance Assessment Validity Baseline Criteria, these issues had to be solved. It was decided that the best approach would be to use the research of the middle of the road psychometricians as the framework for the practical application of the construct.

Principle 1: Domain Adequately Defined

The development of any assessment device should first start with a clear understanding of the purpose of the test (Shepard, 1992). In other words, it should first be determined what the test is attempting to measure. Therefore, the first principle of the criteria evaluates the preciseness with which the domain of the instrument has been defined. Principle 1 asks, "Is the performance domain for the test adequately defined?" It is crucial that a test have an agreed upon underlying construct before one can determine the appropriateness with which the instrument measures it (Mehrens, 1992).

A test is designed to sample the construct for which it is written (Crocker & Algina, 1986). Therefore the assessment instrument is designed to give one an indication of students' capabilities if they were hypothetically tested on the entire domain. The assessment instrument is therefore designed to represent a sample of students' knowledge in the specified content area. Students' scores on the test should represent their proficiency on the specified domain. Psychometricians consider test scores indications of

students' capabilities in that area. If the domain for which the test was written has not been adequately identified, users of the test could not use the test scores to make inference to this amorphous domain.

Psychometricians generally agree that the theoretical identification of domain identification is important to the development of any assessment instrument. However, although domain identification apparently is important, psychometricians seem to somewhat gloss over exactly what domain identification means in a practical application.

Oftentimes when psychometricians write about domain identification, the discussion quickly progresses to the observable student performances one would expect of students who were proficient in that domain. Baker (1991) writes about the importance of domain specification and task identification. Nitko (1995) when writing about test development, comments about the importance of having a clear understanding of what traits one would expect of students who were proficient in the specified performance domain. Crocker and Algina (1986) state that inferences from the assessment to the domain are only meaningful if the scores on the test are an indication of students' desired performance behaviors. They believe this can only occur when the content domain of the assessment has been adequately defined. The apparent intertwining of domain specification and identification of student performance behaviors are joined in the validity matrix of Principle 1. The validity criteria is in Table 2.

TABLE 2 PRINCIPLE 1: VALIDITY CRITERIA MATRIX

DOMAIN ADEQUATELY DEFINED

Is the performance domain adequately defined?

Validity Criteria	Domain Definition	Skill Identification
Level 1	Test developers do not define the domain of the performance assessment test.	Test developers did not identify the skills required in the domain of interest on the performance assessment.
Level 2	The district assembles an expert panel to investigate the claims of the test developers concerning identification of the domain. A majority of the experts agree with the clarity with which the developers have identified the domain.	The district assembled an expert panel to investigate the claims of the test developers concerning identification of the domain. A majority of the experts agree that tasks identify student performances that are a part of the domain of interest.
Level 3	Level 2 requirement satisfied. The district also examines evidence from field-tests or student products to substantiate developers' claims concerning domain identification.	Level 2 requirement satisfied. The district also examines evidence from field-test or student products to substantiate developers' claims concerning task identification.

Principle 2: Domain Adequately Sampled

This principle involves the degree of domain coverage a performance assessment instrument encompasses. Psychometricians consider this issue, content underrepresentation, a major threat to test validity (Cook and Campbell, 1979; Linn, Baker & Dunbar, 1991; Lane, Stone, Ankenmann, & Liu, 1992; Mehrens, 1992). The items or tasks on an assessment should adequately assess all aspects of the domain that are pertinent to identification of the traits considered important (Crocker & Algina, 1986). Unless the major aspects of the domain are tested, the inferences made concerning students' capabilities are suspect. All tests, regardless of format, must assess the areas of the domain that will evaluate all of the behaviors pertinent to that construct (Crocker & Algina, 1986; Mehrens, 1992; Messick, 1989). The adequacy of the sampling frame is critical to the appropriateness of the inferences one wishes to make.

The number of items on a performance assessment are limited because examinees are usually required to answer the question and provide an in-depth explanation of the response. Proponents of performance assessment cite these in-depth responses as being particularly beneficial because they provide insight into students' thinking and understanding (Moss, 1992).

Performance assessment instruments might possibly provide in-depth information, but they also take considerable time to administer. Therefore, the number of items that can be included on a performance assessment is severely curtailed when compared to the number of items on multiple-choice tests. One measure of the validity of a test depends primarily on the adequacy with which the sample of items represents the content, construct or domain of interest (Nunnally, 1978). Psychometricians are concerned with

the coverage of performance assessment instruments because the small number of tasks on the instruments result in a limited sampling frame (Lane, Stone, Ankenmann, & Liu, 1992; Mehrens, 1992). Mehrens (1992) and Baker (1991) cite underrepresentation as one of the major concerns confronting the validity of performance assessments.

It is the representation of the appropriateness of the content sampling procedure that allows one to make inferences from the assessment's domain of behaviors to the larger construct (Messick, 1989). However, if the sampling procedures used in performance assessment instruments underrepresent coverage of the content domain, this invalidates the inferences (Messick).

Principle 2 asks, "Do the performance assessment tasks adequately sample from the content domain?" This issue addresses the delimited sampling frame of performance assessment instruments. The validity criteria matrix for Principle 2 is in Table 3.

TABLE 3 PRINCIPLE 2: VALIDITY CRITERIA MATRIX

SAMPLING FRAME

Does the sampling frame for the performance assessment instrument adequately represent the domain of interest?

Validity Criteria	Representative Sampling Frame
Level 1	Developers do not use a procedure to determine if the tasks on the assessment provide adequate coverage of the intended domain.
Level 2	The district assembles an expert panel to determine if the tasks on the performance assessment sample from the pertinent areas of the domain. The panel develops a conceptual map, test blueprint, or other procedure to determine the breadth of the domain reflected in the tasks. A majority of the panel agree that the tasks adequately represent the domain.
Level 3	The district satisfies all of the Level 2 requirement. In addition, the experts examine student products to determine if the tasks actually produce samples of performance traits they believe should be observed when one has proficiency in this domain. If the pertinent areas of the domain are not covered, this limitation must be acknowledged.

This issue is in the forefront of the concern psychometricians have regarding performance assessment instruments (Baker, 1991; Mehrens, 1992). This validity issue is one that must be addressed by developers of performance assessment instruments if they are going to be used in large scale testing situations.

Principle 3: Directness of Tasks

This validity principle is concerned with the directness with which the skills are measured on the performance assessment instrument. Many advocates of performance assessment believe that the construction of these tasks around *real world issues,* not multiple-choice foils automatically means that the skills are direct measures of student achievement. They often cite this perception of directness as a benefit of performance assessment tasks. Many educational reformers even call performance assessment tasks authentic assessments because they believe the tasks involve performance skills that are valued in their own right and measure students' capabilities directly (Wiggins, 1989).

This directness of measure principle, contrary to common belief, is also advocated by psychometricians. They also believe that multiple-choice items should measure students' capabilities as directly as possible to reduce measurement error. Linn, Baker, and Dunbar (1991) point out that although many proponents of performance assessments cite directly measuring attributes as innovative, measurement specialists have long supported this concept. Linquist (1951) stated that

> ...it should always be the fundamental goal of the achievement test constructor to make the elements of his test series as nearly equivalent to, or as much like, the elements of the criterion series as consequences of efficiency, comparability, economy, and expediency will permit.

Although both psychometricians and educational reformers support measuring tasks as directly as possible, there is little agreement as to exactly what that expression means when one speaks of performance assessment instruments.

Principle 3 asks the question, "Are the tasks on the performance assessment measured as directly as the construct allows?" This principle is concerned with the level of interference that might exist because of the construction or assessment method with which the performance tasks are written and administered.

The validity criteria for Principle 3 is in Table 4.

TABLE 4 PRINCIPLE 3: VALIDITY CRITERIA MATRIX

CONSTRUCT MEASURED

Are the tasks on the performance assessment measured as directly as the construct allows?

Validity Criteria	Skill Interference	Assessment Procedure
Level 1	The tasks are not examined to determine if the wording of the items interfere with the knowledge area being assessed on the test.	The tasks are not examined to determine if the assessment's administration method interferes with the traits being assessed.
Level 2	The district assembles an expert panel to determine if the wording of the items interferes with the knowledge area being assessed on the test. If a majority of the panel judges a task defective, the task is revised or discarded before test administration.	The district assembles an expert panel to determine if the administration procedure interferes with the traits being tested. If a majority of the panel believes the procedure confounds testing content, the procedure is revised.
Level 3	The district satisfies all of the Level 2 requirement. The performance assessment instrument is field-tested. If a majority of the panel after reviewing the field-test data determines the context interferes with the skill being assessed, the task is discarded or revised before test administration.	Level 2 requirement satisfied. The performance assessment instrument is field-tested. An expert collects evidence to determine if the administration procedure might interfere with students exhibiting their knowledge and capabilities. If a majority of the panel believes the procedure is flawed, the procedure is replaced before test administration.

Psychometricians and educational reformers both believe that assessment tasks should not interfere with students' exhibiting mastery of a skill. Applying the procedure defined for principle 3 of the Performance Assessment Validity Baseline Criteria will assist in determining how well the developers adhered to directly measuring the construct. Elimination of skill interference and attention to the methodology used to measure student achievement should result in reduced measurement error. This reduction of measurement error would mean that skills are being assessed in a more direct manner. This increase in the directness of measuring students' capabilities would satisfy both middle of the road psychometricians and educational reformers. The Performance Assessment Validity Baseline Criteria combines these issues into a procedure for achieving that goal.

Principle 4: Cognitive Complexity

Principle 4 addresses the issue of cognitive complexity by investigating the processes and products of student work on the performance assessment instrument. This principle asks, "Do the performance assessment tasks effectively represent cognitive complexity?" The issue of cognitive complexity is especially important for developers of performance assessment instruments. This is one of the major concerns that has fueled much of the controversy surrounding educational reformers support of the move away from multiple-choice tests items as an appropriate format to assess student achievement.

Many performance assessment advocates believe that constructed mathematics assessment tasks will automatically provide activities for students that require cognitively-complex, higher-order thinking skills (Moss, 1992). These researchers, in

holding to the notion that teachers teach toward what is tested, also believe that replacement of multiple-choice items will result in more complex classroom instruction.

Psychometricians caution researchers to be cautious of the enthusiasm surrounding tasks which appear to allow for multiple representation, strategies, and solutions. Linn, Baker, and Dunbar (1991) do not believe that these constructed tasks guarantee that the activities are cognitively complex.

Linn, Baker, and Dunbar (1991) support the use of cognitively complex classroom activities for students. Their caution is for educators and developers of assessments who assume that these performance tasks will guarantee classroom activities that are more conducive to complex learning and thinking processes. Linn, Baker, and Dunbar cite a report released by the National Academy of Education's Committee which stated:

> It is all too easy to think of higher-order skills as involving only difficult subject matter as, for example, learning calculus. Yet one can memorize the formulas for derivatives just as easily as those for computing areas of various geometric shapes, while remaining equally confused about the overall goals of both activities.

They advise researchers to analyze the processes as well as the student products to ensure cognitive complexity. The Principle 4 validity criteria in Table 5 depicts the various levels that should be considered when one addresses the use of cognitive complexity.

TABLE 5 PRINCIPLE 4: VALIDITY CRITERIA MATRIX

COGNITIVE COMPLEXITY

Do the performance assessment tasks effectively represent cognitive complexity?

Validity Criteria	Multiple/Integrated Steps	Depth of Explanation
Level 1	Test developers do not examine the tasks to determine if they are cognitively complex.	Test developers do not examine the tasks to ensure that students are required to supply in-depth explanations of their responses.
Level 2	The school district assembles an expert panel to determine if the tasks are cognitively complex. A majority of panel members agree that the items are cognitively complex and require students to perform multiple or integrated steps to arrive at the correct responses. The tasks which do not satisfy this criteria are either revised or eliminated before test administration.	The school district assembles an expert panel to determine if the tasks require students to supply in-depth responses. A majority of panel members agree that the items require students to supply in-depth explanations of their responses. The tasks which do not satisfy the criteria are either revised or eliminated before test administration.
Level 3	Level 2 requirement satisfied. The instrument is field-tested. An expert panel reviews the field-test data to determine if student products required a multiple-step or integrated process. The tasks which do not satisfy the criteria are either revised or eliminated.	Level 2 requirement satisfied. The instrument is field-tested. An expert panel reviews the field-test data to determine if student performance on the tasks actually required an explanation. The tasks which do not satisfy the criteria are either revised or eliminated.

Cognitive complexity is one of the issues at this development phase in the cycle which researchers cite as important, but have not yet come to consensus about its meaning of application. Attention to these sub-points should help improve the quality of the tasks that are developed for performance assessment instruments.

Principle 5: Concrete Scoring Rubric

The design of performance assessment instruments places scrutiny on the credibility of judgment with which performance assessment tasks are scored. Performance assessment tasks purport to allow students to demonstrate the breadth of their proficiency with subject matter knowledge. Therefore, the scoring rubric must encompass an expansive response range for each item.

Some educational reformers believe that developing scoring rubrics before the assessment is administered limits the scope of the assessment and penalizes students who might tend to view a task in a nontraditional or innovative manner (NCTM, 1989; Wiggins, 1989). Baker (1991) calls the practice, 'til later scoring, because the rubric is designed after student papers are reviewed. She appreciates the position of educational reformers who are trying to encompass the full breadth of students' responses, but does not support the practice of generating tasks without knowing how or whether they can be credibly scored.

Educational reformers believe that performance assessment instruments are suitable for use in accountability issues. The word accountability implies that there is a body of agreed upon expert knowledge for which someone is responsible for teaching. One would want an scoring rubric for a test of accountability to assess students'

knowledge on those agreed upon traits. If the scoring rubric for the assessment instrument depends upon the responses given by the students tested, who is really responsible for the dissemination of the knowledge?

Advocates of performance assessments often criticize multiple-choice test items because the items only have one correct response. On the other hand, they often laud performance assessment's expansive scoring rubrics as proof of the superiority of this assessment method over standardized multiple-choice tests. To the psychometric community, the issue of scoring relates to the clarity of the items. The psychometric community recommends that educators approach test development with definite ideas as to what constitutes a correct or desired response (Mehrens & Lehmann, 1991). The instruction, assessment, and scoring rubric should all reflect that standard of knowledge. Therefore, administering any test in which the scoring criteria being assessed is not previously agreed to might have ambiguous learning consequences for those being tested. Such an educational assessment situation might even cause inequity of instruction and access problems within a single classroom; even more so in a large-scale testing environment.

Principle 5 asks the question, "Did the developers of the performance assessment instrument design a scoring rubric that maximizes the likelihood of student responses being accurately and consistently judged?" Scorer training and rubric discrimination are the two aspects of this concepts in which one determines the adequacy of the validity concept. Table 6 contains the criteria matrix for this principle.

TABLE 6 PRINCIPLE 5: VALIDITY CRITERIA MATRIX

SCORING RUBRIC

Does the scoring rubric maximize the likelihood of student responses being accurately and consistently judged?

Validity Criteria	Training and Scoring	Rubric Discrimination
Level 1	Raters are either not trained or poorly trained to use the scoring rubric. The district does not ensure that the raters clearly understand the scorepoints and are applying the rubric consistently.	The criteria, which determine the scorepoints on the performance assessment rubric, are not clearly defined. The scorers are not given practice tasks, videos, or papers which illustrate the various score points.
Level 2	The district establishes a interrater agreement level of at least 80%. Raters are trained to apply the rubric so as to maintain that standard on a majority of the items scored.	The criteria, which determine the scorepoints on the performance assessment rubric, are clearly defined. The scorers have access to at least one practice task, video, or paper which illustrates each score points.
Level 3	Level 2 requirements satisfied. Raters' performance is checked at the beginning of scoring each day and midway through the session to ensure the specified interrater agreement level.	The criteria, which determine the scorepoints on the performance assessment rubric, are clearly defined. The scorers have access to multiple practice tasks, videos, or papers which illustrates each score point.

