AN EXAMINATION OF THE INSTRUCTIONAL VALIDITY OF THE ARIZONA INSTRUMENT TO MEASURE STANDARDS

by

Megan Eileen Welsh

__________________________
Copyright © Megan Eileen Welsh 2009

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2009
As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Megan Eileen Welsh entitled An Examination of the Instructional Validity of the Arizona Instrument to Measure Standards and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: 8/22/08

Jerome V. D’Agostino

Date: 8/22/08

Joel R. Levin

Date: 8/22/08

Darrell L. Sabers

Date: 8/22/08

Thomas L. Good

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.
I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 8/22/08

Dissertation Director: Jerome V. D’Agostino
STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the copyright holder.

SIGNED: __________________________
Megan Eileen Welsh
 ACKNOWLEDGEMENTS

The confines of this dissertation are too small to properly thank everyone who has contributed to my development. To those overlooked, please accept my heartfelt apology.

My mother has served as a great role model and mentor. She has taught me to challenge myself and to embrace life. My sister, Holly, has been a steadfast friend and a reminder of what’s truly important. I can only aspire to her generous spirit and patient countenance. Finally, my husband Dan has weathered many moods, edited innumerable papers, and supported me even while questioning my choices. His advice, willingness to let me chart my own path, and love have enriched all of my experiences. I am lucky to have found him.

The staff of Policy Studies Associates shepherded me through my early development as a researcher. Brenda Turnbull has been a wonderful mentor and a friend. She and Liz Reisner persuaded me to pursue graduate school. Rick White and Christina Russell are among my first collaborators and best teachers. I cherish having worked with them.

My committee has provided invaluable feedback. Dr. Joel Levin has provided encouragement when I needed it most and has shown me the benefits of learning from a great teacher and mentor. I am fortunate to join the ranks of students he has counseled. Dr. Darrell Sabers has pushed me to think. I have learned more from his tough questions than I have from many a journal article. Dr. Tom Good has provided gentle reminders about the importance of instruction, even in validity research.

I am eternally grateful to Shannon Smith for the hours she spent coding transcripts despite juggling many hats as a mother, school district employee and student. Her generosity can never be repaid. Caity Reddington and Jennifer McLaughlin both helped transcribe interview transcripts, an unpleasant and unrewarding task. Their efforts are greatly appreciated. Adrianna Cimetta and Cecilia Plaza were great teammates, advisors, and listeners. Adrianna provided infinite patience and support, despite working with a difficult colleague. Her graciousness was more than I deserved.

I am indebted to the teachers who agreed to participate in my research and the administrators who supported it. Without their willingness to give their time and thoughtful responses to interview questions, this study would not have happened.

My advisor, Dr. Jerome D’Agostino, has invested innumerable hours to my academic development. In addition to providing financial and intellectual support for the study, he has encouraged me to share my ideas, to work through problems, and has freely offered both accolades and critiques. I am a better researcher for having known him.
DEDICATION

To Mia – You are my inspiration.
# TABLE OF CONTENTS

LIST OF FIGURES .............................................................................................................. 7

LIST OF TABLES .................................................................................................................. 8

ABSTRACT ............................................................................................................................ 9

1. INTRODUCTION ............................................................................................................. 11
   The Test ........................................................................................................................... 12
   Instructional Sensitivity ................................................................................................. 14
   Research Questions ....................................................................................................... 26

2. REVIEW OF THE LITERATURE .................................................................................. 27
   Test Validity .................................................................................................................... 27
   The Evolution of Instructional Validity .......................................................................... 31
   Conclusion ..................................................................................................................... 52

3. METHODS ..................................................................................................................... 54
   Participants ..................................................................................................................... 54
   Materials ....................................................................................................................... 62
   Procedure ....................................................................................................................... 67

4. RESULTS ....................................................................................................................... 85
   Differences Between Grade Levels ............................................................................. 85
   Descriptive Results ...................................................................................................... 88
   Intercorrelations .......................................................................................................... 95
   Multilevel Analyses ..................................................................................................... 98

5. DISCUSSION .................................................................................................................. 113

APPENDIX A: INTERVIEW PROTOCOL ........................................................................... 123

APPENDIX B: TESTLETS REVIEWED BY TEACHERS ...................................................... 127

APPENDIX C: ALIGNMENT RUBRICS ............................................................................. 137

REFERENCES .................................................................................................................... 142
LIST OF FIGURES

Figure 1a. Scatterplot of an Instructionally Sensitive Test ........................................ 19
Figure 1b. Scatterplot of a Moderately Instructionally Sensitive Test ....................... 19
Figure 1c. Scatterplot of a Non-Instructionally Sensitive Test.................................. 19
Figure 2a. Scatterplot of an Instructionally Sensitive Test: Gifted Classrooms ........... 20
Figure 2b. Scatterplot of an Instructionally Sensitive Test: Special Education and Newcomer Classrooms ................................................................................. 20
Figure 3. Scoring Rubric for One AIMS Item ............................................................ 69
Figure 4. Classroom Assignment Example .................................................................. 71
Figure 5. Distribution of Classroom Alignment Ratings, by Grade ............................ 94
Figure 6. Distribution of Classroom Emphasis Ratings, by Grade ............................. 94
Figure 7a. 95% Confidence Interval around Mean Classroom AIMS, Grade 3 ............ 99
Figure 7b. 95% Confidence Interval around Mean Classroom AIMS, Grade 5 ..........  99
Figure 7c. 95% Confidence Interval around Mean Classroom Thirdest ...................... 99
Figure 7d. 95% Confidence Interval around Mean Classroom Fifthest ....................... 99
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1</td>
<td>Mean 2005 AIMS Scale Scores, Study Sample, Participating District and Arizona</td>
<td>55</td>
</tr>
<tr>
<td>Table 2</td>
<td>Demographic Characteristics of Study Sample, District and Arizona, 2004-05</td>
<td>56</td>
</tr>
<tr>
<td>Table 3</td>
<td>Standards-based Report Card-AIMS Hit Rates in Participating District, 2004</td>
<td>58</td>
</tr>
<tr>
<td>Table 4</td>
<td>Teacher Qualifications in Participating Classrooms</td>
<td>61</td>
</tr>
<tr>
<td>Table 5</td>
<td>Performance Objectives Discussed in Interviews</td>
<td>64</td>
</tr>
<tr>
<td>Table 6</td>
<td>Inter-rater Agreement on Overall Alignment Ratings</td>
<td>73</td>
</tr>
<tr>
<td>Table 7</td>
<td>Analysis Variables</td>
<td>76</td>
</tr>
<tr>
<td>Table 8</td>
<td>Descriptive Statistics for Variables Included in HLM Analyses</td>
<td>89</td>
</tr>
<tr>
<td>Table 9</td>
<td>Descriptive Statistics for Variables Included in HLM Analyses, by Grade</td>
<td>90</td>
</tr>
<tr>
<td>Table 10a</td>
<td>Zero-order Correlations Among Variables Used in Analysis</td>
<td>96</td>
</tr>
<tr>
<td>Table 10b</td>
<td>Zero-order Correlations Among Variables Used in Analysis (Continued)</td>
<td>96</td>
</tr>
<tr>
<td>Table 11</td>
<td>Variation in AIMS Scores Explained by the HLM Models, Third- and Fifth-grade Combined</td>
<td>102</td>
</tr>
<tr>
<td>Table 12</td>
<td>Variation in Third-grade AIMS Scores Explained by the HLM Models</td>
<td>102</td>
</tr>
<tr>
<td>Table 13</td>
<td>Variation in Fifth-grade AIMS Scores Explained by the HLM Models</td>
<td>103</td>
</tr>
<tr>
<td>Table 14</td>
<td>Variation in Thrdtest Scores Explained by the HLM Models</td>
<td>103</td>
</tr>
<tr>
<td>Table 15</td>
<td>Variation in Fifthtest Scores Explained by the HLM Models</td>
<td>104</td>
</tr>
<tr>
<td>Table 16</td>
<td>Final Conditional HLM 2-Level Model Predicting Variation in AIMS, Third- and Fifth-grade Combined</td>
<td>107</td>
</tr>
<tr>
<td>Table 17</td>
<td>Final Conditional HLM 2-Level Model Predicting Variation in Third Grade AIMS</td>
<td>107</td>
</tr>
<tr>
<td>Table 18</td>
<td>Final Conditional HLM 2-Level Model Predicting Variation in Fifth Grade AIMS</td>
<td>107</td>
</tr>
<tr>
<td>Table 19</td>
<td>Final Conditional HLM 2-Level Model Predicting Variation in Thrdtest</td>
<td>108</td>
</tr>
<tr>
<td>Table 20</td>
<td>Final Conditional HLM 2-Level Model Predicting Variation in Fifthtest</td>
<td>108</td>
</tr>
</tbody>
</table>
ABSTRACT

The dissertation describes a study of the instructional validity of the Arizona Instrument to Measure Standards (AIMS), a standards-based assessment. The study addresses the third- and fifth-grade mathematics portion of the 2005 AIMS test, focusing on two performance objectives per grade level. The study addresses the following questions: Can variation in students’ mathematics achievement on AIMS be explained by instructional validity measures, namely: (1) alignment between test items and instructional characteristics and (2) by the degree of teacher emphasis on the two skills of interest to the study? Does the relationship between these measures and AIMS performance differ across grade levels? What possible explanations exist to account for grade level differences? Is there a relationship between the instructional validity measures and performance on the objectives of interest to this study?