Attention to these aspects of scoring will maximize the likelihood that student responses can be accurately and consistently judged.

Principle 6: Fairness

The issue of fairness is an important aspect of validity for the measurement community (Linn, Baker, & Dunbar, 1991). One aspect of fairness involves confirming that the inferences drawn from students' scores are only dependent on their ability. Another criterion of fairness speaks to instructional validity or the opportunity to learn (Linn, 1994). Violation of either of these aspects of fairness will compromise the appropriateness of the inferences one could make about the assessment.

Some educational reformers propose that replacing multiple-choice items with performance assessment tasks will yield more accurate assessment of minority students' capabilities. Linn, Baker, and Dunbar do not believe that replacement of multiple-choice testing formats with performance assessments instruments will guarantee the equity of the measurement process. They cite several assessment examples in which the mean group differences between African American and Caucasian students were essentially the same on performance assessment and multiple-choice tests. Feinberg (reported in Linn, Baker, and Dunbar even cites a case where mean group differences between African American and Caucasian students doubled when the test was corrected for the low reliability of graded performance assessments.

Linn, Baker, and Dunbar (1991) hypothesize that differences in mean group achievement might stem in part from the familiarity students have with the tasks. Therefore, the issue of instructional bias is encompassed by the fairness principle. If

students are assessed in a content areas in which they have not been exposed, their low test scores on the criterion might be indicative of their lack of instruction, their lack of content-level mastery or both. Therefore, the validity analysis would have to investigate the reasons for the mean-group differences to discover the reasons behind the results.

Principle 6 asks, "Did the developers of the performance assessment consider fairness in the procedures they used to design, implement, and score the test instrument?" The validity criteria for this principle investigates both contextual and instructional bias concerns. The Principle 6 matrix is in Table 7.

TABLE 7 PRINCIPLE 6: VALIDITY CRITERIA MATRIX

FAIRNESS

Did the developers of the performance assessment consider fairness in the procedures they used to design and implement the test instrument?

Validity Criteria	Contextual Bias	Opportunity to Learn Bias
Level 1	Contextual bias is an issue that is not addressed during performance assessment development.	The opportunity to learn issue is not addressed during performance assessment development.
Level 2	The district assembles an expert panel to review the tasks to determine if the wording might cause bias. Tasks which are deemed biased are either rewritten or replaced with unbiased tasks.	The district assembles an expert panel to review the tasks to determine if students have the opportunity to learn the material. If the experts believe that the majority of the districts students do not have the opportunity to learn the material, the tasks are revised or eliminated.
Level 3	Level 2 criteria is satisfied. The expert panel reviews data from the field-test to determine if the wording might have caused bias. Tasks deemed biased are replaced with unbiased tasks that have been field-tested.	Level 2 criteria is satisfied. The expert panel collects and reviews data to determine if the majority of the districts' students have the opportunity to learn the material being tested. These data might consist of random samples of teachers' lesson plans, homework assignments, or student or teacher surveys. If the reviewers believe that the majority of the students have not had the opportunity to learn an item, it is eliminated from the assessment.

The issue of fairness in performance assessment instruments is one with which the educational community is still grappling. Educational reformers and psychometricians both believe that assessment devices should be an indication of student knowledge and capabilities. Each group however, has its own standard as to the procedures and processes that might help promote the concept. This issue is one that researchers must address if the educational community is going to use performance assessments in large scale testing situations.

Principle 7: Test Security

Stiggins (1989) coined the phrase, "A kid can hit any target he can see." On the surface, this seems like a perfectly reasonable idea. All educators would agree that students should have been exposed to any idea for which they will later be tested. However, in considering issues of validity, student's exposure to identical or nearly identical tasks on an assessment instrument can invalidate the inferences concerning students' knowledge on the criteria. Principle 7 asks, "Is the performance assessment instrument secure?" This principle investigates the issue of test security, task re-use, and teaching to the test.

Some promoters of performance assessment instruments expect and even recommend that the assessment's criteria and in some cases the very tasks by which students will be judged be made public (NCTM, 1989; Wiggins, 1993). They believe it is unfair to judge students by performance standards with which they are not familiar. Others supporters of performance assessment instruments even believe that children should be taught the tasks or a very close approximation of the tasks from the actual

assessment because they represent skills or abilities that are valued (NCTM, 1989; Resnick & Resnick, 1989). Some of the proponents of performance assessment instruments often cite the availability of tasks to examinees as one of the advantages of this test format (NCTM, 1989; Resnick & Resnick; Stiggins, 1989).

These supporters of performance assessment instruments are using arguments which touch on a long-standing controversy in the measurement community, *teaching to the test* (Mehrens & Lehmann, 1991). Teaching to the test is a situation in which instruction is often focused more on the items or tasks being measured than the school's curriculum (Mehrens & Kaminski, 1989). Teaching too closely to test item objectives means that one cannot make accurate inferences to the broader domain, because the test has become the domain of interest (Mehrens & Lehmann). In addition to the validity problem teaching to the test causes, it can also narrow or limit the classroom content and instruction (Bracey, 1987).

The design of performance assessment instruments also causes a validity threat to the security of the test. Performance assessment instruments usually only have a small number of tasks because of the in-depth responses students are expected to supply. Even if one believes these instruments measure higher-order-thinking skills, it might be suspect to reuse the same task to assess the same higher-order-thinking processes in another group of students. For if it is possible, as some believe, to memorize answers to multiple-choice test items, it would also seem possible to memorize a procedure on a performance assessment instrument (Mehrens, 1992). The limited number of items generally included on performance assessment instruments would make it difficult to keep the tasks secure once the test has been administered.

Developers of performance assessment instruments are faced with serious security issues in regards to the validity of performance assessment instruments. The development and scoring costs of this assessment format make it impractical to use them only once before they are discarded. However, threats to the instrument's validity can render performance assessment instruments virtually useless when they are administered in a large scale testing situation for the second time.

The physical prospect of keeping performance assessment instruments secure, is a issue that must be considered in any testing situation. It is important that the actual tests are kept in a location where they are not accessible to the general school population and the public before administration. If the test is going to be used in either a high-stakes or accountability environment, this issue is especially sensitive. This principle investigates the issue of test security, task re-use, and teaching to the test. All of these concerns would affect the general security of the test. The validity criteria matrix for this principle is in Table 8.

TABLE 8 PRINCIPLE 7: VALIDITY CRITERIA MATRIX

TEST SECURITY

Is the performance assessment instrument secure?

Validity Criteria	Task Security	Task Re-use	Teaching to the Test
Level 1	School personnel and the general public have access to the actual test before test administration. The school administration allows copying of the test.	The school district does not have a policy concerning task reuse.	The school district does not address the issue of *teaching to the test* in department or staff meeting. The district appears not to be aware of the issue.
Level 2	School personnel and the general public have access to task items before administration. The school administration does not allow copying of the test.	The district has a policy concerning task reuse. The policy allows for reuse of 50% or more of the non-anchor tasks.	The district has a policy concerning of teaching to the test. Staff documents recommend use of instructional materials that do not duplicate assessment tasks.
Level 3	Access to the test is limited to only a select number of key school personnel before test administration. School personnel and the general public can review task items after administration. Copying of the test is not allowed.	The school district only reuses performance anchor tasks that are used for comparability purposes. The other tasks on the performance assessment are replaced with parallel tasks which are comparable to the original tasks.	Level 2 criteria satisfied. The district routinely samples instructional materials to monitor compliance to the policy. In addition, teaching staff is periodically informed about the *teaching to the test* policy.

The seven principles of the Performance Assessment Validity Baseline Criteria will be critical in the examination of these validity issues.

Summary

The development of the Performance Assessment Validity Baseline Criteria uses a procedure which seeks to incorporate the concerns of both educational reformers and the middle of the road psychometricians. The seven performance principles examine some of the critical issues concerning the validity of performance assessment instruments.

The Performance Assessment Validity Baseline Criteria will not answer all of the psychometric concerns surrounding the validity of performance assessment instruments. It is only a first step of many toward assessing their validity. The zeal with which performance assessments are being adopted in schools, necessitates psychometricians to become involved in participating in the foundational research that is needed to determine if performance assessments offer educational value.

The Oak Hill School District was moving toward replacing standardized multiple-choice tests with performance assessment instruments in the lower elementary grades. They developed an assessment believed to address their district's basic concerns. Although they were concerned with the validity of the Second Grade Mathematics Performance Assessment, they did not have a formal plan for investigating this important concept. The Performance Assessment Baseline Validity Criteria was developed to address a district's validity concerns. Chapter 5 will evaluate the quality of the validity evidence for the Second Grade Primary Mathematical Performance Assessment.

CHAPTER 5

APPLICATION OF THE PERFORMANCE ASSESSMENT VALIDITY BASELINE CRITERIA

Wiggins believes the first step out of the controversy surrounding performance assessment instruments is to provide educators with models and criteria for evaluating performance assessments (Brandt, 1992). He concludes that the next logical step is to apply such a model to an actual performance assessment instrument.

Chapter 4 developed the Performance Assessment Validity Baseline Criteria. This chapter applies this model to the Second Grade Primary Mathematics Performance Assessment in an effort to assess the validity of the instrument. Application of this model might aid practitioners who are in need of guidelines and procedures for evaluating performance assessments. This procedure is in response to the recommendations proposed by Wiggins.

Validity Scoring Index Sheet

Application of the Performance Assessment Validity Baseline Criteria requires an evaluation team to determine the quality of the validity evidence for the seven principles. Strict adherence to the established procedure requires the formation of an expert panel to evaluate the validity evidence. For this project however, it was important to determine how well the procedure might work. Therefore for this study, the researcher served as the

expert panel and evaluated the validity evidence for the Second Grade Primary Mathematical Performance Assessment. If a school district, however, was going to use the Performance Assessment Validity Baseline Criteria to evaluate a performance assessment instrument, the following discussion would help the district in selecting the best experts necessary to do the investigation.

Composition of Expert Panel

The expert panel is used to judge the proficiency with which the validity evidence conforms to the operationalized concepts . This expert panel reviews the validity of evidence and assigns a rating according to the strength of expert agreement. This method, relies on the opinion of an expert panel because of the lack of agreement in the educational community concerning the validity concepts suitable for performance assessment instruments. The judgement each expert makes will reflect their beliefs and theoretical training. Therefore, the composition of the expert committee is most important.

The Performance Assessment Developmental Committee included psychometricians, content specialists, and classroom teachers. All of these individuals were supportive in varying degrees, concerning the use of performance assessment instruments in large scale educational settings. Some of the members believed that performance assessment instruments would solve many of the district's assessment concerns. Others were more reserved in their judgments. All of them, however, were committed to the importance of the task and brought their own personal expertise to the table to work on the process.

One can view an expert panel as a type or focus group for which the topic of discussion is, "How valid is the performance assessment instrument?" Focus groups are usually limited to 15 members. It is thought that a group much larger than this inhibits discussion. Therefore, it is recommended that expert panel for evaluating the validity evidence should have a maximum of 15 members. The panel should include an equal number of psychometricians, content specialists, and class room teachers.

The psychometricians, on the committee, should have a research background similar to the middle of the road researchers. They should hold a degree in psychometrics or measurement, be well read on the various educational positions held in the field concerning the validity of performance assessment instruments, and have practical experience in test design and administration.

The content specialists should have an advanced degree in their profession and extensive experience in the profession. They should also be aware of the latest research in their profession and be actively involved in the field.

The teachers on the panel should have extensive experience in their profession in general and in the district in particular. The teachers should have at least three years experience teaching the grade for which the test is being developed. It would also be beneficial if the teachers had a minimum of five years in the district. Experience with the local district is important because *in-house* knowledge of the district's goals and the type of students in the population will be critical to development of the test instrument. Ideally, the teacher should have experience in teaching the grade levels both preceding and following the grade level of the test. That would give the teachers experience concerning the capabilities students bring into the grade for which the test is being

developed, and knowledge of the content requirements that will be required in the next grade as well. They should also be aware in the current trends and research in their field.

A district might include some external experts, those outside of the district, to staff some of the positions on the panel in which the district lacks expertise. It would be best to limit the number of external experts to only one per category on the committee, since districts developing their own performance assessment instruments are, in most cases, choosing to do so because they want a unique test. Once the district has the expert panel in place, the group can begin to judge the validity evidence for the performance assessment.

Levels of Validity Evidence

The Performance Assessment Validity Baseline Criteria considers evidence for the seven principles. The evidence is categorized into the following three categories: (a) Level 1-the panel ignores or gives cursory attention to the validity principle, (b) Level 2-the panel reviews theoretical aspects of validity principle, and (c) Level 3-the panel reviews student products or other evidence supportive of the validity principle and satisfies Level 2 criteria. If the validity evidence is such that a majority of these experts from diverse fields can agree on a level for the criteria, then that level score is assigned. If the decision of the experts is such that a majority decision cannot be reached, then the next lower level score should be assigned for that aspect of the validity evidence.

Some researchers might not find the use of an expert panel as a means of assigning a score for the validity evidence a very satisfying solution. However, many of the ideas discussed on the Performance Assessment Baseline Validity Criteria, are not ones where there is a consensus as to the terms or their application. Since these ideas are

still being discussed in educational circles, the use of an expert panel to formalize the process is a method of forcing consensus at the local level.

This research study also acknowledges that certain concepts identified on the Performance Assessment Validity Baseline Criteria might not be satisfying to those who disagree with psychometric interpretations of validity. Although disagreements exist in educational circles as to what aspects of validity can appropriately be applied to the study of performance assessment instruments, a research study must be supported by a theoretical framework. The psychometric concerns of the middle of the road researchers form the conceptual framework supporting this dissertation study. Therefore, the project will view the validity of performance assessment instruments from that perspective.

The Performance Assessment Validity Baseline Criteria is comprised of seven principles designed to evaluate the quality of performance assessment instruments. In addition to each major point, the criteria also includes subpoints that should be considered for the further investigation of each validity principle. The validity for the subpoint is rated on a three point scale according to the quality of the validity evidence that accompanies each major principle.

This next section of the chapter presents the analysis of validity evidence that was gathered from applying the Performance Assessment Validity Baseline Criteria to the Second Grade Mathematics Primary Performance Assessment. Each of the principles is discussed as part of the analysis.

Principle 1: Domain Adequately Defined

Principle 1 asks, "Is the performance domain for the test adequately defined?" This issue seeks to determine if the construct of the assessment has been specified in a manner which permits adequate identification of skills, knowledge, or performance behaviors thought critical to this domain.

Validity Evidence for Principle 1

The Oak Hill School Board charged the Primary Assessment Developmental Committee with designing a performance assessment instrument for the district's second graders. The committee's participants were local staff members who had intimate knowledge of the intent and the focus of the Oak Hill School District's second grade mathematics program. They reviewed the Oak Hill Mathematics Curriculum to select the key instructional elements they believed critical for success in mathematics (Pryor, 1994c). The new assessment, the Second Grade Primary Mathematics Performance Assessment, was designed to determine the proficiency with which second graders were learning the Oak Hill Mathematics Curriculum.

Once the instructional objectives were identified, the committee reduced these objectives to student behaviors the committee believed were salient, identifiable, and assessable. The members tried to ensure that all of the identified task areas were ones which adequately addressed the domain (Pryor, 1994a; 1994b).

The composition of the Primary Assessment Developmental Committee aided in the development of the assessment. The committee included mathematics curriculum specialists, who advised the participants as to district goals; teachers, who attested to the level and *teachability* of the tasks; and Research and Development staff, who provided

psychometric expertise. The committee knew the concerns of parents and educators. They were also aware of the skills needed at the second grade level. One of the teachers on the development committee said, "I got on the committee to protect myself. I did not want someone (else) designing a test that I had to live with" (Pryor, 1994g). The linkage between the Oak Hill Mathematics Curriculum and the instructional committee assisted the work of defining and selecting the appropriate assessment domain.

Scoring of Validity Principle 1

The procedure the developers used to determine the domain and develop the tasks, adheres to the criteria specified in level 3. The procedures used by the Oak Hill School district support a high rating of the validity evidence for Principle 1. Table 9 depicts the validity scoring matrix for Principle 1. The rating for the Second Grade Primary Mathematics Performance Assessment is bolded in the table below.

TABLE 9 PRINCIPLE 1: VALIDITY SCORING MATRIX

DOMAIN ADEQUATELY DEFINED

Is the performance domain adequately defined?

Validity Evidence Score	Domain Definition	Skill Identification
	3	3
Level 1	Test developers do not define the domain of the performance assessment test.	Test developers did not identify the skills required in the domain of interest on the performance assessment.
Level 2	The district assembles an expert panel to investigate the claims of the test developers concerning identification of the domain. A majority of the experts agree with the clarity with which the developers have identified the domain.	The district assembled an expert panel to investigate the claims of the test developers concerning identification of the domain. A majority of the experts agree that tasks identify student performances that are a part of the domain of interest.
Level 3	**Level 2 requirement satisfied. The district also examines evidence from field-tests or student products to substantiate developers' claims concerning domain identification.**	**Level 2 requirement satisfied. The district also examines evidence from field-test or student products to substantiate developers' claims concerning task identification.**

The validity evidence supports the conclusion that the efforts of the Performance Assessment Development Committee produced a high score for the validity of this principle.

Principle 2: Domain Adequately Sampled

Principle 2 is concerned with the adequacy of the performance assessment instrument's sampling frame. This principle asks the question, "Does the sampling frame for the performance assessment instrument adequately represent all of the desired elements of the domain of interest?" This is an issue where psychometricians and educational reformers might not agree.

Psychometricians believe a test must sample pertinent areas of a domain of interest in order for one to make inferences to the specified construct. Therefore, the breadth of the assessment is an important issue for psychometricians. Educational reformers believe that it is more appropriate to design a test that covers the depth of certain aspects of the domain. Consequently, educational reformers emphasize the amount of information learned about how well students perform a specific task, more so than how well the tasks represent the domain. However, this research study is based on psychometric principles in general and the work of middle of the road psychometricians in particular. Therefore, the breadth of coverage for the Second Grade Primary Mathematics Performance Assessment will be key in the determination of this validity principle.

Validity Evidence for Principle 2

Many of the district's educators, during this time, believed that the CAT was not assessing the mathematics skills Oak Hill valued (Pryor, 1994a; 1994b; 1994d). The Second Grade Primary Mathematics Performance Assessment was designed to measure the critical mathematical concepts the district believed were responsible for students' success (Pryor, 1994h). The philosophical underpinnings that defined the development of the content for the Second Grade Primary Mathematics Performance Assessment can best be described by the statement of the Director of Research and Development in Oak Hill during the time of its development. He said,

> ...Now, a whole bunch is what we teach, but out of that whole bunch, you can distill it down to a handful that if kids knew these five things in mathematics eventually they would be adequate mathematicians.

His pronouncement, although exhibiting concern for the students in the district, indicates a conceptual framework that limits the sampling frame. This might cause a threat to validity and prevent one from making valid inferences to the domain of interest.

The Performance Assessment Development Committee identified the six key instructional elements they believed produced mathematics success. The behaviors considered salient, identifiable, and assessable were used to write the eleven tasks on the Second Grade Primary Mathematics Assessment. Table 10 depicts the test blueprint for the assessment.

Table 10 Performance Assessment Test Blueprint

Conservation of numbers	1 item
Building a number	1 item
Regrouping with addition	3 items
Regrouping with subtraction	3 items
Problem solving	2 items
Sorting and classifying	1 item
Total of items	**11 items**

Scoring of Validity Principle 2

The evidence for Principle 2 supports assigning a Level 1 score for a representative sampling frame. The Primary Assessment Developmental Committee constructed tasks which appeared to assess the domain of interest. The validity evidence, however, does not indicate that developers analyzed the tasks to determine if the items on the Second Grade Primary Mathematics Performance Assessment adequately covered the issues of concern in second grade mathematics.

Psychometric theory assumes examinees' scores are an indication of how students would perform if they were hypothetically tested on the entire content domain. It is conceivable that the critical mathematics principles second graders needed to be successful in third grade are probably represented in the tasks on the assessment. It is difficult to believe, however, that the eleven items on the Second Grade Primary Mathematics Performance Assessment represent mastery of the entire second grade

curriculum. Therefore student scores on the assessment might not be useful for making valid inferences to the domain of interest.

The committee had been assigned the responsibility of developing an assessment that tested the five or six key elements of the second grade curriculum (Pryor, 1994h). This produced a sampling frame that was possibly too narrow to permit adequate inferences being made from the assessment to the wealth of knowledge covered in the second grade mathematics curriculum. Table 11 depicts the validity scoring matrix for Principle 2. The rating for the Second Grade Primary Mathematics Performance Assessment is bolded below.

TABLE 11 PRINCIPLE 2: VALIDITY SCORING MATRIX

SAMPLING FRAME

Does the sampling frame for the performance assessment instrument adequately represent the domain of interest?

Validity Evidence Score	Representative Sampling Frame
	1
Level 1	**Developers do not use a procedure to determine if the tasks on the assessment provide adequate coverage of the intended domain.**
Level 2	The district assembles an expert panel to determine if the tasks on the performance assessment sample from the pertinent areas of the domain. The panel develops a conceptual map, test blueprint, or other procedure to determine the breadth of the domain reflected in the tasks. A majority of the panel agree that the tasks adequately represent the domain.
Level 3	The district satisfies all of the Level 2 requirement. In addition, the experts examine student products to determine if the tasks actually produce samples of performance traits they believe should be observed when one has proficiency in this domain. If the pertinent areas of the domain are not covered, this limitation must be acknowledged.

The Second Grade Primary Mathematics Performance Assessment, like most other performance assessment instruments, probably suffers from a limited sampling frame. The data do not support the developers' intent to design an assessment which provides an adequate measure of students' proficiency and abilities on the Oak Hill Second Grade Mathematics Curriculum. One would suspect that this assessment probably has a sampling frame that underrepresents the domain. This problem, construct underrepresentation, is a major threat to test validity (Cook and Campbell, 1979; Lane, Stone, Ankenmann, & Liu, 1992; Linn, Baker & Dunbar, 1991; Mehrens, 1992). It is one that developers of performance assessments must address if these instruments are to be used in educational testing situations.

Principle 3: Directness of Tasks

Principle 3 asks, "Are the tasks on the performance assessment measured as directly as the construct allows?" Supporters of performance assessment and psychometricians both believe that assessments should measure cognitive abilities as directly as possible. Although both groups agree in principle, this phrase, measuring task directly, means different things to each constituency. The idea does not enjoy an agreed upon definition in the educational community.

Educational reformers view directly measured tasks as items in which students reproduce or perform the trait being assessed. Linn, Baker, and Dunbar (1991) report that many supporters of performance assessments tend to believe that the construction of performance tasks means that students' capabilities are directly measured. Traditional psychometricians, however, view directly measured tasks as students exhibiting a skill

that is not confounded by other traits that might cause measuring interference or measurement error.

Validity Evidence for Principle 3

The Second Grade Primary Mathematics Performance Assessment was designed to closely mirror the procedures, applications, and skills the district encouraged teachers to teach in second grade. The developmental committee did not evaluate the tasks to determine if the wording used to phrase the tasks might inhibit some children from displaying their capabilities. The developers just assumed that performance tasks *were good* for all students and allowed them to demonstrate their abilities in the chosen subject matter. One of the developers of the assessment said, "The research seemed to indicate that it (performance assessment instruments) was good for kids, teachers, and instruction" (Pryor, 1994a).

The Performance Assessment Development Committee did, however, attend to the issue of assessment interference. The district believed that during classroom instruction, questions were framed in a manner which allowed students to respond correctly. They believed that students' exposure to this classroom inquiry method allowed students the ability to effectively demonstrate their proficiency with mathematical concepts and ideas on an assessment that used a similar format.

The developers, therefore, decided to use an oral administration format for the assessment. The developers believed that this administration procedure was similar to questioning procedures used during classroom instruction. They also wanted to reduce the possible measurement error of students who had reading difficulties (Pryor, 1994a; 1994b). The committee also carefully constructed interview scripts for the tasks and

scripted probes to use in a variety of situations that might occur. The developers of the Second Grade Primary Mathematics Performance Assessment used these methods to reduce any extemporaneous situations that might contribute to students' capabilities being confounded by the methods used to administer the test.

Scoring of Validity Principle 3

The evidence for Principle 3 supports assigning a Level 1 score to skill interference and a Level 2 score to assessment procedure interference. The developers did not examine the items for contextual bias. They did, however, make an effort to ensure that the administration method was free from contributing measurement error to the process. The validity scoring matrix for Principle 3 is in Table 12. The rating for the Second Grade Primary Mathematics Performance Assessment is bolded in the table below.

TABLE 12 PRINCIPLE 4: VALIDITY SCORING MATRIX

CONSTRUCT MEASURED

Are the tasks on the performance assessment measured as directly as the construct allows?

Validity Evidence Score	Skill Interference	Assessment Procedure
	1	2
Level 1	**The tasks are not examined to determine if the wording of the items interfere with the knowledge area being assessed on the test.**	The tasks are not examined to determine if the assessment's administration method interferes with the traits being assessed.
Level 2	The district assembles an expert panel to determine if the wording of the items interferes with the knowledge area being assessed on the test. If a majority of the panel judges a task defective, the task is revised or discarded before test administration.	**The district assembles an expert panel to determine if the administration procedure interferes with the traits being tested. If a majority of the panel believes the procedure confounds testing content, the procedure is revised.**
Level 3	The district satisfies all of the Level 2 requirement. The performance assessment instrument is field-tested. If a majority of the panel after reviewing the field-test data determines the context interferes with the skill being assessed, the task is discarded or revised before test administration.	Level 2 requirement satisfied. The performance assessment instrument is field-tested. An expert collects evidence to determine if the administration procedure might interfere with students exhibiting their knowledge and capabilities. If a majority of the panel believes the procedure is flawed, the procedure is replaced before test administration.

The developers of the Second Grade Primary Mathematics Performance Assessment considered directly measuring tasks as one of the critical aspects of designing the instrument. The general belief that performance assessment instruments were good for students, meant that the developers did not address the issues of skill interference and administration bias directly.

Principle 4: Cognitive Complexity

Principle 4 asks, "Do the performance assessment tasks effectively represent cognitive complexity?" The principle of cognitive complexity is a criterion that does not have an agreed upon definition in the educational community. This ambiguity caused problems in assessing the validity evidence for this principle. The work of Linn, Baker, and Dunbar (1991) recommended that one investigate multiple/integrated steps as a key to this topic. This concept of cognitive complexity will be investigated from that perspective.

Validity Evidence for Principle 4

The developers of the Second Grade Primary Mathematics Performance Assessment discussed the complexity of the tasks for the test. One of the district's mathematics consultants, who was active in the NCTM, advised the developers about the types of tasks mathematicians believed supported complex learning.

The Second Grade Mathematics Primary Performance Assessment also required students to explain their responses. Students were supplied with various manipulatives to use in demonstrating their answers. Tasks that were either considered too simplistic or could be answered with a simple *Yes* or *No* were revised or eliminated (Pryor, 1994b).

Scoring of Validity Principle 4

The evidence for Principle 4 supports assigning a level 2 score for both subcategories. The Performance Assessment Development Committee spent considerable time discussing the tasks on the Second Grade Mathematics Performance Assessment (Pryor 1994a). The Performance Assessment Developmental Committee met over the course of several months to develop the conceptual framework of the test (Pryor, 1994h). Once the framework was in place, the committee used an additional five to six meetings to develop the tasks and scoring rubric. Finally, the committee examined the tasks to ensure that students were required to furnish in-depth and complete explanations and that the rubric could adequately score student responses (Pryor, 1994a). The rating for the Second Grade Primary Mathematics Performance Assessment is bolded in Table 13.

TABLE 13 PRINCIPLE 4: VALIDITY SCORING MATRIX

COGNITIVE COMPLEXITY

Do the performance assessment tasks effectively represent cognitive complexity?

Validity Evidence Score	Multiple/Integrated Steps	Depth of Explanation
	2	**2**
Level 1	Test developers do not examine the tasks to determine if they are cognitively complex.	Test developers do not examine the tasks to ensure that students are required to supply in-depth explanations of their responses.
Level 2	**The school district assembles an expert panel to determine if the tasks are cognitively complex. A majority of panel members agree that the items are cognitively complex and require students to perform multiple or integrated steps to arrive at the correct responses. The tasks which do not satisfy this criteria are either revised or eliminated before test administration.**	**The school district assembles an expert panel to determine if the tasks require students to supply in-depth responses. A majority of panel members agree that the items require students to supply in-depth explanations of their responses. The tasks which do not satisfy the criteria are either revised or eliminated before test administration.**
Level 3	Level 2 requirement satisfied. The instrument is field-tested. An expert panel reviews the field-test data to determine if student products required a multiple-step or integrated process. The tasks which do not satisfy the criteria are either revised or eliminated.	Level 2 requirement satisfied. The instrument is field-tested. An expert panel reviews the field-test data to determine if student performance on the tasks actually required an explanation. The tasks which do not satisfy the criteria are either revised or eliminated.

Supporters of performance assessment instruments often claim these instruments require students to use higher-order-thinking skills. Performance Assessment Developmental Committee considered the most current research on cognitive complexity when designing the instrument. Their efforts relate to the rating on this criteria (Pryor, 1994c; 1994h). Unfortunately, this issue is another in the realm of performance assessment instruments that must be more completely addressed and defined by educators.

Principle 5: Concrete Scoring Rubric

Principle 5 asks, "Did the developers of the performance assessment design a scoring rubric that maximizes the likelihood of student responses being accurately and consistently judged?" In the psychometric community, it is a standard test development practice to design the scoring rubric before the test is administered. In the realm of performance assessment instruments, however, having a predetermined scoring rubric is not often practiced. This is one issue where different opinions exist between the two educational communities.

Validity Evidence for Principle 5

The developers of the Second Grade Primary Mathematics Performance Assessment developed the scoring rubric as they were developing the performance assessment tasks (Shiffler, Beyer, & Sperling, 1992). Most committee members were familiar with *Math Their Way,* a conceptually oriented instructional program making heavy use of concrete manipulatives. Therefore, the teachers on the Performance

Assessment Development Committee were familiar with the type of responses students typically give to these kinds of tasks.

The committee met during the summer to develop the tasks and the scoring rubric (Pryor, 1994a). The developers also field-tested some of the tasks from the Second Grade Mathematics Performance Assessment on summer school students to determine the proficiency with which the written tasks could be scored.