The dissertation discusses the evolution of thinking about instructional validity as standardized testing has changed. The study method, including developing alignment measures from interview transcripts and classroom assessment examples collected from 16 third-grade teachers and 20 fifth-grade teachers in one school district are also described. Findings of the study are presented. Although the method of using qualitative data to gauge instructional validity yields rich information about instructional practice, there was little instructional variation between classrooms in the district studied. This may have occurred because the district requires teachers to provide instruction exactly as specified in the district-adopted mathematics text. Some between-grade level differences
do exist in the curricular alignment with AIMS; however, teachers attempted to overcome this in their instruction despite district mandates to the contrary. Results support the instructional sensitivity of AIMS at third grade, but not at fifth grade. Differences in instructional sensitivity across grade levels might be linked to curricular differences; some third-grade teachers reported supplementing the curriculum to address the state standards while fifth-grade teachers largely reported that this was not necessary. Interestingly, the degree of alignment at third- and fifth-grade did not vary, although fifth-grade teachers placed more emphasis the study objectives. This speaks to third-grade teacher commitment to address the standards, and the challenges in emphasizing them when district-adopted curricula are not well-aligned with state standards. Only two third-grade teachers solely taught the curriculum. The remaining third-grade teachers either addressed the standards in a way that mimicked the test or presented concepts to students in multiple ways in an attempt to ensure that students could generalize their knowledge to a variety of situations.
Standards-based reform has had a strong influence on the American educational system over the past twenty years. In the mid-eighties a government-sponsored commission declared our nation to be at risk due to weaknesses in our educational system. The commission called on states to implement rigorous standards that require high expectations of all students (National Commission on Excellence in Education, 1983). Since that time, the federal government has adopted policies that require states to develop standards outlining what students should know and be able to do, to assess student progress, and to hold schools accountable for student performance (National Educational Goals Panel, 1991; Goals 2000: Educate America Act, 1994; No Child Left Behind Act, 2001). Despite massive federal support for standards-based reform, there has been some debate about its effectiveness in improving classroom instruction (Darling-Hammond & Wise, 1985; Elmore & Furhman, 2001; Spillane, 2004).

One criticism is that standards-based reform compels teachers to focus their instruction to the format of specific test items rather than instructing students in ways that help them learn to apply what they learn to a variety of situations (Smith, 1991). Others have contended that it compels teachers to change their instruction in ways that are beneficial to students. They theorize that, by outlining what students should know and be able to do at each grade level, testing students on those skills, and holding schools accountable for performance on standards-based tests, teachers will be persuaded to
improve their instruction (Ravitch, 1996).

This dissertation study examines the mathematics teaching practices of 16 third-grade and 20 fifth-grade teachers in one school district that emphasizes standards-based instruction. It determines how teachers responded to working in this environment; whether they “taught to the test” or addressed skills more generally in helping students to meet state standards. It uses this information to explore whether one of these approaches was associated with better than expected scores in mathematics on Arizona’s state assessment, the AIMS test.

The Test

Arizona’s standards-based assessment was first administered in 1999 to provide feedback on student progress towards state standards. Currently administered in grades 3 through 8 and in grade 10, the test assesses student content knowledge in reading, writing, and mathematics. The reading and mathematics tests are comprised of between 76 and 84 multiple choice items, some of which are used to generate criterion-referenced scores (proficiency levels, which label students as “Falls Far Below,” “Approaches,” “Meets,” or “Exceeds” state standards) and some of which are used to generate norm-referenced scores (percentiles that rank student performance in relation to national norms).¹ Most criterion-referenced items were generated by Arizona teachers hired to participate in test development while norm-referenced items were created by the AIMS vendor. In addition to subject-level scores, subtest (or testlet) scores are generated which

¹ At tenth grade, AIMS items are used only to generate criterion-referenced scores.
reflect the organization of the state standards. The standards are arranged in three

hierarchical levels, the broadest level, called “strands” in Arizona, represents a wide array of skills. Strands are divided into more specific concepts (level 2), which in turn contain the most specific descriptions of what students are supposed to know and be able to do (performance objectives). At third grade, there are 84 mathematics objectives, 17 concepts, and five strands. Although some objectives are not assessed by the AIMS test, test items address 16 of the 17 concepts and all five strands. In addition, 16 testlet scores are generated, for all five strands and for 11 concepts or combinations of concepts. While some testlet scores are based on a relatively large number of items (25 number sense items at third grade in 2005) others are based on relatively few items (four items each comprise six testlet scores at third grade in 2005—structure and logic, algorithms, data analysis, probability, discrete math, and estimation).

The test has evolved over time, with changes in item format, grades assessed, and scores generated. Originally, the test was administered only in grades 3, 5, 8, and 10 and was comprised of both multiple-choice and constructed response items. In addition, the test has become more high-stakes over time. Although high school graduation was supposed to be linked to AIMS performance as early as 1999, it was first used to make graduation determinations in Spring 2006. The test is also used to meet the requirements of the No Child Left Behind Act of 2001 (NCLB), to influence merit pay decisions in some districts, and to determine state accountability ratings. Because of its strong association with state and federal accountability systems, teachers are keenly aware that

2 While the format of reading and mathematics items has changed, the writing assessment has always been an essay test.
the test is used to evaluate their instruction. Therefore, they may be motivated to provide instruction on test items instead of developing conceptual understanding of subject matter. However, such behavior would jeopardize the validity of test score (and accountability rating) interpretations if AIMS lacks instructional sensitivity.

Instructional Sensitivity

Standards-based reform, as outlined by O’Day and Smith (1993), represents a systematic approach to ensuring that all students gain proficiency on a uniform set of skills at each grade level. By establishing the order with which topics are presented across grade levels, states hope to ensure that students are exposed to a meaningful progression of skills, that all students are exposed to challenging material, and that students will experience continuity in their education if they transfer schools. The system that O’Day and Smith (1993) describe includes several components, most designed to compel teachers to implement the reform. These include: establishing academic standards that describe what students should know and be able to do, gauging student progress towards attaining the standards, and holding schools accountable for student performance. They also recommended professional development for teachers to help them develop the skills necessary to implement the reform.

Although standards-based reform is designed to ensure that students learn a uniform set of skills, O’Day and Smith (1993) also specify that teachers know best how to convey information to students and leave specific instructional strategies up to them. This provides teachers with a great deal of latitude, both in the method of instruction used
and in deciding which aspects of the standards are most important. Such instructional leeway can have a variety of consequences. While some teachers may erroneously interpret standards documents, others may construe the standards in a multitude of valid ways. If teachers have different yet plausible interpretations of what the standards mean, their instructional forms likely will vary (Baker & Herman, 1983), resulting in differential experiences for students across classrooms.

This phenomenon is well documented. In their study of nine teachers implementing language arts reforms, Spillane and Jennings (1997) noted that, even when professional development and other sources of support are in place, teachers’ instruction varies widely based on their knowledge, skills, and experience. In addition, using observations of district curriculum committees working to align state standards and district curriculum materials, Hill (2001) concluded that teachers interpreted the same objective quite differently and have difficulty coming to consensus about what is intended by standards documents.

The challenge for a state testing system, therefore, is to determine whether students have mastered the requisite objectives regardless of the examples or approaches used. They assume that instructional content will generalize to test items. That is, as long as teachers appropriately transmit the skill, students should do well on the assessment. Instructional sensitivity studies examine whether this assumption holds. With regard to standards-based assessments, they gauge the ability of tests to detect a wide array of standards-based instruction.

It seems self-evident that instructionally sensitive assessments are needed in high-
stakes testing situations. Assessment is the lynchpin of standards-based reform as it connects standards implementation with accountability measures. For the reform to work, the assessments must detect adequate instruction on the standards. Otherwise, teachers who faithfully implement standards-based instruction may experience poor test scores and change instructional practices in search of ones that can be registered by the test. In this case, test scores easily influenced by test preparation practices would motivate teachers to show students how to answer specific types of items in lieu of developing the conceptual understanding necessary to apply knowledge in a variety of ways.