During the field-tests, two committee members administered the Second Grade Primary Mathematics Performance Assessment. One rater conducted and scored the interview while the other observed and also rated the student's performance. At the end of the assessment, the raters discussed the score each of them had assigned to the student's work on each task. The committee established this two-person process because they believed being able to discuss a student's work immediately after the test was beneficial (Pryor, 1994f). These field-tested student performances were videotaped for future discussion with the committee.

When the Performance Assessment Developmental Committee reconvened, they viewed the videotapes of the assessment and discussed the tapes along with the interviewers' scores. The committee members, during these discussions, would adjust the rubric if necessary. Once the committee agreed on the scoring of a task, videotapes that clearly illustrated the various scoring points were kept for training purposes. All of the persons responsible for administering the 1990-91 Second Grade Primary Mathematics Performance Assessment were trained from the tapes of the field-test data.

The psychometricians and the classroom assessment specialist on the Primary Assessment Developmental Committee were responsible for training the raters who

would administer the 1990-91 Second Grade Primary Mathematics Performance Assessment. Raters were systematically checked to determine the degree of inter-rater reliability and their adherence to the scoring rubric. The committee wanted to ensure that each rater had a clear understanding of the rubric and could apply it equitably. During training sessions on the Second Grade Primary Mathematics Performance Assessment, raters achieved an agreement score of 80% or above on seven tasks and an inter-rater agreement of 60% on the other four tasks (Shiffler, Beyer, & Sperling, 1992).

The proficiency of the raters in scoring the Second Grade Primary Mathematics Performance Assessment was probably due to their intimate knowledge of the district's curriculum and goals. Most of them had assisted in the development of the Oak Hill Mathematics Curriculum, the Second Grade Primary Mathematics Performance Assessment, and the scoring rubric for the assessment. This knowledge along with their professional training and experiences probably added to their scoring proficiency.

Scoring of Validity Principle 5

The validity evidence for Principle 5 is essentially concerned with investigating the quality of rater training and the clarity with which the rubric demonstrated each scoring point. This principle considers the training and scoring of the assessment, and the construction of the rubric. The validity evidence supports a Level 2 score for Training, and Scoring and a Level 3 score for Rubric Discrimination. The validity matrix for Principle 5 is Table 14. The rating for the Second Grade Primary Mathematics Performance Assessment is bolded.

TABLE 14 PRINCIPLE 5: VALIDITY SCORING MATRIX

SCORING RUBRIC

Does the scoring rubric maximize the likelihood of student responses being accurately and consistently judged?

Validity Evidence Score	Training and Scoring	Rubric Discrimination
	2	3
Level 1	Raters are either not trained or poorly trained to use the scoring rubric. The district does not ensure that the raters clearly understand the scorepoints and are applying the rubric consistently.	The criteria, which determine the scorepoints on the performance assessment rubric, are not clearly defined. The scorers are not given practice tasks, videos, or papers which illustrate the various score points.
Level 2	**The district establishes a interrater agreement level of at least 80%. Raters are trained to apply the rubric so as to maintain that standard on a majority of the items.**	The criteria, which determine the scorepoints on the performance assessment rubric, are clearly defined. The scorers have access to at least one practice task, video, or paper which illustrates each score points.
Level 3	Level 2 requirements satisfied. Raters' performance is checked at the beginning of scoring each day and midway through the session ensure the specified interrater agreement level.	**The criteria, which determine the scorepoints on the performance assessment rubric, are clearly defined. The scorers have access to multiple practice tasks, videos, or papers which illustrates each score point.**

The process the Primary Developmental Committee used to develop, adjust, and ultimately train scorers was very thorough. Their procedures produced a level of inter-rater reliability that is not often reached when educators score locally produced assessments (Baker, 1991). The Oak Hill School District devised a procedure that satisfied most of the validity considerations.

Principle 6: Fairness

Principle 6 asks, "Did the developers of the performance assessment use procedures which promoted fairness and eliminated design bias?" This issue of fairness, or equity of representation for students, was one of the underlying beliefs that persuaded the Oak Hill School District to develop the Second Grade Primary Mathematics Performance Assessment (Pryor, 1994b). The issue of fairness, however, is another concept that does not have an agreed upon definition by educational reformers and psychometricians. The middle of the road psychometricians considered opportunity to learn and contextual bias as two issues that pertained to the fairness issue in the validity of performance assessment instruments. Therefore, for this research study, this term is defined in the manner prescribed by Linn, Baker, and Dunbar, (1991).

Validity Evidence for Principle 6

Many of the primary teachers believed that the CAT was not assessing skills the district valued. The district's teachers had anecdotal evidence that all students were learning, yet the CAT did not seem to indicate gains in achievement for minority students (Pryor, 1994h). The Performance Assessment Development Committee also believed that a change to a performance assessment instrument would indicate that African

American students were learning as well as Caucasian students. They perceived that mean group differences evidenced on the CAT would not be evident on the Second Grade Primary Mathematics Performance Assessment. Their definition of fairness, however, did not include an aspect of bias which can affect all students; the opportunity to learn. The developers of the Second Grade Primary Mathematics Performance Assessment apparently believed that performance assessment instruments offered reliable assessment measures for students and expanded instruction methods for teachers (Pryor, 1994e). This general attitude that performance assessments were good for students, teachers, and instruction was held by most of those on the committee (Pryor, 1994b). Therefore, the committee did not concentrate on the validity subpoints considered important for Principle 6.

Fairness: Test Bias

The developers of the Second Grade Primary Mathematics Performance Assessment believed that performance assessment tasks ensured that all students would be tested only on their subject-matter knowledge. The Primary Assessment Developmental Committee met over the course of several months to investigate the mathematical content of the tasks (Pryor, 1994f). The developers, however, did not consider the possibility that the wording of the tasks might cause contextual bias and possibly interfere with students' ability to display their knowledge of subject-matter. Although the processes the Performance Assessment Development Committee used to develop mathematical tasks was extensive, no attempt was made to investigate possible contextual bias.

Fairness: Opportunity to Learn

The developers of the Second Grade Primary Mathematics Performance Assessment did not use a procedure to determine if the students assessed had an opportunity to learn the material being tested. One might suspect that the opportunity for students to have been exposed to the tested material should have been high given that some of the developers were actually primary classroom teachers. The inclusion of primary teachers on the development team, however, was not enough to ensure that the test encompassed skills that were being taught in the class.

The district realized that there might be some initial resistance from teachers who might not be willing to adapt their teaching methods to adhere to the new curriculum. Three members of the development team said that all of the districts teachers were not completely convinced of the merits of the revised Oak Hill Mathematic Curriculum (Pryor, 1994a; 1994b; 1994c). Some research has shown that a district's accountability assessment determines to some extent the breadth and scope of classroom instruction (Resnick & Resnick, 1992; Silver, 1990). Developers believed that once the assessment was in place, the responsibility of administering the Second Grade Primary Mathematics Performance Assessment would solicit compliance from the district's teachers to move toward teaching the revised curriculum (Pryor, 1994a; 1994b). It was hoped that changing the assessment format would increase students' exposure to the revised mathematics curriculum on which the Second Grade Primary Mathematics Performance Assessment was based.

Scoring of Validity Principle 6

The validity evidence indicates that the Primary Assessment Developmental Committee failed to investigate contextual bias and opportunity to learn bias. Therefore, these issues were not addressed in the development of the Second Grade Primary Mathematics Performance Assessment. The evidence supports assigning a Level 1 score for both contextual bias and opportunity to learn bias. The validity scoring matrix for this principle is bolded in Table 15.

TABLE 15 PRINCIPLE 6: VALIDITY SCORING MATRIX

FAIRNESS

Did the developers of the performance assessment consider fairness in the procedures they used to design and implement the test instrument?

Validity Evidence Score	Contextual Bias	Opportunity to Learn Bias
	1	1
Level 1	**Contextual bias is an issue that is not addressed during performance assessment development.**	**The opportunity to learn issue is not addressed during performance assessment development.**
Level 2	The district assembles an expert panel to review the tasks to determine if the wording might cause bias. Tasks which are deemed biased are either rewritten or replaced with unbiased tasks.	The district assembles an expert panel to review the tasks to determine if students have the opportunity to learn the material. If the experts believe that the majority of the districts students do not have the opportunity to learn the material, the tasks are revised or eliminated.
Level 3	Level 2 criteria is satisfied. The expert panel reviews data from the field-test to determine if the wording might have caused bias. Tasks deemed biased are replaced with unbiased tasks that have been field-tested.	Level 2 criteria is satisfied. The expert panel collects and reviews data to determine if the majority of the districts' students have the opportunity to learn the material being tested. These data might consist of random samples of teachers' lesson plans, homework assignments, or student or teacher surveys. If the reviewers believe that the majority of the students have not had the opportunity to learn an item, it is eliminated from the assessment.

It is ironic that the Second Grade Primary Mathematics Performance Assessment received a low score for this principle. The fairness issue was one of the driving forces for the development of the assessment. Many supporters of performance assessment instruments believe that this format is a protection against bias. Unfortunately, the Primary Assessment Developmental Committee believed that supported rhetoric.

Principle 7: Test Security

Principle 7 examines the test security issue by posing the following questions, "Is the performance assessment instrument secure?" This principle encompasses the issue of task security, task re-use and teaching to the test. The psychometric community believes that knowledge of the items or tasks on an assessment prior to testing might influence the validity of the inferences one might make concerning student knowledge. Therefore, the tasks or items on an assessment instrument should be secure. This consensus is not held by all educational reformers. Wiggins (1993) so disagrees with the policy concerning test security that he wrote a rebuttal called, *The Immorality of Test Security.*

Validity Evidence for Principle 7

The 1990-91 administration of the Second Grade Primary Mathematics Performance Assessment was a unique situation because it was the pilot year for the test. The district was interested in determining the feasibility of using performance assessment instruments for large scale assessment purposes. Therefore, the policies concerning test security had not really been established. The district was waiting to receive the results from this pilot administration before it decided to make the assessment part of the accountability standard for primary students. Given these conditions, it is inappropriate

to apply *task re-use and teaching to the test* issues in this situation. For that reason, Principle 7 will only be rated on the security of the tasks.

The teachers, students, and the general community knew that the Oak Hill School District had revised their mathematics curriculum. The educational community also knew that these new standards were designed to be more in line with those proposed by the NCTM. The new curriculum had been publicized, but many in the community were still not aware that a test was actually being developed to assess these newly revised mathematics goals and objectives (Pryor, 1994f).

The unique situation caused the security of the tasks on the Second Grade Primary Mathematics Performance Assessment to be high. The actual tasks were known only to the developers of the assessment and individuals trained to score students' responses.

Scoring of Validity Principle 7

The validity evidence for task security is rated a Level 3. Only the test developers and those involved with scoring the tests knew which details concerning items. The test was also not administered by classroom teachers. Only the teams trained by the committee were aware of the tasks. That made the assessment tasks very secure. The validity scoring matrix for Principle 7 is in Table 16.

TABLE 16 PRINCIPLE 7: VALIDITY SCORING MATRIX

TEST SECURITY

Is the performance assessment instrument secure?

Validity Evidence Score	Task Security	Task Re-use	Teaching to the Test
	3	**NA**	**NA**
Level 1	School personnel and the general public have access to the actual test before test administration. The school administration allows copying of the test.	The school district does not have a policy concerning task reuse.	The school district does not address the issue of *teaching to the test* in department or staff meeting. The district appears not to be aware of the issue.
Level 2	School personnel and the general public have access to task items before administration. The school administration does not allow copying of the test.	The district has a policy concerning task reuse. The policy allows for reuse of 50% or more of the non-anchor tasks.	The district has a policy concerning teaching to the test. Staff documents recommend use of instructional materials that do not duplicate assessment tasks.
Level 3	**Access to the test is limited to only a select number of key school personnel before test administration. School personnel and the general public can review task items after administration. Copying of the test is not allowed.**	The school district only reuses performance anchor tasks that are used for comparability purposes. The other tasks on the performance assessment are replaced with parallel tasks.	Level 2 criteria satisfied. The district routinely samples instructional materials to monitor compliance to the policy. In addition, teaching staff is periodically informed about the *teaching to the test* policy.

Conclusion

The Performance Assessment Validity Baseline Criteria was developed to assist educators, researchers, and school administrators assess the validity evidence of performance assessment instruments. Currently, there is not a mechanism by which those interested in the validity of performance assessment instruments can have a common discussion of their merits. The Performance Assessment Validity Baseline Criteria could be a useful structure from which those with disparate views of the validity of performance assessment instruments can focus their attention to the pertinent aspects of validity and generate a common discussion.

The Performance Assessment Validity Baseline Criteria is also useful from a practical point of view. It can benefit practitioners trying to select a performance assessment instrument from among many alternatives. School administrators could compare various scores on the aspects of validity pertinent to the district. This way they could easily compare and contrast various instruments.

The Performance Assessment Validity Baseline Criteria could also be used by those developing performance assessment instruments. They could readily determine which areas of their tests needed adjustments. This criteria might be an instrument that could be used several times at different stages over the course of the developmental process.

The procedure was presented in a somewhat finished form. However, like any other tool, it could be improved or adapted to fit one's personal needs. The format and procedure used by the Performance Assessment Validity Baseline Criteria might be of

assistance to others interested in using performance assessment instruments to assess student achievement. Researchers could expand the baseline criteria to include other areas of validity principles of interest. They could develop a validity scoring matrix which addressed their concerns. The same method of collecting and examining evidence used by this study could be used to rate the evidence for the new principles.

Chapter 5 examined the quality of the validity evidence for the Second Grade Primary Mathematics Performance Assessment, using a qualitative procedural model based on a psychometric validity concerns. Chapter 6 examines the validity of the Second Grade Primary Mathematics Performance Assessment using traditional psychometric measures. Most educational reformers believe that traditional measurement techniques cannot be used to evaluate these assessment instruments. Other facets of the educational community believe that these techniques can measure any test that has been properly developed. The analysis in Chapter 6 was designed to investigate that issue.

CHAPTER 6

APPLICATION OF TRADITIONAL MEASUREMENT TECHNIQUES

There is a controversy in the educational community concerning the appropriateness of traditional psychometric measures in evaluating the validity of performance assessment instruments. Proponents of performance assessments often claim that psychometric theories concerning validity are incapable of evaluating creative, multi-faceted performance assessment instruments (Moss, 1992). These supporters of performance assessments might be correct in their criticism of traditional psychometric concepts for evaluating the validity of these mostly hypothesized, multi-dimensional, measurement devices (Moss). Most measurement experts would probably agree that psychometric techniques were not developed to evaluate these individualized, free-form instruments often envisioned by some educational reformers (Williams, Phillips, & Yen, 1991). Psychometric procedures, however, might be useful for analyzing the validity evidence from the Second Grade Primary Mathematics Performance Assessment, a more standardized performance assessment instrument.

The Second Grade Primary Mathematics Performance Assessment incorporated some of the similarities of multiple-choice testing formats. The administration of the assessment required raters to read students the tasks from a scripted format. Each task had only one correct response, although students could use multiple methods to arrive at

that answer. The developers designed a scoring rubric that allowed for the tasks to be reliably scored. These similarities between the Second Grade Primary Mathematics Performance Assessment and multiple-choice testing formats might allow one to use traditional psychometric techniques in evaluating the validity of a conservative performance assessment instrument.

Validity Discussion

Psychometric theory considers validity to be one of the most important aspects of test design and construction (Allen & Yen, 1979). This principle states that one can use the test scores from the assessment instruments to make inferences of students' knowledge to the construct or domain of the test (American Educational Research Association, American Psychological Association, & National Council of Measurement in Education, 1985). In other words, pronouncing a test valid for particular purposes, means that the test measures what it purports to measure. The study of validity involves making an evaluative judgment concerning the degree to which the empirical evidence and theoretical structure supporting the inferences made on the basis of a test (Messick, 1989). The concept of validity is an integrative construct that is generally composed of various types of validity evidence. The validity of inferences from an assessment instrument might change as more evidence becomes available to either support or refute the inferences that have been made. The validation procedure requires one to continuously evaluate the inferences as new evidence from theoretical and practical concerns becomes available (Messick, 1989). It is also important to note that the concept of validity is assigned to the use of the test, not the test itself (Allen & Yen, 1979).

Messick (1989) considers validity to be a unitary concept. Traditionally, the evidence in support of this concept has been classified into three related categories; content-related validity, criterion-related validity, and construct-related validity. Although these traditional groupings still exist in examining validity, Messick (1989) points out that all forms of validity actually support construct validity. The current psychometric practice of separating the types of validity related evidence into these categories, is a matter of convention that relates to the types of supporting documentation from which the inferences are made.

Content-Related Validity Evidence

Content-related validity is based on the expert judgment concerning the relevance of the test's substance to the desired behavior domain being assessed. This type of validity is concerned with the appropriateness of the domain and the tested sample in supporting the inferences made from test scores. Content-related evidence supports how well the items or tasks on the assessment actually allow one to make inferences to subject matter knowledge (Messick, 1989). The determination of the quality of the content-related evidence indicates how well the test samples from the specified content domain (Mehrens, & Lehmann, 1991).

The issue of content validity is particularly important in academic environments. Consideration of the validity evidence requires examining the content domain and the inferences one wishes to make concerning student test scores. Psychometric theory does not have a formula which allows one to judge content validity. This procedure requires

the subjective analysis of the items by expert judges as to how well items represent the domain of interest.

Criterion-Related Validity

Psychometricians base criterion-related validity on the statistical strength of the relationship between test scores on the assessment and an agreed upon external measure of the construct of interest (Messick, 1989). Criterion related evidence is generally divided into two categories; predictive validity and concurrent validity. When evaluating predictive validity, one examines how well test scores predict future performance on the criterion. In concurrent validity, one examines whether test scores are an adequate substitution for another measurement taken at the same time (Mehrens, & Lehmann, 1991). The procedures for assigning the strength of the relationship are the same, the time-period when the evidence is collected determines which type of criterion related validity one is evaluating.

The examination of criterion related validity starts with the identification of an external criterion that adequately represents the domain of interest (Crocker & Algina, 1986). The strength of the relationship between the criterion and the test is indicated by the correlation coefficient. An adequate external criterion should be (a) relevant to the process, (b) reliably measurable, and (c) free from contamination (Mehrens & Lehmann, 1991). Selection of an adequate criterion is often more difficult than the development of the test instrument. However, the strength of the criterion-related validity rests upon clear consideration of these issues.

Construct-Related Validity

Messick (1989) believes that construct validity incorporates all forms of validity which assist in the understanding, interpretation, meaning, or inference made from test scores. He also claims that construct validity subsumes all other forms of validity. This broad definition of construct validity allows Messick (1989) to argue that all information which supports the appropriateness of the inferences being made contribute to the construct validity of the assessment.

The measurement of construct validity starts with a hypothesized theory about how the construct will affect test performance. If the expected relationships between the construct and the test are not found, several problems might be evident. Possibly, the hypothesized definition of the concept might not capture the relevant idea. Sometimes it might be possible that the test is not an accurate portrayal of the concept. If the evidence is not supportive of the construct, it is often difficult to determine which problem exists in the analysis (Mehrens & Lehmann, 1991).

The reliance of construct validity on the purpose surrounding the development of the Second Grade Primary Mathematics Performance Assessment, was essential in designing this study. The validation procedures examined the types of evidence supportive of test developers' claims.

The Importance of Purpose to Validity

The purpose of the Second Grade Primary Mathematics Performance Assessment provided the basis of the statistical analysis. As described in Chapter 3, the purpose of the assessment was to (a) assess individual second graders' mathematics capabilities or

skills as defined by the curriculum, (b) provide information on individual students' mathematical abilities which were not tested on the CAT; and (c) provide district level accountability standards on the successful implementation of the district's mathematics curriculum (Shiffler, Beyer, & Sperling, 1992). In addition to these stated purposes, many educators in the district believed that the Second Grade Primary Mathematics Performance Assessment would eliminate the achievement gap between African American and Caucasian students (Pryor, 1989a; 1994h).

The study used the following data to determine if traditional statistical methods were capable of yielding evidence which could be used to examine the validity of the Second Grade Primary Mathematics Performance. These data were (a) individual task level raw student scores on the 11 tasks on Second Grade Primary Mathematics Performance Assessment; (b) individual item level raw, and scaled student scores from for the CAT subtests in Mathematics Computation and Conceptual, Reading, and Vocabulary; and (c) student ethnicity and gender data.

Content Validity of Performance Assessment

Psychometricians determine content-related validity by having content experts determine whether the items or tasks on the assessment are relevant to the desired behavior domain being assessed. Given that the Oak Hill School District did not assemble a panel of experts to review the validity issues connected to the Second Grade Primary Mathematics Performance Assessment, this research study will consider the Performance Assessment Developmental Committee as the expert panel.

The Performance Assessment Developmental Committee included experts in

psychometrics, mathematics, curriculum, and instruction. Most of them had been involved with the revision of the Oak Hill Second Grade Mathematics Curriculum. The committee developed the Second Grade Primary Mathematics Performance Assessment in response to the district's attempt to identify the five or six areas of the curriculum thought to measure critical mathematical concepts (Pryor, 1994h). The test blueprint that the developers used to design the assessment, is in Table 17.

Table 17 **Performance Area Description of Task Items**

Performance Area	**# of Items**	**Item Numbers**
Conservation of numbers	1 item	S1
Building a number	1 item	S2
Regrouping with addition	3 items	S31, S32, S33
Regrouping with subtraction	3 items	S41, S42, S43
Problem solving	2 items	S51, S52
Sorting and classifying	1 item	S6
Total of items	**11 items**	

The committee designed the assessment to encompass all of the performance areas identified in the blueprint. Therefore, if an analysis of the Second Grade Primary Mathematics Performance Assessment produced scales which matched the constructs identified by the developers, this evidence would be supportive of content-related validity. The hypotheses tested was, "The analysis of the Second Grade Primary Mathematics Performance Assessment will produce the same performance scales

identified by the developers." A factor analysis was used to test this hypothesis.

Factor Analysis

Factor analysis is a statistical method used to identify a smaller set of indicators from a larger set of constructs. This procedure can be used to represent relationships among sets of interrelated variables. Psychometricians can use this method to examine the underlying constructs of an assessment. These indicators or factors reduce larger constructs to a smaller albeit more efficient number of relationships. The resultant factors represent a linear combination that supposedly will reflect the composition of the larger domain (Norusis, 1991). Factor analysis assumes that relationships between interrelated factors will be identified from observed correlations between variables that share the same construct.

Exploratory factor analysis is a method that allows one to investigate the relationships of the items without any preconceived notions or hypothesis as to what one might expect from the analysis (Norusis, 1991). The correlations or factors indicate the way that student performances covary in the construct of interest (Messick, 1989). An exploratory factor analysis of the 11 items of the Second Grade Primary Mathematics Performance Assessment, was performed to determine the relationship of the items in the test. The rotated analysis of the data is presented in Table 18.

Table 18 **Second Grade Primary Mathematics Performance Assessment Factor Analysis**

Items	Factor 1	Factor 2	Factor 3
S41	.88148	.23603	.12392
S42	.87949	.26222	.13785
S31	.85473	.30777	.17703
S32	.84797	.28887	.20733
S33	.72268	---	.19342
S1	.64718	.30943	---
S2	.59114	---	.15000
S6	.21441	.93087	.14897
S52	.21510	.87101	.23061
S43	.25728	.14915	.93527
S51	.19242	.23688	.92734

The *rule of thumb* for this method indicates that correlations of less than .3 should not be included as part of the factor (Schmidt,1991). Therefore, the data analysis indicates the existence of three, not six constructs on the Second Grade Primary Mathematics Performance Assessment.

Once the factors are identified, they are interpreted to determine if they *fit* the description of the test developers. Table 19 lists the factors.

Table 19 **Tasks in Each Factor**

Factor 1

Problem #	**Description of Task**
S1	Conservation of numbers
S2	Building a number
S3-1	Regrouping w/addition solution
S3-2	Regrouping w/addition model
S3-3	Regrouping w/addition explaining
S4-1	Regrouping w/subtraction solution
S4-2	Regrouping w/subtraction model

Factor 2

Problem #	**Description of Task**
S5-2	Problem solving: with a picture
S6	Sort and Classify

Factor 3

Problem #	**Description of Task**
S4-3	Regrouping w/subtraction explanation
S5-1	Problem solving without a picture

The following section logically interprets the three factors of the Second Grade Primary Mathematics Performance Assessment.

Factor 1

The seven tasks in the first factor appear to involve several basic skills second graders would need to be successful in mathematics. These items assessed students' abilities in understanding numbers, and explaining addition and subtraction models. This factor accounted for 54% of the variability in the scores on the Second Grade Primary Mathematics Performance Assessment.

Although some experts might not consider the explanation and modeling of mathematical tasks a basic skill, educators who have adopted the conceptual mathematical paradigm might consider these skills essential to understanding.

Many of the Oak Hill's teachers were teaching mathematics from a conceptual viewpoint which used more of a holistic approach (Pryor, 1994a; 1994b; 1994d). Therefore, they would be less inclined to just present the algorithms and insist on rote memorization of mathematics facts. For that reason, the modeling and explanation of addition and subtraction would be a part of Factor 1, the basic building block factor.

Factor 2

The two items in this scale included the story problem solving task and a Venn diagram sorting task. This factor accounted for 13.7% of the variance on the Second Grade Primary Mathematics Performance Assessment. The successful execution of these skills would most probably be considered a demonstration of higher-order-thinking skills by mathematics teachers.

The problem solving task required students to conceptualize a solution with the aid of a graphic story card which outlined the problem. The Venn diagram sorting task required students to develop appropriate procedural rules for sorting various wooden pieces of different shapes and colors into their appropriate group. Although the problems were abstract, students were allowed to use manipulatives to represent the problem and explain their responses.

Factor 3

Factor 3 included two items which seemed to require a more complex level of mathematical ability than Factor 2. The first item required students to explain subtraction regrouping. The second task was a version of the previously discussed story problem. This task, however, was more difficult because students had to abstractly conceptualize the problem without the aid of a story card. This factor accounted for 10.3% of the variance.

The factor analysis of the Second Grade Primary Mathematics Performance Assessment revealed three distinct factors that were interpretable. Psychometricians however, routinely determine the stability of these factors or scales by examining them with a reliability analysis procedure.

Reliability Analysis

Reliability analysis is used by psychometricians to determine the degree of internal consistency of the scales of an assessment instrument. This procedure produces a reliability coefficient, of between .0 and 1.0, that theoretically evaluates the quality of the items and the degree to which they represent the same performance domain. The reliability coefficient

represents the extent to which the scales on the test are free from error variance.

If examinee scores are judged to have item homogeneity, the content sampling of the items from the larger construct have minimum sampling errors. If the coefficient indicates that the factors or subscores of an assessment are reliable, then it is permissible to use the factors as the appropriate unit of analysis (Borg & Gall, 1983). The reliability coefficient estimates the *trueness of fit* for the factors or scales in question. Table 20 lists the reliability coefficient for each factor of the Second Grade Primary Mathematics Performance Assessment.

Table 20 **Reliability of Factors**

Factors	**Reliability Index**
Factor 1	.9173
Factor 2	.9056
Factor 3	.9538

Helmstadter (cited in Borg & Gall, 1983) reported representative reliability coefficients for standardized achievement batteries. He indicated a .92 as being a moderate rating and .98 as being high. Factor 1 and Factor 3 have a moderate reliability rating. Factor 2 is slightly below the moderate level suggested.

The strength of the reliability coefficients of the three factors on the Second Grade Primary Mathematics Performance Assessment was surprising. Other research has generally indicated low inter-item correlations for performance assessment instruments (Shavelson, Baxter, & Pine, 1991a). This study however, has shown very encouraging

results for the reliability of the factors. The data for the factor analysis and the reliability analysis are in Appendix C.

Limitations of the Factor Analysis Procedure

Factor Analysis attempts to determine an expected pattern that coincides to the theory of the construct under evaluation (Messick, 1989). The procedure is usually confined to investigation which samples a higher number of students and items. The method works well when there are at least 100 cases and a larger pool of items (Nunnally, 1978). The size of the sample and the fact that the Second Grade Primary Mathematics Performance Assessment only included eleven items does make the results somewhat tenuous.

The resultant factor analysis extrapolated three factors. Two of the factors only included two items. The residual variance for factors with two items is usually not very reliable. However, the reliability analysis indicated that Factor 2 and Factor 3 had moderate reliability indexes. Even though the analysis did not support the statistical expectations, the method produced factors that could be used in other analyses.

Validity Implications

The factor analysis reveals three factors instead of the six scales indicated by the Second Grade Primary Mathematics Performance Assessment Test Blueprint. The factor analysis did not support the construct validity claims of the developers of the Second Grade Primary Mathematics Performance Assessment. The factor analysis, although not agreeing with the validity evidence, did further the general understanding of the performance assessment.

An analysis of these scales revealed a reasonable theory as to why the items clustered together. Borg and Gall (1983) caution researchers to only analyze test instruments at their factors or subscales level if the reliability coefficients are available for those scales. The interpretative analysis and the reliability analysis support using the three scales of the Second Grade Primary Mathematics Performance Assessment as the unit of analysis for the rest of the statistical studies. Therefore, the remainder of the statistics will analyze the Second Grade Primary Mathematics Performance Assessment using the factors as the unit of analysis.

Construct Validity of the Assessment

Construct validity is based on the evidence which supports the reason, intention, or purpose for which a test instrument was designed. This investigation of construct validity will examine developers' belief that the Second Grade Primary Mathematics Performance Assessment would indicate that there would not be an achievement gap between African American and Caucasian students. If the Second Grade Primary Mathematics Performance Assessment serves this purpose, one would expect mean group differences between African American and Caucasian students to be nonsignificant.

In the early 1980s, Oak Hill's Research and Development Department ascertained that achievement levels for African American and Caucasian students differed significantly from each other on the CAT. Some of the educational research during that period, indicated that Afro American Cognitive Style and the multiple-choice testing format of the CAT were responsible for observed mean group differences.

Teachers assured the school board and local administors that *all* students were

learning despite stagnant CAT scores. Many district level administrators and educators believed that changing the format to a performance assessment instrument would reveal that learning was occurring for all of the districts' students. An examination of the construct-related validity evidence tests the hypothesis, "The three factors of the Second Grade Primary Mathematics Performance Assessment will indicate equal mean group performance for African American and Caucasian students."

Psychometricians believed that mean group differences should be consistent with performance differences on the criterion. Therefore, when tests indicate mean group differences, psychometricians examine the items to determine if the test is assessing the same construct for the two groups. A statistical indication of performance differences on a task does not necessarily indicate that the item is biased. However, this incident will usually cause test developers to investigate the item for possible bias. This in turn would cause one to question the construct-related validity of the test.

Researchers must then judge if the performance difference on the criterion are relevant or irrelevant to the construct being investigated (Shepard, 1992). If higher students scores are related to higher achievement levels on the criterion, then the results would not support test bias. However, if the scores assess irrelevant information which is not directly connected to desired behavior on the criterion, then the test results could be biased.