The literature is rife with examples of instruction focused on test items rather than course content, indicating that teachers have little faith in the instructional sensitivity of standardized tests (Stecher, Barron, Chun, & Ross, 2000; Madaus & Clarke, 2001; Schorr, Bulgar, Stickel Razze, Monfils, & Firestone, 2004). For example, Darling-Hammond and Wise (1985) conducted interviews with 43 teachers working in a variety of school districts and reported that teachers narrowed their curriculum to address only the tested subjects (e.g., reducing instruction in science and social studies) and taught “the precise content appearing on the test rather than the concepts underlying the content” and “skills as they are to be tested rather than as they are to be used in the real world” (p. 320). Taylor, Shepard, Kinner, and Rosenthal (2003) found similar patterns based on a survey of 1000 teachers from across Colorado. In addition, they documented that teachers in low-performing schools were more likely to place great emphasis on test preparation practices than teachers in high-performing schools. Teachers in low-performing schools also more frequently reported that they had reduced the instructional time devoted to
projects that extend over several days and devoted to conducting research in response to Colorado’s high-stakes testing program.

Based on this research, it seems that standards-based reform may have resulted in weakened instructional programs even though this clearly was not the goal. Therefore, it is necessary to determine if these assessments indeed reward teachers for providing instruction that mirrors test items. The potential for test scores to be biased by the use of specific instructional examples is not only a policy or a fairness concern. It also brings into question the validity of the test, or the correctness of test score inferences. As Messick (1989) has argued, test consequences are manifestations of test inferences and therefore must be considered a form of validity evidence. However, negative consequences do not necessarily indicate that a test is invalid. Rather, we must determine whether the consequence stems from sources of invalidity such as construct under-representation or construct-irrelevant variance (Messick, 1995). Put into the context of this study, documenting that teachers mirror test items in their instruction does not constitute evidence that a test is invalid. Validity is threatened only if doing so results in improved test scores.

The results of an instructionally sensitive assessment, therefore, would not be influenced by such approaches. Students in classrooms that addressed tested content would perform well on the assessment, regardless of the similarity between instructional examples and test items. To help elucidate this point, Figures 1a-c present the relationships between mean classroom performance on a standards-based assessment for a hypothetical instructionally sensitive test, a test of modest instructional sensitivity, and
a non-instructionally sensitive test. With the instructionally sensitive test (Figure 1a), classrooms that address the skill outperform classrooms that are not exposed to tested content. In addition, classrooms that learn instructional examples that mimic test items and those that learned the skill with examples dissimilar to the test perform similarly. In Figure 1b, the modestly sensitive test, there is a linear relationship between being exposed to the skill and test performance, with classrooms that are not exposed to the skill performing worst, those that learn the skill in a manner different from the way it is presented on the test performing moderately, and those taught with examples that mirror test items performing the best. In Figure 1c, the non-instructionally sensitive test, alignment between instruction and test content is not related to test performance.

These examples should apply to classrooms where grade level instruction is appropriate for the majority of students. In other classrooms, where grade level instruction is not appropriate, the relationship between alignment and performance would look different for tests that are instructionally sensitive in most classrooms (Figures 2a-b). Figure 2a presents one specialized category of classrooms. Teachers in gifted classrooms would not teach test content as it would be inappropriate for their students (who presumably enter the classroom surpassing grade level standards), but their students would still perform well on the test. Other teachers, such as instructors of special education or newcomer classrooms that help transition recent, non-English speaking immigrants into school, might not teach the standards because their students are not yet ready for them. These classrooms would be expected to perform quite poorly on an instructionally sensitive standards-based assessment, but for reasons having little to do
Figures 1a-c show the relationship between test scores and the instruction-test alignment for three tests. The first test (Figure 1a) is instructionally sensitive because students who receive instruction in the tested skill, regardless of the similarity between instructional examples and test items, do better than students who are not instructed in the test’s content. In Figure 1b (Modest Instructional Sensitivity), students who learn with examples that mirror the test outperform other students, although students who do not receive instruction in tested content perform the worst. The non-instructionally sensitive test (Figure 1c) does not detect any instruction on tested content, so there is no relationship between instruction-assessment alignment and test scores.
Figures 2a-b shows the relationship between test scores and the instruction-test alignment for an instructionally sensitive test administered to classrooms filled with special populations. Gifted classrooms are represented in Figure 2a. Teachers in these classrooms may not teach state standards because their students have exceeded grade level expectations, but their students would still do very well on the test. Classrooms of non-English speakers (Newcomers) and of special education students may also not teach state standards because it is inappropriate. When these classrooms participate in standards-based assessments, students would perform poorly on the test for reasons that extend beyond instructional alignment (Figure 2b).
with alignment (Figure 2b). Many states have provisions allowing these students to opt out of testing altogether. Some special education students also do not participate in standards-based assessments, while others may take an alternate assessment that is based on a separate set of standards. Therefore, few classrooms may fall into this category.

Despite the importance of instructional sensitivity evidence in determining the validity of standards-based assessments, little work has been done to determine how vulnerable tests are to teaching that mirrors test items. While the United States Department of Education requires states to provide validity evidence in the form of content alignment studies (U.S. Department of Education, 2004), it does not require states to evaluate the instructional validity of their assessments. It assumes that the skills students are exposed to in class will generalize to test items, regardless of the instructional form.

Instructional validity is also not mentioned in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). This might seem to indicate that documenting instructional sensitivity is not an important validation process. However, sensitivity is a core requirement if one is to make proper score inferences. As Airasian and Madaus (1983) have argued, test sensitivity represents one form of construct validation in that standards-based assessment scores should be influenced by instruction on the standards. Applying Cronbach and Meehl’s (1955) group comparison method, construct validity can be documented by showing that students who are instructed on the standards (the “treatment” group) perform better than students who
do not receive standards-based instruction (the “comparison” group) on a standards-based assessment.

This (and other) approaches have also been suggested by Popham (2007) and implemented by Niemi, Wang, Steinberg, Baker, and Wang (2007), who evaluated the instructional sensitivity of a school district performance based assessment by comparing scores on the assessment for students whose teachers were randomly assigned to one of three instructional emphasis groups for an eight-day unit: writing, literary emphasis, and teacher choice of emphasis. They found that the students exposed to the literary emphasis unit outperformed other students in their ability to describe conflict in a reading selection and that students in the writing group wrote stronger introduction and thesis statements to organize their writing. In addition, students in the literary analysis group outperformed other students in terms of the overall score on the assessment, which focused on writing an essay that analyzed a piece of literature.

This study attempts to evaluate the instructional validity of the Arizona Instrument to Measure Standards (AIMS). To do so, it addresses the operational match between how subject matter is conveyed on AIMS and in classrooms and the degree of emphasis placed on key test objectives during instruction. These measures are examined in relation to student performance on AIMS (adjusted for student background and for classroom characteristics) to determine whether teachers whose instruction mimics AIMS items and who emphasize tested content have higher-performing classrooms than other teachers.

This approach extends upon previous work in instructional validity, which has
focused either on the content of instruction or on pedagogies used. For example, Porter and Smithson (2001) developed a questionnaire to examine the match between the content of instruction and the content of assessments, called Surveys of Enacted Curriculum (SEC). To implement the SEC, teachers provide detailed information on the amount of time devoted to different topics at various levels of cognitive demand. State standards documents and assessments are also reviewed and scored according to the relative emphasis placed on each content area and level of cognitive demand. Alignment measures compare teachers’ responses with the assessments ratings. SEC analyses do not evaluate how teachers convey course content, nor do they ascertain how teachers interpret what is meant by the skills or content areas included in the questionnaires. Because they do not take an in-depth look at how the standards are operationalized, they are able to address all of the objectives comprised in standards documents.

Yoon and Resnick (1998) took a different approach to examining the instructional validity of a performance-based assessment. They evaluated whether middle school teachers implemented instructional strategies consistent with those required by a performance-based assessment without ascertaining the content of instruction. They asked middle school teachers who were and were not implementing the California Mathematics Renaissance program to complete a survey of instructional strategies aligned with the program, such as having students work in groups, conduct lab and field work, do oral presentations, and complete portfolios. Using hierarchical linear modeling and controlling for student socioeconomic status, they found that students whose teachers reported frequent use of methods included on the performance-based New Standards
Mathematics Reference Exam outperformed students in other classrooms.

There are several limitations to the approaches used in these studies. First, because they focus on either the content of instruction or the methods used, they cannot fully capture the complexities of classroom instruction and therefore suffer from construct under-representation in that they do not fully address standards-based instruction (Messick, 1995). In their discussion of the difficulties of fully portraying the alignment between instruction and assessment, Baker and Herman (1983) assert that the complexity of instruction makes it difficult for most studies to fully depict instructional processes. This study does not purport to fully gauge the complexities of classroom instruction. Rather, it represents one attempt to expand the types of instructional variation accounted for by considering both the content of instruction and the methods used to convey information to students. It does so by taking an in-depth look at the ways that teachers address specific objectives rather than cataloguing the variety of skills taught.