Findings of bias would compromise the construct validity of the test instrument. This analysis will use t-tests to investigate mean group differences between African American and Caucasian students on the Second Grade Primary Mathematics Performance Assessment.

T-Tests Analysis

The two sample t-test for independent means was used to investigate the following hypothesis, "The three factors of the Second Grade Primary Mathematics Performance Assessment will indicate equal mean group performance for African American and Caucasian students." The analysis seeks to discover if the populations means are statistically equal.

Methodology

The data set of 70 students were categorized into the two pertinent sample groups; Caucasian students and African American students. Students from other ethnic backgrounds were eliminated from this analysis. The resultant sample contained 22 African American and 24 Caucasian students.

The analysis compared mean group differences between African American and Caucasian students on the factors of the Second Grade Primary Mathematics Performance Assessment. The developers of the performance assessment believed that performance differences observed on the CAT would not be evident on the Second Grade Primary Mathematics Performance Assessment. Table 21 depicts the results for each factor on the Second Grade Primary Mathematics Performance Assessment.

Table 21 **T-test Performance Assessment Data**

	Factor 1	Factor 2	Factor 3
African American Mean	13.8	6.7	4.3
African American SD	8.5	2.8	3.4
Caucasian Mean	25.0	9.1	7.3
Caucasian SD	9.5	1.5	3.5
T statistic	(4.3)	(3.54)	(3.05)
95% Confidence Interval	(-6, -16.5)	(-1.1, -3.7)	(-1, -5.1)
Significance	Yes	Yes	Yes
Effect Size (Standard Dev.)	(1.18)	(1.6)	(.86)

The detailed results of the t-tests for the factors of the Second Grade Primary Mathematics Performance Assessment are in Appendix D.

The data indicate that the mean group differences were significant for African American and Caucasian students on every factor of the Second Grade Primary Mathematics Performance Assessment. These results are very different than those expected by the test developers. This analysis does not support the hypothesis, "The three factors of the Second Grade Primary Mathematics Performance Assessment will indicate equal mean group performance for African American and Caucasian students." This validity study also performed a comparable study of the CAT to determine if mean group differences persisted on this assessment. Table 22 depicts the results of the t-test on the CAT mathematical subtests.

Table 22 **T-test CAT Data**

	Computation	**Conceptual**
African American Mean	16.6	25.5
African American SD	4.7	7.3
Caucasian Mean	20.5	32.5
Caucasian SD	3.9	3.7
T statistic	-3.2	-4.1
95% Confidence Interval	(-1.43, -6.5)	(-3.6, -10.3)
Significance	Yes	Yes
Effect Size (Standard Dev.)	(1)	(1.89)

Limitations of T-Test Procedure

The t-test requires the computation of a test statistic in order to perform a test of the hypotheses. It is therefore necessary to examine the assumptions for this procedure. The sampling distribution of t-ratios when the null hypothesis is true in this situation are (a) the scores for African American and Caucasian students are normally distributed, (b) the population variances for the two samples are equal, and (c) that the scores for African American and Caucasian students on the assessments are independent (Glass & Hopkins, 1984).

Glass, Peckham, and Saunders' study (as cited in Glass & Hopkins, 1984) indicates that the violation of the assumption of normality has almost no practical consequences in this procedure. The issue concerning the equality of the population

variances is also of little significance in a practical sense for this data set given the African American and Caucasian ratios of the population sizes and the corresponding variances. The last assumption, independence of test scores for African American and Caucasian students is satisfied. Therefore, the use of the t-test is an appropriate procedure.

Previous achievement data on the distribution of CAT scores for African American and Caucasian students indicated that the underlying distribution for each group was not normal (Pryor, 1994h). The standard deviation for each population was also unknown. However, although the data do not conform to the assumptions posed for small samples, the procedure is fairly robust (Bhattacharyya & Johnson, 1977). Therefore, although the results do indicate mean group differences on both the Second Grade Mathematics Performance Assessment and the CAT, the analysis should be cautiously accepted.

Discussion of Validity Evidence

The developers of the Second Grade Primary Mathematics Performance Assessment believed that mean group differences between African American and Caucasian students were the result of a format artifact of the CAT. Therefore, if that belief were true, and if the Second Grade Primary Mathematics Performance Assessment is valid, one would expect the t-test to yield insignificant results. The evidence from the t-tests however, demonstrates that both the Second Grade Primary Mathematics Performance Assessment and the CAT have significant mean group differences. This construct-related evidence does not support the beliefs of the test developers. The Second Grade Primary Mathematics Performance Assessment did not support the test developers'

claim that the assessment eliminated mean group performance differences between African American and Caucasian students.

Proponents of performance assessment instruments have often theorized that mean group differences between ethnic groups would be eliminated with the adoption of performance assessment instruments. Empirical research has generally not supported that finding (Shavelson, Baxter, & Pine, 1991b). This research becomes another study which indicates that changing the format of an assessment does not automatically eliminate mean group differences.

Although this statistical analysis indicated that performance differences exist for African American and Caucasian students on the Second Grade Primary Mathematics Performance Assessment, it does not indicate why that difference existed. A more comprehensive analysis would require researchers to determine if the observed performance differences on the criterion were relevant or irrelevant to the construct being investigated. Unfortunately, the data do not exist to complete that portion of the analysis. One can only state that construct-related validity evidence does not support the developers' claims. The scores indicate that the Second Grade Primary Mathematics Performance Assessment does not eliminate the hypothesized standardized testing artifact which produced mean group difference between African American and Caucasian students.

The results from the t-tests prompted the following question, "Since mean group performance differences exist between African American and Caucasian students on the Second Grade Primary Mathematics Performance Assessment and the CAT, is it possible that the two assessment devices are measuring different constructs for each sub group?"

This is the type of question that can be investigated with a correlation analysis.

Correlation Analysis

Correlation analysis is a statistical technique which is useful for determining the linear relationship between variables. It does not, however, allow one to determine the direction or causal nature of the relationship. Psychometricians sometimes use correlation to analyze the relationship of students' scores on two different test instruments. Therefore, this method was used to examine the linear relationship of the subtests of the CAT to the factors of the Second Grade Primary Mathematics Performance Assessment. Since the developers believed that the two assessments tested different constructs, the evidence should indicate a minimal linear relationship if the evidence is to support the developers' beliefs.

Methodology

This correlation analysis examined the linear relationship of students' scores on the three factors of the Second Grade Primary Mathematics Performance Assessment to the raw scores on the CAT subtests for mathematics conception and computation, vocabulary, total reading, and total mathematics. Table 23 depicts the linear association between the factors on the Second Grade Primary Mathematics Assessment and the CAT.

Table 23 Correlation of Performance Assessment to the Cat Subtests

	Total Mathematics	Total Reading	Vocabulary	Math Computation	Math Conception
Factor 1	.6318	.5902	.5910	.6212	.5517
Factor 2	.5011	.6049	.6886	.4102	.4902
Factor 3	.3027	.3212	.4217	.2126	.2199

Factor 1 of the Second Grade Primary Mathematics Performance Assessment, basic mathematical building blocks, was moderately correlated to every subtest of the CAT. Factor 1 was more correlated to the mathematical subtests than to the other tests. Factor 2, higher-order-thinking skills, was moderately correlated to four of the five possible subtest groupings. It is also interesting to note that Factor 2, which was interpreted as incorporating more higher order thinking skills, was more correlated to the reading and vocabulary test rather than the mathematics. Some might consider this evidence of invalidity. However, given that the developers of the Second Grade Primary Mathematics Performance Assessment believed that this instrument assessed mathematical skills and the ability to explain them, one could reasonably argue that a student had to possess mathematical and vocabulary skills to do well on Factor 2. Factor 3, complex higher order thinking skills, has a low correlation to the CAT subtests. Given that the developers of the performance assessment believed that it measured a different construct than the CAT, one would not expect the correlation between the test assessment instruments to be very high. Factor 3, interpreted as having the most complex

mathematical skills, had the lowest correlation. The correlation analysis appears to support the interpretation of the factor analysis.

Factor 1 of the Second Grade Primary Mathematics Performance Assessment consisted of seven items and Factor 2 and Factor 3 had two items. As a basis of comparison, two samples were randomly selected from among the CAT Subtests. One sample had four items and the other had two items. This test was done to determine if the relation of the factors of the Second Grade Primary Performance Assessment would have a similar correlation to the CAT as these randomly selected items. The correlation of the seven item sample to the CAT test was .6685. The sample of the two item sample to the CAT was .5322. These data tend to support the results that Factor 1 and Factor 2 on the Second Grade Primary Mathematics Performance share a moderate linear relationship to the CAT. This finding would cause one to question the developers claim that Factor 1 and Factor 2 of the Second Grade Primary Mathematics Performance Assessment were measuring a different construct than the CAT.

The correlation of Factor 3 to the CAT is less than the random samples of sets of two items each. This seems to indicate that Factor 3 might actually be measuring a different construct than the CAT and the other factors on the Second Grade Primary Mathematics Performance Assessment.

Discussion of Validity Evidence

If the Second Grade Primary Mathematics Performance Assessment is assessing different constructs than those tested on the CAT, one would expect the correlation of the factors on the performance assessment to have a low correlation to the CAT. If the factors of the Second Grade Primary Mathematics Performance Assessment were either

moderately or highly correlated to the CAT, these data would not support test developers claim concerning the Second Grade Primary Mathematics Performance Assessment measure different capabilities than the CAT. This would invalidate one of the beliefs as to the purpose of the performance assessment. Given the results of the correlation analysis, the claims of the test developers might be questionable for Factor 1 and Factor 2.

Supporters of performance assessment tests have long claimed that higher-order-thinking skills incorporated a multifaceted approach to knowledge. Factor 3 was previously identified as containing tasks that exhibited more complex higher-order-thinking skills. The highest variance explained in Factor 3 was only 18% while the lowest was 5%. These results would lend some support to the claims that the more complex higher-order-thinking skills might not be assessed by traditional multiple-choice tests.

Limitations of the Correlation Analysis

Correlation as a statistical method, is also a low power procedure. The correlation statistic will denote the strength of a linear relationship between variables, but it does not denote directional or causal relationships. The analysis of the Second Grade Primary Mathematics Performance Assessment, however, did lend some understanding of the relationship of the factors of the performance assessment and the CAT. The analysis also indicates that the correlation analysis can be useful for evaluating construct-related validity evidence of performance assessment instruments.

Conclusion

The research question investigated in Chapter 6 was, "Are traditional psychometric techniques able to yield useful information for evaluating the various types of validity concerns on the Second Grade Primary Mathematics Performance Assessment?" Supporters of performance assessments have long claimed that statistical methods were not capable of yielding information useful for evaluating the content, criterion, and construct validity of performance assessment instruments.

The factor analysis indicated that the structure of the Second Grade Primary Mathematics Performance Assessment had three, not six scales as proposed by the test developers. The interpretative analysis was also in conflict with the developers' claims. This analysis resulted in content-related validity evidence which did not support the inferences. The t-test analysis indicated significant mean group differences between African American and Caucasian students on all three factors of the Second Grade Primary Mathematics Performance Assessment and the CAT. Consequently, this construct-related evidence did not support the claim that the performance assessment instrument would eliminate the mean group differences and therefore support the claim that the achievement gap was a testing format artifact. The correlation analysis of the Second Grade Primary Mathematics Performance Assessment was intended to determine the degree of correlation between the performance assessment and the CAT. The correlation analysis produced mixed results in support of the validity evidence. However, these issues were a part of the overall question, "Are psychometric procedures capable of yielding useful information for a conservative performance assessment instrument like the Second Grade Primary Mathematics Performance Assessment?"

This preliminary empirical study was not able to answer all of the concerns raised about the appropriateness of using traditional psychometric measures to evaluate the validity of performance assessment instruments. However, this chapter demonstrated that there are some techniques which can contribute to the collection of validity evidence for a conservative performance assessment instrument like the Second Grade Primary Mathematics Performance Assessment. Although the statistical procedures alone were not able to investigate all of the validity issues, the procedures still provided validity evidence.

As with any study, one needs to reflect upon the benefit of the work. Chapter 7 summarizes the lessons learned from this empirical study.

CHAPTER 7

WHAT WAS LEARNED

This empirical study grew from a personal concern about the use of performance assessment instruments for accountability. Over the last few years, the rhetoric and appeal surrounding these assessment instruments have seemed almost surreal.

The call for educational reform in the 1990s, placed performance assessment instruments in the middle of an educational controversy. Supporters and critics of performance assessments had been debating the merits of these measurement instruments. The conversation usually centered around a mostly hypothetical instrument. This study was atypical because the issue of validity was examined with respect to the Second Grade Primary Mathematics Performance Assessment, an actual assessment instrument.

Some educational reformers proposed replacing multiple-choice tests with performance assessment instruments. They believed the standardized multiple-choice tests currently used to assess student achievement had to change if schools were to be transformed into paragons of learning. Psychometricians, on the other hand, were concerned because few of the discussions on performance assessment instruments addressed technical issues. Some educational reformers believed that the psychometric definitions of validity were irrelevant for performance assessments. Psychometricians however, believed that validity was an idea that was germane to the development of any test instrument.

Psychometricians and educational reformers, for the most part, had entrenched themselves in the rhetoric of their research methodologies and doctrines. Few researchers on either side were willing to listen to the concerns, theories, judgments, or expertise of the other. However, in the midst of this accusatory environment, a unique group of psychometricians emerged. These middle of the road researchers, attempted to merge the two ideologies into a blended approach which might address the concerns of the educational reformers in a psychometric framework. This empirical study attempted to follow the path set out by these middle of the road psychometricians.

This dissertation study investigated several research topics. The general purpose of the research involved analyzing the validity of performance assessment instruments. In that quest, this dissertation study developed a conceptually based validity procedure, the Performance Assessment Validity Baseline Criteria. This method capsulated some of the validity concerns expressed by middle of the road psychometricians. The Performance Assessment Validity Baseline Criteria was then used to analyze the validity of the Second Grade Primary Mathematics Performance Assessment. This study also evaluated the usefulness of traditional psychometric techniques in assessing the validity evidence of the Second Grade Primary Mathematics Performance Assessment.

In addition to the examination of the validity evidence, the empirical study also chronicled the experience of the Oak Hill School District in designing the Second Grade Primary Mathematics Performance Assessment. Knowledge of Oak Hill's procedures are important because their work probably mirrors the logical approach other school districts might make in attempting to develop a performance assessment instrument. Educators contemplating designing a local performance assessment might benefit from knowledge

of Oak Hill's experience. The remainder of this chapter will briefly discuss the results learned from each research question and the lessons learned from this dissertation study.

Research Question Number One

The first research question examined Oak Hill's reason for developing the performance assessment instrument. This question was, "What was the purpose of the Second Grade Primary Mathematics Performance Assessment?" This issue was important because it formed the crux of the validity evaluation.

Any school district attempting to develop a performance assessment instrument must first know the purpose for which the test is being designed. The assessment should be developed in response to a perceived void in the district's testing program.

The Oak Hill School District developed the Second Grade Primary Mathematics Performance Assessment because it was convinced that tests marketed during that time did not effectively measure Oak Hill's curriculum. The developers of the Second Grade Primary Mathematics Performance Assessment designed the assessment to (a) assess individual second graders' mathematics capabilities or skills as defined by their curriculum, (b) provide information on individual students' mathematical abilities which were not tested on the CAT, and (c) provide district level accountability standards on the successful implementation of the district's mathematics curriculum (Shiffler, Beyer, & Sperling, 1992). In addition to these stated purposes, many educators in the district

believed that mean group differences between African American and Caucasian students would be eliminated if the district-wide accountability test was changed to a performance assessment instrument. These were the purposes that were evaluated in the collection and analysis of validity related evidence.