In addition to better capturing the ways that information is conveyed to students, this study also extends beyond prior work by determining how teachers interpret state standards. Past studies assumed that teachers uniformly decode what is intended by state standards by simply asking them whether they teach a particular skill and aggregating responses across classrooms. However, teachers may have very different concepts in mind when reading standards documents. Requiring teachers to explain how they address the standards can help elucidate what they believe is required of them.

This dissertation builds upon previous comparisons of how teachers and test items operationalize state standards (see D’Agostino, Welsh, & Corson (2007) for a complete
description). That study based its findings on survey data. The questionnaire asked teachers to draw examples of how they presented two mathematics objectives to students and to describe the key components of each skill, in addition to asking teachers to indicate how much emphasis they placed on a range of mathematics objectives. Two subject matter experts then examined the state test items relating to those objectives and coded the alignment between teacher responses and test items on a three-point rubric. Alignment ratings and the interaction of alignment with the degree of emphasis teachers placed on the skills significantly predicted test performance after controlling for student background characteristics. These results represent what this study describes as a modest degree of instructional sensitivity. That is, students who were exposed to the skill outperformed those who were not exposed to the skill, and those who were exposed to examples that mirrored test items outperformed those who learned the skill in a way that differed from the way it was presented on the state test.

Although that study addressed the way that objectives were operationalized in classrooms and on the state test, the descriptions generated by teachers may represent only one of many approaches implemented. Teachers were asked to provide one or two examples that they typically would present to students during instruction. However, the array of typical instruction could be so broad as to be not adequately captured by a limited number of examples. The present study attempts to expand on D’Agostino et al. (2007) by ascertaining the full range of instructional examples used in conveying two mathematics objectives.
Research Questions

The present study collected rich instructional descriptions of how third and fifth grade teachers in one school district taught two mathematics objectives. These descriptions were based on 90-minute interviews and also on examples of classroom assessments, collected and discussed during the interviews. More specifically, teachers were asked to describe each lesson relating to the performance objectives and to describe the progression of lessons over the course of the school year. They were also asked to discuss all of the ways that they assessed student understanding of these objectives and to explain what types of student responses constituted full understanding. The number of performance objectives examined was limited in an attempt to address the full array of examples and instructional approaches implemented.

These data were analyzed qualitatively and quantitatively to address several research questions. First, the study explores whether variation in students’ math scores is explained by the alignment between test items and instructional characteristics and by teacher emphasis on the two skills discussed (which are also predominant on the state math assessment), after accounting for student background and classroom characteristics. Second, the study determines whether the relationship between instructional validity measures and AIMS performance differs across grade levels, and explores possible explanations for grade-level variations. Finally, to determine whether the instructional validity measures (alignment and emphasis) predict student performance on the objectives of interest (in contrast with overall math scores), analyses are repeated using objective-specific testlet scores.
CHAPTER 2
REVIEW OF THE LITERATURE

This study attempts to gauge the instructional validity of a standards-based assessment by predicting scores with measures of instructional alignment, emphasis on tested topics, and use of instructional approaches that foster conceptual understanding of mathematical content. While there is an extensive body of literature devoted to ascertaining the instructional validity of minimum competency tests, criterion-referenced tests, and performance-based assessments, little work has been done to define instructional validity of standards-based assessments. To do so, this review discusses the role of instructional validity in the context of test validity and examines how instructional validity has evolved to reflect a range of assessments. Finally, opportunity to learn (OTL) will be addressed. OTL is an important concept within the framework of standards-based reform as it pertains to the resources that ensure students learn. It is especially relevant to the validity of standards-based assessments as certain aspects of OTL must be accounted for to accurately gauge instructional validity.

Test Validity

Test validity is discussed extensively in the *Standards for Educational and Psychological Testing* (1999) (*Standards*), developed as a joint effort between the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. First published in 1954, the
Standards have been revised multiple times, with the most recent version written in 1999. Although recommended approaches to studying validity have evolved over time, the definition of validity has remained fairly consistent. “Validity information indicates to the test user the degree to which the test is capable of achieving certain aims. Tests are used for several types of judgment, and for each type of judgment, a somewhat different type of validity is involved” (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1954, p. 13). That is, a test score itself is never valid or invalid, rather the validity of test score inferences are examined.

The 1954 Standards described four aspects of validity: content validity, predictive validity, concurrent validity and construct validity. Content validity is appraised by determining how well a test samples from across the domain of skills of concern, focusing on the breadth and depth of coverage. Predictive validity is examined by gauging how well a test predicts performance on a separate criterion measured at some point in the future. Therefore, predictive validity studies involve administering a test, waiting to see how test takers perform on a future criterion, and determining whether the test was an adequate predictor of the criterion score. Concurrent validity is similar to predictive validity, except that performance can be gauged simultaneously with the assessment. For this reason, predictive validity and concurrent validity are often grouped together under the term “criterion-related validity” (Allen & Yen, 1979). Finally, construct validity is evaluated by studying the degree to which test performance can be explained by the characteristics it was intended to measure. Cronbach and Meehl (1955)
recommended that construct validity be addressed by first outlining a nomological network defining relationships between observable properties, between observable properties and theoretical constructs, and among constructs. Construct validity studies then test whether the relationships between test scores and other observed variables behaved as predicted. Campbell and Fiske (1959) specified an approach for testing nomological networks by examining convergent and discriminant validity evidence. That is, construct validity studies should both document that interrelated variables correlate (convergent evidence) and that theoretically unrelated variables do not correlate (divergent evidence).

Recently, validity theorists have debated the appropriateness of this trinitarian view of validity, which treats content, criterion-referenced, and construct validity as distinct (Guion, 1980). Messick (1989) argued that validity is a unitary concept in that all validity evaluations address whether a test measures the concepts or skills of interest and therefore make judgments about construct validity. Criterion-related and content validity approaches remain, but are considered aspects of construct validity.

The 1999 version of the Standards (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999) adopts this unitary approach and outlines various sources of evidence that can be used in evaluating the appropriateness of test score interpretations. These include: evaluating the content of the test in relation to the construct it is intended to measure, evaluating the response processes test takers use, examining internal structural evidence to determine if the relationships among test items are consistent with those intended by
the test user, and evidence based on the relationship of test scores to other variables (such as criterion-related measures, analyses of convergent/discriminant variables, etc.).

Instructional validity studies ascertain whether test scores detect instruction on the relevant skills, regardless of the instructional examples used. Although certain approaches may be more effective than others, and should therefore result in higher test scores, teachers may also be uniform in their instructional effectiveness, but emphasize different aspects of tested content. These variations in instructional form, which Baker and Herman (1983) define as “the substantive features of the task, other than topic, over which the learner is expected to generalize” (p. 155), may be more or less similar to those presented on an assessment. As long as teachers have effectively presented a concept, student understanding should generalize to test performance. When test scores vary as a result of instructional form, inferences about the quality of education provided to students are jeopardized.

Instructional validity is not addressed in either the 1954 or the 1999 Standards. However, under the unified view of validity, instructional validity is an aspect of construct validity and calls to investigate the ability of tests to detect instructional differences date back at least thirty years. Airasian and Madaus (1983) argued, “If we want to make inferences about differential school, program, or instructional effectiveness, then the processes underlying performance on achievement measures need to be closely linked to instruction. Evaluating these links is a construct validity issue, one that has been largely ignored by those who build standardized achievement tests” (p. 106). Efforts to examine the instructional validity of tests are discussed below.
The Evolution of Instructional Validity

Minimum competency tests

Instructional validity gained prominence with the Debra P. v. Turlington case, brought against the State of Florida to challenge its use of a minimum competency test to make high school graduation decisions. Of particular concern was the disparate performance of minority and non-minority students on the test. Plaintiffs claimed that students of color performed poorly because they received inferior instruction, which did not address tested concepts. Therefore, the plaintiffs argued that students should not be held accountable for passing (644 F.2d 397 [1981]).

The court agreed that it was unfair to require students to pass a test of concepts absent from instruction. In addition, they found that, if the tested concepts truly were not taught in Florida’s classrooms, then the test was invalid. It required the State to prove that schools addressed tested concepts. The resulting study included several data sources: a survey of all Florida’s teachers that asked which objectives were taught, school district reports detailing the curriculum and instruction implemented, classroom observations, and surveys of eleventh-graders asking about the concepts and skills learned (Langenfeld & Crocker, 1994). Based on this information, the court concluded that students had the opportunity to learn tested objectives and that test validity was documented (730 F.2d 1405 [1984]).

In his writings about the case, McClung (1979) argued that two types of validity were pertinent: curricular validity and instructional validity. He described curricular
validity as the match between skills and knowledge addressed on a test and in curricular materials, while he described instructional validity as “a measure of whether schools are providing students with instruction in the knowledge and skills measured by the test” (McClung, 1979, p. 683). The Debra P. case shaped ensuing measurement studies in many ways. First, states examined whether students had the opportunity to learn the skills contained in minimum competency tests. Second, researchers developed new methods for gauging instructional and curricular validity. Finally, new tests were developed to measure student progress towards specific criterion instead of determining competence on more general skills.

Masters (1988) provides an example of a state effort to validate a minimum competency test. In this two-pronged study, he first used questionnaire data collected from third-, fifth-, and eighth-grade teachers and from principals across the state of Pennsylvania to determine whether teachers addressed tested objectives in their instruction. After determining that the majority of teachers did indeed cover the material included on the state assessment, he conducted Differential Item Functioning (DIF) analyses to compare the item \( p \)-values for students whose teachers did and did not provide instruction in the concepts of interest. Since DIF analyses were only conducted for selected objectives, it is difficult to draw conclusions about the instructional sensitivity of the overall test. However, he did find some \( p \)-value differences, with students who had been taught a skill outperforming those who had not been exposed to the skill, and also found more relationships for the mathematics objectives than for the reading objectives that were selected.
Masters (1988) interpreted these results as providing evidence of the usefulness of DIF analyses in reviewing test items, but did not argue for or against the instructional sensitivity of the test. Rather, he concluded that Pennsylvania’s assessment was instructionally sensitive because it tested the taught skills. As with the Debra P. studies, Masters’ (1988) research did not evaluate test items or test results to make conclusions about instructional validity. For minimum competency tests, instructional validity judgments seem to have focused on instructional content.

After the Debra P. decision, researchers developed methods for examining the curricular and instructional sensitivity of tests to extend beyond the state efforts described above. Mehrens and Phillips (1987) conducted a curricular validity study in one school district using schools matched on student demographic characteristics and test performance. Although the schools served similar student populations, they used one of three different texts, allowing cross-text comparisons. The content of the texts and of the Stanford Achievement Test were analyzed by first assigning each test item (and problems in the texts) to one of 180 categories that captured the cognitive process (concepts, skills, or applications), operation (addition, subtraction, etc), and nature of the materials (fractions, decimals, etc.). Then the number of items per category was computed for the test and for each text, the degree of text-test alignment summarized, and item \( p \)-values aggregated across all items that focused on a specific cognitive process for each curricular group. The \( p \)-value differences across texts were small, indicating that text use was not associated with differing levels of test performance. The authors also supported this conclusion by correlating \( p \)-value differences with disparities in cross-text content.
Since these two variables were not correlated, the authors surmised that variations in test performance could not be attributed to differences among the texts. Finally, they conducted three separate Rasch analyses (one for each text group) and found that item level logit scores were similar for all three groups (Mehrens & Phillips, 1987).

This methodology stands in stark contrast to Nitko’s (1995) discussion of curriculum-driven assessment, which argues that instructionally sensitive tests must be based on the curriculum. That is, instead of comparing curriculum content with an established test that is already in use, assessments should be based on the curriculum already in place. Therefore, different tests would be administered for each text.

Another approach involved asking teachers to review test items and report whether they believed students could correctly answer each item. Leinhardt and Seewald (1981) asked approximately 400 first- and third-grade teachers to review each item on the mathematics and reading portions of the Comprehensive Test of Basic Skills (CTBS) and, for each item, to estimate the percentage of students who had been taught enough of the content assessed to provide a correct answer. In addition, the CTBS was administered as both a pretest and posttest to students in each classroom. The researchers summed teacher estimates across items to derive an overall estimate of test-instruction overlap. The overlap measure was used in classroom-level regression analyses, which predicted student performance on the CTBS posttest while controlling for pretest student achievement levels. The authors found that overlap significantly predicted student performance at both grade levels in both reading and math, indicating that the test was instructionally sensitive (Leinhardt & Seewald, 1981).
Criterion-referenced tests

Minimum competency tests were originally developed with the assumption that the assessments measured progress on a broad domain of skills that students should develop regardless of instructional emphases. However, concerns about a mismatch between the skills tested and the skills taught buoyed implementation of criterion referenced tests, which provide information about student performance on a specific set of skills, presumably those addressed in instruction (Popham, 1978). Since criterion referenced tests narrowed the domain (from general mathematics understanding to a specific set of math skills), instructional validity was reconceptualized. While minimum competency test validation efforts required proof that classroom instruction addressed tested skills, validation for criterion-referenced tests began during test development and focused on selecting those items most sensitive to instruction. It was assumed that, by including instructionally sensitive items, the test score would be instructionally valid.

The most common approach to gauging the instructional sensitivity of test items involved administering the same set of items to students before and after instruction, or to equivalent groups of students who had or had not received instruction. Those items whose characteristics changed between the pretest and the posttest were assumed to have changed as a result of instruction and were therefore considered to be instructionally sensitive (Hambleton, Swaminathan, Algina, & Coulson, 1978). Hence, for criterion-referenced tests, instructional sensitivity studies have based validation findings on characteristics of test items instead of on schools or instruction.

Many statistical indices were created to capture item sensitivity, defined by
Haladyna and Roid (1981) as “the tendency for an item to vary in difficulty as a function of instruction” (p. 40). Based on Cronbach and Meehl’s (1955) group comparison method, instructional sensitivity analyses compared the test performance of two groups: those exposed to a treatment that increases the attribute of interest and a control group. According to the group comparison method, if the treatment group outperforms the control group, construct validity is established. Item sensitivity analyses were used to identify test items that, in combination, create an instructionally sensitive test and, by extension, a test of high construct validity.

Item sensitivity measures were developed using a variety of measurement theories. The most straightforward measures, based in classical test theory, included a simple comparisons of item $p$-values (developed by Cox & Vargas, 1966), a phi-coefficient analysis that correlates pretest and posttest responses (Popham, 1971), and a review of point-biserial correlations between item responses and total test scores for tests administered to students with varying levels of instruction (Haladyna, 1974). In addition, a $t$ test of item response theory (IRT) item calibration differences (Wright & Stone, 1979), and Bayesian methods were also used (see Haladyna & Roid, 1981, for a review of these methods).

Finally, Hanson, McMorris, and Bailey (1986) proposed an approach based on generalizability theory. They categorized teachers as “low,” “middle,” or “high” implementers based on teacher estimates of the number of lessons devoted to an instructional program over the course of the school year. Total scores were then generated using every possible combination of items (e.g., creating one item tests with
each item, creating two-item tests of each possible combination of items, creating three item tests, etc.). The variance in each score was analyzed to estimate the percentage of variance explained by three facets—student, classroom, and implementation. Results were used to identify the combination of items that maximized the implementation variance component, which would then comprise the assessment. The authors argued that, by applying this method, instructional validity would be optimized.

Interestingly, although these methods assume that instruction in a skill will increase proficiency with that skill, instructional sensitivity analyses did not gather information on the instructional methods used. Not only were pedagogical approaches ignored, but information on the content of instruction also was not collected. Furthermore, the studies did not directly address the validity of test scores since analyses were conducted during item selection. Rather, the research assumed that selecting instructionally sensitive items ensured the validity of the total test scores.

Performance-based assessments

As criterion-referenced tests gained prominence, researchers documented and became concerned about teacher efforts to “teach to the test,” or mimic test items in their instruction. This can take many forms. One concern is that teachers respond to standardized assessments by narrowing the curriculum to only teach content measured on the test (Darling-Hammond & Wise, 1985). In addition, some have found classrooms in which instruction solely mirrors the format of standardized test items and does not challenge students to apply concepts in a variety of ways (Smith, 1991). Teachers also
attempt to increase test scores by instructing students in test taking skills instead of focusing on content. This continuum of test preparation activities can range from no special preparation for testing to teaching test taking skills, to providing students with the correct answers to test items (Mehrens & Kaminski, 1989). Unfortunately, this pattern of teaching test-taking strategies in lieu of subject matter is most prevalent among teachers who serve minority and low-socioeconomic status student populations, and among secondary teachers who work in states, like Arizona, with high school exit exams (Dorr-Bremme & Herman, 1986; Madaus & Clarke, 2001).

Performance-based assessments gained prominence, at least in part, due to this perception that instruction focused on correctly answering test items instead of nurturing thinking skills or conveying important content. To motivate teachers to improve their instruction, performance-based assessment tasks required students to exhibit their facility with problems that mirrored promising instructional approaches (Linn, 1993). Therefore, they required students to participate in activities that reformers hoped would occur in classrooms. Students were scored on their performance during the activity or on the product created. The rationale for these assessments was straightforward. If teachers were going to mirror assessments in their instruction, then test developers should create tests that require teachers to use sound pedagogy (Frederiksen & Collins, 1989). Because of this change in test format, standard validation techniques were reexamined and methods for capturing the appropriateness of performance-based assessment score inferences proposed (see Moss, 1992 for a review of such approaches).