Research Question Number Two

Research question number two chronicled the events surrounding the development of the Second Grade Primary Mathematics Performance Assessment. This question was, "What procedure or process did the Oak Hill School District use to develop and implement the Second Grade Primary Mathematics Performance Assessment?" This question was important because of the guidance it could provide for other school districts attempting to develop a local assessment.

The commitment and resources needed to develop a performance assessment instrument was substantial. Oak Hill School District invested approximately 600 hours over nearly two years to develop the 1990/91 Second Grade Primary Mathematics Performance Assessment. This assessment was designed to be the centerpiece of accountability testing at the lower primary grades. The work, energy, and money the Oak Hill School District invested in this process demonstrated their dedication to the project.

Any district attempting to develop a performance assessment must first be committed to the project. They must be willing to invest the time, energy, and resources

to bring their theoretical testing model into reality. A portion of the task includes identifying and developing a conceptual framework which encompasses a specific curriculum, class, or content domain the assessment will measure. The developers should know the goals and behaviors valued by the district's educators.

The Oak Hill School District designed the Second Grade Primary Mathematics Performance Assessment to assesses specific goals valued by the district. The behaviors for successful performance were defined. The developers' understanding of these performance domains enabled them to write tasks which directly assessed Oak Hills' mathematical goals. A well specified domain is a necessary precursor to the development of a valid test instrument.

Any school district attempting to develop a performance assessment instrument should have a test development committee that includes experts in psychometrics, curriculum, and the discipline being assessed. The committee should also include classroom teachers who have experience at the grade level being measured by the assessment. The professional experiences of these individuals should increase the likelihood of developing a valid performance instrument.

Oak Hill's Primary Assessment Development Committee included experts in psychometrics, mathematics, and curriculum development. The committee also included primary grade teachers. All of these experts had a critical part to play in the development of the assessment. The quality of the instrument largely depends on the expertise used in

the development stage. The knowledge of these professions contributed to the quality and integrity of the Second Grade Primary Mathematics Performance Assessment.

Research Question Number Three

The third question developed a validity procedure which incorporated the research theories of the middle of the road psychometricians. This research question was, "Does the Second Grade Primary Mathematics Performance Assessment address some of the validity concerns posed by the middle of the road psychometricians?"

This empirical study developed the Performance Assessment Validity Baseline Criteria to frame a response to that issue. This method incorporated some of the concerns of the middle of the road psychometricians into the Performance Assessment Validity Baseline Criteria, a psychometric framework for evaluating validity evidence. This procedure was then used to evaluate the validity of the Second Grade Primary Mathematics Performance Assessment.

Extracting Practical Applications from Theory

This dissertation study intended to develop a procedure whereby local educators could examine the validity of performance assessment instruments within the context of a conceptual psychometric framework. The methodology first required extracting the essence of traditional validity concepts from the statistical procedures in which they were encased. The middle of the road psychometricians had already started to work on this

process. The difficulty of this procedure occurred when this study attempted to push these theoretical constructs into an operational level whereby an actual system would evolve that would permit one to evaluate a performance assessment instrument.

It was ascertained that some psychometric icons, i.e., domain definition and representative sampling frame, although concepts most measurement professionals would agree are important for the validity of any test instrument, are not clearly defined. The literature, unfortunately, is not forthcoming as to how one might operationalize these concepts.

The misspecification of domain definition in the operational sense is particularly troubling. Psychometricians consider a test valid for a particular purpose when one is able to make correct inferences to the domain of reference from student test scores. The literature states the importance of this concept and then promptly proceeds to develop items that will test the level of ideas test developers wish to assess. This circuitous logic establishes the items as the definition of the domain for which one is going to use the items to make inferences to the domain.

A similar problem occurred when this study tried to ascertain the exact meaning of a construct. Test developers often speak of the importance of a construct and readily admit that knowledge of a construct is important in the design of any test. However, the literature is less than clear as to how one explicitly defines a construct. Discussions of a construct quickly change into a definition of observable behaviors one believes

individuals who possess this construct will exhibit. This tautological logic establishes observable behaviors as the construct for which the observable behaviors are going to be assessed.

The lack of an agreed upon theoretical basis for these ideas caused considerable difficulty in designing and operationalizing a criteria matrix for Principle 1 and Principle 2 of the Performance Assessment Validity Baseline Criteria. The study pushed on these theoretical concepts in an effort to arrive at an acceptable and practical solution to this problem. The results, although responsive and in keeping with the current literature, were not completely satisfying. The dissertation study indicates that there are some concepts, that although agreed upon universally by psychometricians in a theoretical manner, need to be more clearly developed into operational concepts.

Any school district interested in developing a performance assessment instrument should be concerned that the assessment meets a reasonable validity standard. If the test developers are unable to substantiate the inferences made as the result of administering the test, what good is the assessment? The use of the Performance Assessment Validity Baseline Criteria might enable one to make a determination of the quality of the evidence pertaining to the validity concerns surrounding performance assessment instruments.

Research Question Number Four

The final research question investigated the appropriateness of evaluating the validity of performance assessment instruments with traditional statistical procedures. This research question was, "Are traditional psychometric techniques able to yield useful information for evaluating the validity of the Second Grade Primary Mathematics Performance Assessment?" The study showed that it was possible to use psychometric techniques to examine the validity of the Second Grade Primary Mathematics Performance Assessment. These methods were useful in the assessment of the Second Grade Primary Mathematics Performance Assessment.

The factor analysis indicated that the Second Grade Primary Mathematics Performance Assessment had three factors instead of the six that were posed by developers in the test blueprint. The t-test analysis indicated that the assessment did not eliminate mean group differences between African American and Caucasian students. The correlation analysis indicated that Factor 1 and Factor 2 on the Second Grade Primary Mathematics Performance Assessment might be assessing skills that were being assessed by the CAT. This evidence indicated that certain statistical procedures were able to provide content-related and construct validity evidence.

School districts attempting to develop a performance assessment instrument should consider using traditional psychometric techniques as a measure of validity. These procedures might not be appropriate for all performance assessment instruments,

however, these methods should be applied whenever possible.

The statistical analysis used for this study was of a low power because of the type of data available. Districts who are planning to develop a performance assessment instrument could determine in advance the type of evidence and data that need to be collected for the validity study. Preplanning the analysis would permit districts to gather data that would enable the use of psychometric procedures with more statistical power. Preplanning the validity study might lend itself to a more thorough analysis of the issues of concern.

Advice for Future Developers of Performance Assessment Instruments

This dissertation study was designed to investigate several issues related to the development and validity of the Second Grade Primary Mathematics Performance Assessment. The analysis, however, pointed to some basic advice for future developers of performance assessment instruments.

The research and empirical base for performance assessment instruments is still relatively small. Consequently, one can speculate that educators might not have enough knowledge about performance assessments to support using them in large scale testing situations. The general advice for those attempting to travel along the path toward developing performance assessment instruments for accounting purposes is, "Proceed with caution." The appropriate assessment of student learning is important to the education process. This matter should not be taken lightly. It is equally important,

however, that the instruments used to measure students' achievement are valid for that purpose.

This dissertation study only examined some of the basic tenets of validity. Although these issues were important, they just scratched the surface of the technical problems that should be addressed before performance assessments become *traditional* assessment instruments. Researchers have not been able to solve the delimited domain problem inherent in performance assessment instruments. The issue of item security is another problem that has generally not been satisfactorily addressed. These and other yet to be solved problems, should make educators cautious when the suggestion is made to use performance assessment instruments as district-wide accountability measures. There have not been enough technical or empirical studies to even consider that recommendation.

Conclusion

Educators in the past, have started too many projects and then stopped midstream because of some important issue that was never considered. The Validity of the Second Grade Primary Mathematics Performance Assessment, was intended to examine some of the psychometric concerns surrounding performance assessment instruments. The educational community has been discussing the merits of these instruments for some time. This research study contributed to the knowledge base concerning the validity of

performance assessment instruments.

The study developed the Performance Assessment Validity Baseline Criteria, a methodology which combined research of both educational reformers and psychometricians. The research also demonstrated how this procedure might be used by district administrators or researchers to (a) assist in the development of performance assessments, (b) evaluate the validity of an instrument one is considering using, or (c) provide a forum for discussing the validity of performance assessment instruments. The statistical analysis used to investigate the validity of the Second Grade Primary Mathematics Performance Assessment demonstrated that these methods are capable of yielding content and construct related evidence on a conservative performance assessment instrument.

This study also chronicled the events and procedures used by the Oak Hill School District in developing the Second Grade Primary Mathematics Performance Assessment. An inside look as to how another school district developed an *inhouse* performance assessment might be of interest to practitioners.

Lastly, this empirical study demonstrated how the assessment community might move towards more of a collaborative research model instead of taking the combative stance that has permeated performance assessments discussions. This work also

illustrated how a blended approach to validity might be used to advance the knowledge concerning their possible quality and use.

Perhaps this pattern of blending research theories within a collaboration model might be useful to future researchers interested in evaluating the utility of performance assessment instruments. This Performance Assessment Validity Baseline Criteria can be expanded to address other validity concerns and issues by future researchers.

School districts are developing and implementing performance assessment instruments. State Boards of Education are incorporating performance tasks on their competency and exit exams. These actions necessitate that psychometricians become involved with the development of these assessment instruments. The educational community will benefit from the considerable knowledge and expertise psychometricians have in developing tests. This project was designed as a collaborative framework and provided an objective lens by which to address the concerns of educational reformers and psychometricians.

APPENDIX A

APPENDIX A

SECOND GRADE PRIMARY MATHEMATICS PRIMARY ASSESSMENT

S-1 CONSERVATION OF NUMBER

I am going to make two rows of beans.

Have a pile of beans on the table.
Make horizontal row of 10 beans, then make another row.
Beans should be placed in 1 to 1 correspondence.

Are there the same number of beans in each row?

If the students says yes, spread out the beans in 1 row to make it look longer

Are there the same number of beans in each row?

Have the student explain her/his answer.

Can you tell my why?

S-2 BUILD A NUMBER

Place a pile of beans on the table. Show the number card sixteen.

Ask the student what is the number on the card?

Correct the student if he/she gives the incorrect answer.

Use the beans to show me this number. Point to the six and ask what this number means? Please represent this number with the beans.

If the student performs correctly, go to S-3
If the students performs incorrectly, probe the response.

Point to the one and repeat the same procedure.

APPENDIX A

S-3 REGROUPING WITH ADDITION

Place the base ten materials on the table.
Have the student clear his/her board.
Show the card with 16 + 17 written on it.
Have pencil and paper available.

Please add these numbers. You may do it in your head or on paper.
Have the student explain how he/she got his/her answer.

Please explain how you got that answer?

Would you please show me with one of these materials a way you can solve the problem?
Find out which kind of base ten materials the student is comfortable using.

Explain what you are doing/did?
Have the student explain, using the base 10 material and the procedure he/she had just described.

S-4 REGROUPING WITH SUBTRACTION

Note: Have the materials the student preferred on the table. Show the card with 34 - 17 written on it to the student.

Please subtract these numbers. You may do it in your head or on paper.
Have the student explain his/her answer.
Can you tell me how you got your answer?

Show me 34 with your materials.

Show me, as you explain, how you subtracted 17.

APPENDIX A

S-5 PROBLEM SOLVING WITHOUT A PICTURE

I am going to tell you a story. You can use your paper and pencil, if you wish.

Place the story card in front of the student so he/she can see it as you read loud.

There are two trees in a yard. One tree has five baby birds in it. The other tree has three baby birds in it. How many feet are there altogether?

Give the student time to work.

Please explain how you solved the problem?

S-6 SORT AND CLASSIFY

Place grandma's buttons and a sorting mat in front of the student. Withold the small blue square, the small yellow circle, the large yellow square, and the small red triangle.

A. You are going to do some sorting. This circle is for the yellow pieces. This circle is for squares. Here are some colored shapes. Please choose any three pieces and place them on your sorting mat.

Probe if child places any incorrectly.

Now I am going to give you some pieces to place. Decide where each piece goes. Put it there. Then tell me why it goes there.

Hand the student one piece at a time in this order: Small blue square, the small yellow circle, the large yellow square, and the small red triangle.

Probe/prompt when necessary.

Ask the student to rethink if there is an error.

Use additional pieces if you feel you need to reassess something.

B. This time you get to sort any way you would like.

When the student has decided the labels, make out cards for the student and place above the circles over previously used labels.

Please place five pieces.

Have the student give you reasons for why he/she placed pieces where he/she did.

Select two or three pieces and have the student place them for you and have the student explain why they go there.

APPENDIX B

APPENDIX B

SCORING RUBRIC

S-1 CONSERVATION OF NUMBER

0	1	3	5
No response	Does not conserve at 10	Conserves at 10 Insufficient explanation	Conserves at 10 Provides valid explanation

S-2 BUILDING A NUMBER

0	1	3	5
No response	Cannot create place value model	Creates place value model Difficulty explaining relationship of digits	Builds and explains model Explains the relationship of the digits

S-3 REGROUPING WITH ADDITION-SOLUTION

0	1	3	5
No response	Incorrect answer	Minor error	Correct

MODELING PROBLEM

0	1	3	5
No response	Cannot model regrouping or place value model	Regrouping with prompt Creates place value model Difficulty building model	Models regrouping Creates place value model Correctly builds numbers

EXPLANATION

0	1	3	5
No response	Cannot explain regrouping or digit relationship	Difficulty demonstrating regrouping, modeling and explaining digit relationship	Explains regrouping and digit relationship Models relationship

APPENDIX B

S-4 REGROUPING WITH SUBTRACTION-SOLUTION

0	1	3	5
No response	Incorrect answer	Minor error in execution	Correct answer

MODELING

0	1	3	5
No response	Cannot model regrouping or create place value model	Models regrouping with prompt Creates place value model Difficulty building model	Models regrouping Creates place value model and correctly builds both numbers

EXPLANATION

0	1	3	5
No response	Cannot explain digit relationship, regrouping, or modeling	Difficulty showing regrouping, explaining model, or digit relationship	Explains regrouping, modeling, and digit relationship to numerical value

S-5 PROBLEM SOLVING SOLUTION

0	1	3	5
No response	Incorrect answer	Minor error	Correct answer

STRATEGY

0	1	3	5
No response	Inappropriate strategy	Partially correct	Correct strategy and explanation

S-6 SORT AND CLASSIFY

0	1	3	5
No response	Incorrect procedure	Sorts main attributes	Sorts correctly

APPENDIX C

APPENDIX C

FACTOR ANALYSIS

FACTOR MATRIX:

ITEMS	FACTOR ONE	FACTOR TWO	FACTOR THREE
S31	.90049	-.21233	
S32	.89819	-.19671	
S41	.88649	-.27433	
S42	.87162	-.29726	
S1	.69448	-.16166	-.11218
S33	.65192	-.28694	.23103
S52	.63796	.44620	-.50201
S6	.63246	.41888	-.59934
S43	.62681	.56548	.50051
S2	.54453	-.22466	.15810
S51	.60834	.63749	.42026

FINAL STATISTICS

Variable	Commun.	Factor	Eigenvalue	% Var.
S1	.521	1	5.94	54.0
S2	.372	2	1.51	13.7
S31	.857	3	1.13	10.3
S32	.845			
S33	.561			
S41	.861			
S42	.848			
S43	.963			
S51	.953			
S52	.868			
S6	.935			

APPENDIX C

VARIMAX ROTATION

Varimax converged in 5 iterations.