Frederickson and Collins (1989) presented principles for the systematic design of
valid tests, which outlined key considerations for selecting assessment tasks, determining
the primary traits each task is designed to measure, developing a library of exemplar
responses, and providing a training system for scoring. In addition, they outlined a set of
validity standards for performance-based assessments: directness (the skill of interest is
measured directly because it is performed on the assessment task), scope (the
performance task contains all aspects of a skill), reliability (consistency of scores across
students with the same level of understanding), and transparency (test takers understand
the criteria by which they are being judged). Finally, they listed acceptable methods for
improving test scores, which were quite different from those typically considered
acceptable for multiple-choice tests (e.g., including opportunities to practice taking the
test and allowing students to take the assessment multiple times).

Haertel (1990) also suggested an approach to designing performance assessments
to ensure validity. The strategy he suggested reflected many principles of good
assessment planning, regardless of test format, including: determining the purpose of
assessment, defining the aspects of a skill to be evaluated, selecting tasks or problems to
gauge that skill, outlining the performance task parameters (e.g., will students have
calculators?, will they work in groups or individually?), writing scoring criteria, selecting
scorers, and training them.

Like instructional sensitivity measures developed for criterion-referenced tests,
the measures mentioned above were used during test development to maximize the
validity of assessments. However, methods for evaluating the validity of performance-
based assessments after they were administered were also developed. Concerned that
traditional validity approaches would result in unfavorable reviews of performance assessments, Linn, Baker, and Dunbar (1991) also proposed an expanded set of validity criteria to emphasize the kinds of information that performance assessments provide. They argued that performance assessments provide a unique window into student learning in that they help gauge students ability to apply their learning to new situations, to solve complex problems, and to work in ways that are similar to what will be required of them in life outside of school. The authors developed validity criteria to better capture these roles of performance assessments. These included: 1) the consequences of testing, 2) fairness (equitable access to resources, rater bias, and construct irrelevant variance), 3) generalizability (across raters and tasks), 4) evaluations of the quality of performance tasks (the cognitive complexity required, quality of content, content coverage across the domain, and meaningfulness of the task to students), and 5) cost and efficiency of testing.

In that same year, Haertel (1991) also discussed the validity criteria he considered most relevant to evaluating performance assessments in the context of performance-based teacher certification tests. More in line with traditional validity approaches, he identified the most important aspects of validity for performance based tests instead of developing new terms, narrowing the concept of validity for these assessments. He was concerned with issues most pertinent to these new tests, namely: replicability and generalizability (across raters and tasks, or test-retest and parallel forms reliability), criterion-related validity, test bias and establishing defensible standards for test performance. He did not address other aspects of validity, which might be more difficult to gauge with performance assessments, such as examining the internal structure of tests (Cronbach &
Meehl, 1955), or reviewing convergent and discriminant evidence by applying multitrait-multimethod approaches (Campbell & Fiske, 1959).

Methods for examining instructional validity also changed for performance-based assessments. Since criterion-referenced tests focused on course content, instructional sensitivity measures were largely unconcerned with pedagogy. In contrast, performance-based assessments focus on the method of instruction, requiring that instructional validity efforts examine classroom activities. That is, instructional validity of performance-based assessments addresses whether the similarity between classroom activities and performance tasks affects test scores. This is similar to instructional validity of minimum competency tests, which gathered data on the concepts and skills emphasized in instruction. However, validation of performance assessments focuses on the instructional approaches used (cooperative grouping, use of manipulatives, etc.) instead of the content of instruction.

For example, Yoon and Resnick (1998) provide an example of attempts to study the instructional validity of performance-based assessments. They correlated use of classroom activities similar to those encountered on one performance-based assessment with test outcomes, and argued that the positive correlation between these two variables indicates that the test is instructionally valid (see the introduction for a detailed discussion of this study).

Lane, Park, and Stone (2002) adapted this approach to their study of the impact of the Maryland School Performance Assessment Program (MSPAP) on teaching practices. Using questionnaire data collected from teachers, students, and principals
across the state, they compared students and teachers at tested and untested grade levels on several constructs including the alignment of current math instruction with MSPAP (Current Instruction) and MSPAP instructional impact (Impact). The authors examined the relationship between school-level changes in student math performance on MSPAP over a five year period and school level aggregate Impact, Current Instruction, and the percentage of students eligible for free lunch. They found that Current Instruction and free lunch eligibility predicted student performance in the year in which the data were collected, and that Impact was positively associated with test score gains. This indicates that use of aligned instructional practices affects test scores in the year the practices were used and that teachers who reported that MSPAP greatly affected their instruction helped students improve over time. Both findings seem to provide evidence for the instructional validity of MSPAP.

These findings contrast with research on the impact of test-based reforms on instructional practice. Although testing-related modifications to instruction are well documented (Smith, 1991; Koretz, Baron, Mitchell, & Stecher, 1996; Lane, Park, & Stone, 2002), some have noted that these changes are unlikely to improve students’ conceptual understanding. However, this may be affected by the ways in which teachers adjust their instruction. There is some evidence to suggest that many who report using reform-oriented instruction only make superficial changes in their approach. For example, Schorr et al. (2004) found that, while teachers did increase the use of manipulatives in response to one state’s assessment, they had students follow step-by-step instructions with the manipulatives rather than independently explore mathematical
properties. Teachers also did not ask students to come up with their own solutions to mathematical problems, a necessary step if they are to develop conceptual understanding. Spillane and Zeuli (1999) observed 25 elementary- and middle-school mathematics teachers, selected because they reported implementing mathematics reforms consistent with the National Council of Teacher of Mathematics Standards (NCTM, 1991). The observations revealed a great deal of variation in implementation of the NCTM standards, leading them to conclude that, “at least 11 of our teachers have adapted the reforms in ways that undermine their spirit or their core intent” (Spillane & Zeuli, 1999, p. 20).

Others have debated whether high-stakes tests are effective in getting teachers to adjust their instruction. As Mehrens (1998) argued, “the evidence we do have is inadequate with respect to drawing any cause/effect conclusions about the consequences [of high-stakes tests]. If instruction changes concomitant with changes in both state curricular guidelines and state assessments, how much of the change was due to which variable?” (p. 20).

These issues highlight the importance of consequential validity in determining the appropriateness of test score interpretations. Messick (1989) argued that the consequences of testing are an essential component of test validity because test scores are only useful decisionmaking tools when accurately interpreted. If harmful consequences arise from policies based on inappropriate inferences, then the validity of that interpretation is jeopardized. It is important to note, however, that Messick (1995) also argued that harmful consequences arising from accurate test score interpretations do not endanger the validity of those interpretations. That is, test interpretations are not invalid
just because a teacher mimics test items in her instruction. Test interpretations are only invalid if student performance improves as a result of such behavior.

Standards-based assessments

Despite concerns about the impact of testing on instruction, the era of performance-based assessments did not last long. States abandoned performance-based assessments and instituted more traditional multiple-choice tests. The expense of scoring such tests (and of administering them to the number of students mandated by No Child Left Behind) was perhaps the biggest obstacle to implementing performance-based assessment systems. Stecher and Klein (1997) estimated that performance-based assessment systems could cost as much as 20 times the cost of multiple-choice tests. In addition, the tests were unpopular with some parents, who wanted their children to take nationally-normed tests that could provide information about how they performed relative to other students in the nation (Lawton, 1996).

Most recently, states have adopted standards-based assessments as part of their reform efforts. Standards-based assessments have blended some characteristics of minimum competency, criterion-referenced, and performance-based tests in that they are sometimes used to make high school graduation decisions, are intended to provide feedback on a specific set of skills (as outlined in state standards), and may include rather involved constructed-response items. However, they are distinct from previous testing efforts in that they represent one component of a larger reform effort (Smith & O’Day, 1990). Under standards-based reform, testing contributes to accountability systems. To
encourage schools to provide instruction on state standards, such systems require that corrective actions be taken against schools with poor test performance. There is quite a range of corrective actions taken against schools, from requiring them to implement new curricula or instructional programs to demoting the principal, reassigning school staff, or having a third party (private organization or the state) manage schools (Padilla, Woodward, Lash, Shields, & Laguarda, 2005). Because of the severe consequences associated with poor test performance, it is likely that teachers have intensified efforts to focus their instruction on state assessments.

Teachers might increase their focus on state assessments in many different ways. Some will respond by providing instruction on state standards, the intent of the reform and an approach that is generally considered appropriate (Mehrens & Kaminski, 1989). Others may implement methods that are inappropriate and potentially harmful, such as focusing on item formats without addressing the concepts tested, teaching the concepts tested using only examples that are likely to be presented on tests, or narrowing the curriculum to address only tested concepts, (Darling-Hammond & Wise, 1985; Mehrens & Kaminski, 1989). Finally, teachers may address standards in their classrooms, but present information in a way that differs from the way that standards are operationalized on the assessments. When this occurs, student performance on the assessment should not be jeopardized as long as the information has been presented effectively.

Just as the Yoon and Resnick (1998) study assumed that teachers could accurately report their use of instructional strategies consistent with the NCTM standards, most attempts to examine the instructional validity of standards-based assessments have
assumed that teachers similarly interpret and operationalize state standards. To gauge implementation, researchers typically ask teachers to indicate how much emphasis they place on certain skills or concepts. However, teachers tend to be provided with little information in determining what is intended by each standard. Therefore, teachers may give similar responses to represent a wide array of ways that the skill is presented in class. While some teachers may incorrectly interpret what is intended, others may come up with different but equally accurate ways of presenting a concept, leading to similar treatment of quite different behaviors during analysis. Alignment measures are often derived by comparing teacher responses with content area expert ratings of the degree of emphasis placed on the standards in state tests. Content area experts face the same challenges as teachers in interpreting what is intended by the standards, causing inter-rater reliability problems across content experts.

The Surveys of Enacted Curriculum (SEC), a research effort that now provides content alignment information to state and local education professionals throughout the United States, presents one example of this type of analysis. Teachers participating in the SEC provide detailed information about the amount of time devoted to specific topics (such as phonemic awareness, phonics, vocabulary, etc.) at various levels of cognitive demand (e.g., recall, demonstrate, analyze). State standards and assessments are also reviewed and scored according to the relative emphasis placed on each content area/cognitive demand combination. By comparing teachers’ responses with the ratings of assessments and/or standards, the SEC method generates measures of alignment between instruction and assessments and also presents visual displays that highlight areas
of overlap and divergence. Since the SEC is intended as a tool to help teachers improve
the alignment between state standards, their instruction, and assessment, it provides
content alignment information, but does not evaluate instructional validity. That is, the
relationship between alignment and test performance is not examined (Porter &
Smithson, 2001; Blank, 2002; Smithson & Porter, 2004).

In addition, such methods do not address whether certain instructional methods
are particularly effective. Rather, it is assumed that if the standards are emphasized
equally across classrooms, students will perform equally well on the test. In reality, it is
likely that both emphasis on the tested skills and use of effective instructional methods
affect test performance, in combination with the background knowledge and skills that
students bring with them to the classroom. These, and other resources provided to schools
and classrooms (teacher qualification, instructional resources, etc.) comprise students’
opportunity to learn tested material.

**Opportunity to learn**

Opportunity to learn is a concept that is related to instructional validity in that it is
concerned with the relationship between educational experiences and test outcomes.
Carroll (1963) described opportunity to learn (OTL) as the amount of instructional time
devoted to a concept in his seminal work, which asserted that all students can learn and
that learning is a function of aptitude, the ability to understand instruction, perseverance,
quality of instruction, and the amount of instructional time devoted to a concept. Hence,
OTL differs from instructional validity in that it focuses more on students’ educational
experiences and less on test validation properties.

Since Carroll’s (1963) work, the concept of OTL has broadened and has been applied to a variety of contexts. In educational policy settings, it often refers to educational equity—whether all students are provided with the resources necessary to help them learn—and culminated in efforts to define Opportunity-to-Learn standards, which outlined the minimum amount of resources that should be provided to schools to ensure an adequate education (Porter, 1995). These included efforts to gauge whether instructional materials were of satisfactory quality and provided in sufficient quantities, teachers were properly qualified for their teaching assignments, and instructional content was appropriate for a given grade level of content area (Floden, 2002; McDonnell, 1995).

Awareness that some students could not access adequate educational resources lead to requirements that states receiving Goals 2000 funding develop OTL standards (Goals 2000: Educate America Act, 1994) and to discussions about what comprises “equitable” opportunity to learn (Guiton & Oakes, 1995). A definitive set of OTL standards was never developed, perhaps because of the difficulties associated with specifying all of the components of opportunity to learn.

Instead, research focused on the impact of specific aspects of OTL or used OTL measures as covariates when comparing student outcomes. For example, Wang (1998) attempted to examine the effect of four OTL measures on science achievement: content coverage, content exposure, content emphasis, and instructional quality. To arrive at these measures, she collected information on the ways that six eighth-grade science teachers addressed circuits. Although she found a relationship between three OTL
measures (content coverage, content exposure, and instructional quality) and science achievement, the measures used were weak approximations of the constructs of interest. For example, instructional quality was not gauged using classroom observations, which Wang (1998) notes is the preferred method for obtaining this measure. Instead, she determined whether teachers submitted a lesson plan, whether students had textbooks they could take home, whether teachers taught Ohm’s law, whether students were familiar with the equipment used in the science tests, and the match between student texts and science tests in terms of: concepts addressed, format of presentation, depth of materials and objectives. Therefore, teachers rated high in instructional quality could have presented incorrect information or used ineffective instructional strategies. Other OTL measures presented in this study had similar weaknesses, supporting the assertion that it is difficult to measure some OTL constructs well.

The International Association for the Evaluation of Educational Achievement’s (IEA) First International Mathematics Study presents one attempt at using OTL measures as covariates. The IEA study was designed to internationally compare mathematics achievement on a uniform assessment while controlling for cross-national curricular and instructional differences. Therefore, teachers participating in the study were asked to review each test item and to estimate the percentage of students who had the opportunity to learn the type of problem it represented. This information was then aggregated up to the test level, with mean classroom test performance controlled for by mean teacher ratings across all items on the test (Floden, 2002). Subsequent IEA studies (such as the Second International Mathematics Study) expanded on this line of questioning by asking
national educational representatives to rate the appropriateness of each item for the students in their nation (Schmidt, Wolfe, & Kifer, 1992). By treating curricular differences as covariates, “true” differences in student achievement were better estimated.

OTL research can inform instructional validity studies as some measures used to gauge opportunity-to-learn can also be used to ascertain whether tests are sensitive to instruction. Content coverage and content emphasis are of particular importance to both fields (Stevens, 1993; Wang, 1998) in that test scores of instructionally valid tests should be higher for those students who are exposed to the tested material and for whom the degree of emphasis placed on different topics matches the test. Coverage, which refers to the curricular topics taught by teachers, is commonly gauged by presenting teachers a list of topics and asking them to report the ones that they addressed (for a detailed description of this method, see Popham & Yalow, 1982). Since teachers can cover a topic, but place little emphasis on it, it is also important to obtain an emphasis indicator. This can be measured directly by asking teachers to report the degree of emphasis placed on various topics, or by asking them to report how much instructional time (minutes, days, or class periods) was devoted to the topics (Muthén et al., 1995).

Although some have found that coverage and importance alone tend to be fair predictors of students’ achievement levels (Wang, 1998), others have noted that the variables better predict achievement when combined as a composite measure (Winfield, 1993). In their study of the effect of remedial math programs on high school student achievement, Gamoran, Porter, Smithson, and White (1997) examined differences across courses in the instructional time devoted to topics [referred to as emphasis in the present
study and as level of coverage by Gamoran et al. (1997) and the alignment between instruction and an assessment (which the authors referred to as configuration of coverage). They found college-preparatory math courses to be most rigorous and to devote the most time to topics covered on a test comprised of publicly-released National Assessment of Educational Progress (NAEP) items. In addition, these courses best-matched the test in terms of relative emphasis placed on tested topics. In contrast, courses designed to prepare students for college preparatory math were less-well matched in terms of coverage and emphasis and general math courses were even less-well aligned.

Gamoran et al. (1997) also used these measures to predict student performance on the NAEP-like test. They found that an index of the level of coverage (emphasis) and configuration of coverage (alignment), which they termed content coverage, was associated with test performance. The authors used these results to argue that low-performing students are capable of learning more that is normally required of them in general math courses and that high schools should consider allowing a wider array of students into college preparatory math, or create transition classrooms that expose low-performers to more rigorous content to prime them to take more advanced coursework.

The OTL measures common to these studies (emphasis and alignment) can also be applied to studies of the instructional validity of standards-based assessments. A form of criterion-referenced tests, standards-based assessments are designed to detect instruction on tested content (state standards). While useful for test development purposes, instructional sensitivity measures used with criterion-referenced tests do not address whether test scores reflect successful instruction on state standards. Therefore, it
is essential that instruction be taken into account in studies of the instructional validity of
standards-based assessments. By relating test performance to the instructional emphasis
placed on standards and the alignment between the degree of emphasis during instruction
and on a test, the instructional validity of standards-based assessments can be examined
(D’Agostino et al., 2007).

Conclusion

Instructional validity is an important aspect of test validity in that it examines the
ability of tests to reflect what happens in classrooms. In this era of high-stakes testing,
instructional validity has become especially important in that accountability systems
hinge on the assumption that standards-based assessments can gauge instructional efforts.
While the approaches used in instructional validity studies have evolved as the purpose
and format of assessments have changed, few methods specific to standards-based tests
have emerged. Those that have been conducted tend to use measures of the opportunity to
learn state standards, namely emphasis on state standards and alignment between the
content of instruction and assessment.