ROTATED FACTOR MATRIX:

Items	Factor 1	Factor 2	Factor 3
S42	.88148	.23603	.12392
S41	.87949	.26222	.13785
S31	.85473	.30777	.17703
S32	.84797	.28887	.20733
S33	.72268		.19342
S1	.64718	.30943	
S2	.59114		.15000
S6	.21441	.93087	.14897
S52	.21510	.87101	.23061
S43	.25728	.14915	.93527
S51	.19242	.23688	.92734

FACTOR TRANSFORMATION MATRIX:

	Factor 1	Factor 2	Factor 3
Factor 1	.82077	.43025	.37579
Factor 2	-.55900	.46935	.68354
Factor 3	.11771	-.77110	.62574

APPENDIX C

RELIABILITY ANALYSIS

Scale One

	Mean	Std. Dev	Cases	Alpha if Item Deleted
S42	2.3143	1.7409	70	.8924
S41	2.3429	1.8089	70	.8910
S31	2.5143	1.7672	70	.8911
S32	2.6143	1.8589	70	.8915
S33	3.0429	1.9740	70	.9199
S1	3.5714	1.7074	70	.9151
S2	4.0286	1.7526	70	.9275

Alpha=.9173

Scale Two

	Mean	Std. Dev	Cases	Alpha if Item Deleted
S6	4.0435	1.2300	69	-
S52	4.0290	1.2123	69	-

Alpha=.9056

APPENDIX C

Scale Three

	Mean	**Std.Dev.**	**Cases**	**Alpha if Item Deleted**
S43	2.9714	1.9410	70	
S51	3.2857	1.9046	70	

Alpha=.9538

APPENDIX D

APPENDIX D

T-TEST ANALYSIS

T-tests for Independent Samples of Ethnicity

Factor 1

Variable	# of Cases	Mean	Std. Dev.	SE of Mean
African American	23	13.7816	8.480	1.768
Caucasian	24	25.0417	9.471	1.933

Mean Difference = -11.2591

Levene's Test for Equality of Variances: F=1.106 P=0.299

Variances	T-value	2-Tail Significance	SE of Difference	Deg. of Freedom
Equal	-4.29	.000	2.626	45
Unequal	-4.3	.000	2.620	44.8

T-tests for Independent Samples of Ethnicity

Factor 2

Variable	# of Cases	Mean	Std. Dev.	SE of Mean
African American	22	6.7727	2.827	.603
Caucasian	24	9.1250	1.541	.315

Mean Difference = -2.3523

Levene's Test for Equality of Variances: F=10.323 P=0.002

Variances	T-value	2-Tail Significance	SE of Difference	Deg. of Freedom
Equal	-3.54	.001	.664	44
Unequal	-3.46	.002	.680	31.84

APPENDIX D

T-tests for Independent Samples of Ethnicity

Factor 3

Variable	# of Cases	Mean	Std. Dev.	SE of Mean
African American	23	4.2609	3.374	.704
Caucasian	24	7.3333	3.522	.719

Mean Difference = -3.0725

Levene's Test for Equality of Variances: F=.565 P=.456

Variances	T-value	2-Tail Significance	SE of Difference	Deg. of Freedom
Equal	-3.05	.004	1.007	45
Unequal	-3.05	.004	1.006	45

APPENDIX D

T-tests for Independent Samples of Ethnicity

MCPMPRAW

Variable	# of Cases	Mean	Std. Dev.	SE of Mean
African American	23	16.5652	4.727	.986
Caucasian	24	20.5417	3.901	.796

Mean Difference = -3.9764

Levene's Test for Equality of Variances: F=1.22 P=0.295

Variances	T-value	2-Tail Significance	SE of Difference	Deg. of Freedom
Equal	-3.15	.003	1.262	45
Unequal	-3.14	.003	1.267	42.69

T-tests for Independent Samples of Ethnicity

MCONRAW

Variable	# of Cases	Mean	Std. Dev.	SE of Mean
African American	23	25.5217	7.279	1.518
Caucasian	24	32.4583	3.659	.747

Mean Difference = -6.9366

Levene's Test for Equality of Variances: F=21.020 P=.000

Variances	T-value	2-Tail Significance	SE of Difference	Deg. of Freedom
Equal	-4.15	.000	1.67	45
Unequal	-4.10	.000	1.692	32.14

REFERENCES

REFERENCES

Allen, M. J., & Yen, W. (1979). Introduction to measurement theory. Monterey: Brooks/Cole Publishing Company.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Baker, E. L. (1991). Expectations and evidence for alternative assessment. Paper presented at the annual meeting of the American Education Research Association, Chicago, IL.

Baxter, G. P., Shavelson, R. J., Goldman, S. R., & Pine, J. (1992). Evaluation of procedure-based scoring for hands-on science assessment. Journal of Educational Measurement, 29, 1-17.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Quality and diverse student impact. Journal for Research in Mathematics Education, 24, 190-216.

Bhattacharyya, G. K., & Johnson, R. A. (1977). Statistical concepts and methods. (pp. 186-333). New York: John Wiley & Sons.

Borg, W. R., & Gall, M. D. (1983). Educational research: An Introduction (4th ed.). (pp. 270-316). New York: Longman Inc.

Bracey, G. W. (1987). Measurement-driven instruction: Catchy phrase, dangerous practice. Phi Delta Kappan, 9, 683-86.

Brandt, R. (1992). On performance assessment: A conversation with Grant Wiggins. Educational Leadership, 49, 5-37.

CTB/McGraw-Hill. (1985). California Achievement Test. Monterey, CA: McGraw-Hill, Inc.

Cannell, J. J. (1988). How public educators cheat on standardized achievement tests. Albuquerque: Friends for Education.

Cook, T. D., & Campbell, D. T. (1979). Quasi-Experimentation: Design and analysis issues for field settings. (pp. 37-91). Boston: Houghton Mifflin Company.

Cizek, G. J. (1991). Innovation or enervation? Phi Delta Kappan, 72, 695-99.

Cole, B. P. (1987). College admissions and coaching. The Negro Educational Review, 38, 125-135.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: CBS. College Publishing.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), Intelligence: Measurement, theory, and public policy (pp. 147-171). Urbana: University of Illinois Press.

Dossey, J. A. (1989). Transforming mathematics education. Educational Leadership, 47, 22-24.

Fairtest: National Center for Fair and Open Testing. (1988). What is Authentic Evaluation (Research rep.). Cambridge: Author.

Frechtling, J. A. (1991). Performance assessment: Moonstruck or the real thing? Educational Measurement: Issues and Practice, 10, 23-25.

Glass, G. V., & Hopkins, . D. (1984). Statistical methods in education and psychology (2nd ed.)

Engelwood Cliffs, N.J.: Prentice-Hall.

Hilliard, A. (1989). Teachers and cultural styles in a pluralistic society. NEA Today, 7, 65-69.

Hipps, J. A. (1993). Trustworthiness and authenticity: Alternate ways to judge authentic assessments. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Kane, M. T. (1992). An argument-based approach to validity. Psychological Bulletin, 112 (3), 527-535.

Kirst, M. (1991). Interview on assessment with Lorrie Shepard. Educational Researcher. 20, 21-23, 27.

Lane, S., Stone, C. A., Ankenmann, R. D., & Lui, M. (1992). Empirical evidence for the reliability and validity of performance assessments. Paper presented at annual meeting of the American Educational Research Association, Chicago, IL.

LeMahieu, P. G. (1984). The effects on achievement and instructional content of a program of student monitoring through frequent testing. Educational Evaluation and Policy Analysis, 6, 175-187.

LeMahieu, P. G. (1992). Using student portfolios for a public accounting. School Administrator, 49, 8-15.

LeMahieu, P., & Leinhardt, G. (1985). Overlap: Influencing what's taught: A process model of teachers' content selection. Journal of Classroom Interaction, 21, 2-11.

LeMahieu, P., & Leinhardt, G. (1986). Up against the wall: Psychometrics meets praxis. Educational Measurement: Issues and Practice, 5, 12-16.

Lesh, R., & Lamon, S. J. (Eds.). (1992). Assessment of authentic performance in school mathematics. Washington, D. C.: American Association for the Advancement of Science.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. Educational Evaluation and Policy Analysis, 15 (1), 1-16.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. Educational Researcher, 23 (9), 4-14.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20 (8), 15-21.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), Educational Measurement (pp. 119-158). Washington, D.C.: American Council on Education.

Madaus, G. F., & Tan, A. G. (1993). The growth of assessment. In G. Cawelti (Ed.), Challenges and achievements of American Education. Alexandria, VA: Association for Supervision and Curriculum Development.

McKnight, C. C., Crosswhite, F. J., Dossey, J. A., Kifer, E., Swafford, J. O., Travers, K. J., & Cooney, T. (1987). The Underachieving Curriculum: Assessing U.S. School Mathematics from an International Perspective. A National Report on the Second International Mathematics Study. Washington, D. C.: National Science Foundation.

McPhail, I. P. (1978). A Psycholinguistic approach to training urban high school students in test-taking strategies. Journal of Negro Education, 47, (2) , 168-76.

McPhail, I. (1981). Why teach test-wiseness. Journal of Reading, 25, 32-38.

Mehrens, W. A. (1992). Using performance assessment for accountability purposes. Educational Measurement: Issues and Practice, 11, 3-9, 20.

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? Educational Measurement: Issues and Practice, 8, 14-22.

Mehrens, W. A., & Lehmann, I. J. (1991). Classroom testing: The planning Stage. Measurement and Evaluation in Education and Psychology. (pp. 49-80). Fort Worth: Harcourt Brace Jovanovich College Publishers.

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012- 1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). New York: MacMillan.

Millman, J., Bishop, C. H., & Ebel, R. L. (1965). An Analysis of test wiseness. Educational and Psychological Measurments, 25, 707-26.

Mimms, E. M. (1988). An interview with Irving P. McPhail. Break Through: Programs for Educational Opportunity Newsletter, 15, 1-10.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62 (3), 229-258.

National Council for Teachers of Mathematics. (1989). Curriculum and evaluation standards for school mathematics. Reston: NCTM.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), Educational Measurement (3rd ed., pp. 447-72). New York: MacMillan.

Nitko, A. J. (1995). Is the curriculum a reasonable basis for assessment reform? Educational Measurement: Issues and Practice, 35, 5-10.

Norusis, M. J. (1991). The SPSS guide to data analysis for SPSS/PC+. (2nd ed.). Chicago, IL: SPSS Inc.

Nunnally, J. C. (1978). Psychometric Theory. (2nd ed.). New York: McGraw-Hill.

Peterson, P. L., & Knapp, N. F. (1993). Inventing and reinventing ideas: Constructivist teaching and learning in mathematics. In G. Cawelti (Ed.), Challenges and achievements of American Education. Alexandria: Association for Supervision and Curriculum Development.

Resnick, L. B. (1987). Education and learning to think. Washington, DC: National Academy Press.

Resnick, L. B., & Klopfer, L. E. (1989). Toward the thinking of curriculum: Current cognitive research. 1989 Association for Supervision and Curriculum Development Yearbook. Alexandria: Association for Supervision and Curriculum.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), Future assessments: Changing views of aptitude, achievement, and instruction. Boston: Kluwer.

Robitaille, D. F., Schroeder, T. L., & Nicol C. C. (1991). Mathematics '90: A Status Report on School Mathematics in British Columbia. Provincial Report. Victoria: British Columbia Department of Education.

Rothman, R. , New tests based on performance raise questions. Education Week. September, 12, 1990. 1, 10, 12.

Rudman, H. (1987). The future of testing is now. Educational Measurement: Issues and Practices, 6, 5-11.

Shade, B. J. (1982). Afro-American cognitive style: A variable in school success? Review of Educational Research, 52 (2), 219-244.

Shade, B. J. (1986). Is there an Afro-American Cognitive Style. Journal of Black Psychology, 13, 13-16.

Shavelson, R. J., Baxter, G. P. (1992). What we've learned about assessing hands-on science. Educational Leadership, 48, (8), 20-25.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991a). Performance assessment in science. Applied Measurement in Education, 4, 347-362.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991b). Performance assessment: Political rhetoric and measurement reality. Educational Researcher, 21, 22-27.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessment: political rhetoric and measurement reality. Educational Researcher, 21, 22-27.

Shepard, L. A. (1992). Evaluating test validity. Review of Research in Education, 19, 405-450.

Shiffler, N. L., Beyer, A. L., & Sperling, D. (1992). Primary Mathematics Performance Assessent at the District Level: There's No Substitute for Practical Experience. Paper presented for the Michigan Educational Research Association Winter Conference, Ann Arbor, MI. .

Silver, E. (1990). Assessment and Mathematics Education Reform in the United States. Paper presented at the annual meeting of the American Educational Research Association, Boston, Mass.

Slack, W. V., & Porter, D. (1980). The Scholastic Aptitude Test: A critical appraisal. Harvard Educational Review, 50, 392-410.

Stiggins, R. J., (1991). Facing the challenges of a new era of educational assessment. Applied Measurement in Education, 4, 263-267.

Webb, N. L., & Romberg, T. A. (1994). Reforming Mathematics Education in America's Cities. New York: Teachers College Press.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70, 703-713.

Wiggins, G. (1993). The immorality of test security. Educational Policy, 8 (2), 157-183.

Williams , P. L. , Phillips, G. W., Yen, W. M. (1991). Measurement Issues in High Stakes Performance Assessment. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Wittrock, M. C. (1991). Testing and recent research in cognition. In M. C. Wittrock & E. L. Baker (Eds)., Testing and Cognition (pp. 5-16). Englewood Cliffs, NJ: Prentice-Hall.

Wolf, D. W ., Bixby, J., Glenn III, J., & Gardner, H. (1990). To use their minds well: Investigating new forms of student assessment. Review of Research in Education, 17, 31-74.

INTERVIEWS AND MEETINGS

Linn, R. (Speaker), 1995, Michigan State University Assessment Symposium.

Pryor, O. (1989*a*, June) [Staff Meeting] Oak Hill School District.

Pryor, O. (1989*b*, July) [Interview with Classroom Assessment Specialist 1 and Research and Development staff person]. Oak Hill School District.

Pryor, O. (1989*c*, June) [Staff Meeting] Oak Hill School District.

Pryor, O. (1989*d*, October) (Interview with Irving P. McPhail)

Pryor, O. (1993, June) [Interview with Research and Development psychometrician]. Oak Hill School District.

Pryor, O. (1994*a*, February) [Interview with Research and Development psychometrician]. Oak Hill School District.

Pryor, O. (1994*b*, February) [Interview with Classroom Assessment Specialist 1 and Research and Development staff person]. Oak Hill School District.

Pryor, O. (1994*c*, February) [Interview with Classroom Assessment Specialist 2 and Research and Development staff person]. Oak Hill School District.

Pryor, O. (1994*d*, March) [Interview with Mathematics Coordinator]. Oak Hill School District.

Pryor, O. (1994*e*, March) [Interview with Teacher 1]. Oak Hill School District.

Pryor, O. (1994*f*, March) [Interview with Teacher 2]. Oak Hill School District.

Pryor, O. (1994*g*, April) [Interview with Teacher 3]. Oak Hill School District.

Pryor, O. (1994*h*, February) [Interview with the 1990-91 Research and Development Director of the Oak Hill School District].

Schiller, R. (Speaker). (1992). State of assessment in Michigan. Speech given at annual Michigan Educational Research Association Meeting, Novi, MI.

Schmidt, W. (1991). [Psychology class notes on test constructionClass Notes on Test Construction] Michigan State University. East Lansing, MI.

Stiggins, R. J. (1989) [Teacher Inservice at Oak Hill School].