While OTL measures are critical to instructional validity investigations, they tend
to be rather simplistic, asking teachers to report which standards they have taught and
with what frequency. More nuanced measures of instruction, which account for how
teachers interpret and operationalize the standards, are needed. The present study builds
upon previous instructional validity research by gauging the full range of instructional
forms used to teach state standards, and, in doing so, determines how teachers interpret
the standards.
CHAPTER 3
METHODS

Participants

Thirty-six teachers participated in the study: 16 third-grade teachers and 20 fifth-grade teachers. The teachers represented one suburban Arizona school district that serves approximately 13,000 students in 17 schools. Its students perform slightly above average on AIMS, are of moderate to high socio-economic status, and are mostly Caucasian compared to other districts in Arizona. Students in the study sample outperformed district and statewide AIMS averages on every measure, often only slightly. However, the sample mean was greater than one-third of a standard deviation better than the state mean in reading and mathematics at third grade (Tables 1 and 2).

The district can be considered a high-stakes testing environment because of its focus on standards-based grading. It has implemented a standards-based report card system since 2002-03, which requires teachers to grade students according to the same performance levels as AIMS, in effect predicting how students will perform on the test. According to a description of standards-based report cards posted on the district website, the goal is to encourage teachers to focus on the academic skills laid out by the state standards instead of incorporating other factors, such as attendance, class participation, and homework completion, in their assessment of students. Comparisons of end-of-year standards-based report card grades and AIMS results provide some evidence that variation exists in teacher understanding of what it means to meet state standards. Hit
Table 1. Mean 2005 AIMS Scale Scores, Study Sample, Participating District and Arizona

<table>
<thead>
<tr>
<th></th>
<th>Math</th>
<th>Reading</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
</tr>
<tr>
<td>Sample, Grade 3</td>
<td>472</td>
<td>47.1</td>
<td>319</td>
</tr>
<tr>
<td>District, Grade 3</td>
<td>464</td>
<td>48.6</td>
<td>976</td>
</tr>
<tr>
<td>Arizona, Grade 3</td>
<td>448</td>
<td>50.3</td>
<td>77,443</td>
</tr>
<tr>
<td>Sample, Grade 5</td>
<td>511</td>
<td>47.2</td>
<td>465</td>
</tr>
<tr>
<td>District, Grade 5</td>
<td>506</td>
<td>48.0</td>
<td>988</td>
</tr>
<tr>
<td>Arizona, Grade 5</td>
<td>501</td>
<td>54.3</td>
<td>76,719</td>
</tr>
</tbody>
</table>

Sources:

Table 2. Demographic Characteristics of Study Sample, District and Arizona, 2004-05

<table>
<thead>
<tr>
<th></th>
<th>Sample (N=784)</th>
<th>District (N=12,994)</th>
<th>Arizona (N=1,043,298)</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>70%</td>
<td>69%</td>
<td>48%</td>
</tr>
<tr>
<td>African American</td>
<td>3%</td>
<td>3%</td>
<td>5%</td>
</tr>
<tr>
<td>Hispanic</td>
<td>24%</td>
<td>24%</td>
<td>38%</td>
</tr>
<tr>
<td>American Indian</td>
<td>2%</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>Asian</td>
<td>2%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>English language learners</td>
<td>3%</td>
<td>4%</td>
<td>19%</td>
</tr>
<tr>
<td>Students with disabilities*</td>
<td>9%</td>
<td>22%</td>
<td>18%</td>
</tr>
<tr>
<td>Percentage of Students Eligible for Free or Reduced-Price Lunch</td>
<td>--</td>
<td>34%</td>
<td>48%</td>
</tr>
</tbody>
</table>

*Note: Although the percentage of students with disabilities seems large, this figure includes students with a wide range of disabilities including: speech/language impairments, mild mental retardation, specific learning disabilities, emotional disabilities, moderate mental retardation, visual impairments, hearing impairments, other health impairments, orthopedic impairments, traumatic brain injury, multiple disabilities, multiple disabilities with severe sensory impairments, autism, and severe mental retardation.

Sources:

Sample and District: *Arizona Instrument to Measure Standards 2005 district results* [Data file]. Phoenix, AZ: Arizona Department of Education.
rates, or the proportion of time teachers assigned end-of-year grades consistent with AIMS scores varied a great deal across classrooms. In particular, teachers assigned students the same performance level as was indicated by their AIMS score between 17% and 81% of the time in third grade mathematics and between 7% and 67% of the time in fifth grade mathematics. On average, teacher grades matched AIMS performance levels for fewer than half of students, with the exception of third grade reading, where the mean hit rate was 0.566 (Table 3).

Participants with at least three years of teaching experience were initially selected for the study to ensure that study measures would not be confounded with inexperience. The goal was to interview forty-eight teachers; twelve in each of four categories: experienced teachers new to (with two or fewer years teaching) third grade, experienced teachers new to fifth grade, teachers experienced at (with 3 or more years teaching) third grade, and teachers experienced at fifth grade. Given the relatively small size of the district, this required participation of the entire population in some categories.

There are several reasons for selecting this group of teachers. First, at the elementary level, the AIMS test had only been administered at grades 3 and 5 in Spring 2003 and 2004. Teachers at these grade levels would be familiar with AIMS and are much more likely to present course content similarly to AIMS items than teachers at other grade levels, who were not exposed to the test until 2005. Second, teachers need to be familiar with state standards to participate in the study. Experience teaching the same grade level for several years ensures some familiarity with the standards. Teachers new to the grade level were interviewed for comparative purposes.
Table 3. Standards-based Report Card-AIMS Hit Rates in Participating District, 2004

<table>
<thead>
<tr>
<th></th>
<th>3rd Grade</th>
<th></th>
<th>5th Grade</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>N</td>
<td>Mean (SD)</td>
<td>N</td>
</tr>
<tr>
<td>Mathematics</td>
<td>.479 (.500)</td>
<td>702</td>
<td>.429 (.495)</td>
<td>767</td>
</tr>
<tr>
<td>Reading</td>
<td>.566 (.496)</td>
<td>719</td>
<td>.424 (.495)</td>
<td>753</td>
</tr>
<tr>
<td>Writing</td>
<td>.451 (.498)</td>
<td>661</td>
<td>.437 (.496)</td>
<td>750</td>
</tr>
</tbody>
</table>

Note: Hit rates are computed as the proportion of time teachers assigned end-of-year grades consistent with AIMS scores.

Recruiting teachers proved quite difficult, resulting in a four-stage approach. First, letter and e-mail contacts informed teachers of the study. Second, teacher interest was fostered by presenting the convergence between standards-based report card grades and AIMS scores at each school, coupled with classroom-level convergence reports. Teachers were asked to sign up for interviews at the end of the presentations. Third, each teacher received an origami flower with a lifesaver stamen and a note that said “Please Be A Lifesaver (and participate in the study).” The note included contact information to sign up for interviews and a postcard teachers could return with an interview time and location. Finally, teachers who did not set an interview time were contacted via phone.

Forty-one teachers were interviewed, but analyses focused on 36 mathematics classrooms. One experienced third-grade teacher was removed from the analysis because she taught a classroom serving students in grades one through three. Only five third-graders from this classroom had both 2005 AIMS mathematics scores and 2004 SAT-9 scores, a requirement for analysis. Because of the small number of third-graders served and the unusual classroom configuration, this classroom was omitted. An experienced fifth-grade teacher was also omitted because of technical difficulties that arose during the interview. The batteries in the tape recorder used to document the interview stopped functioning and the back-up batteries brought along to the interview also did not work. Therefore, field notes were the only record of this interview and were not detailed enough to provide comparable information about instructional strategies.

Finally, fifth grade is departmentalized in three schools so that mathematics is taught to all fifth-graders by the same teacher and reading is taught to all students by
another teacher. At these schools, two teachers were interviewed—the reading and the mathematics teacher—to ensure that students’ exposure to mathematics content outside of math time was captured. In these cases, the two classrooms were treated as one large mathematics classroom, with students from both classrooms linked to the same teacher, and both interviews contributing to alignment and emphasis ratings. Other teacher-level measures (e.g., teacher quality indicators) were generated by taking the mean of the two teachers. In every case, the teacher pairs had similar levels of experience and education; the mean was not much different from the individual ratings.

The resulting sample varied a great deal in terms of overall teaching experience, experience at the grade levels of interest, and in education level. On average, teachers had more than 10 years of classroom experience and more than five years experience at their grade level. However, years of total teaching experience ranged from first-year teachers to 30-year veterans, with between one and 27 years spent at their current grade level. Despite efforts to omit first-year teachers from the study, two first-year third-grade teachers were interviewed because their status was not discovered until interviews were underway. Due to the already small number of participants, they were included in analyses. Finally, although only one teacher specialized in mathematics, approximately one-third possessed master’s degrees (Table 4). These measures were used to create teacher qualifications composite, which was used as a covariate to control for differences between teachers in their level of training and experience. The composite was generated through principal components analysis; it is the factor score generated by forcing a one factor solution to an analysis of the three measures.
Table 4. Teacher Qualifications in Participating Classrooms

<table>
<thead>
<tr>
<th></th>
<th>3rd Grade Mean (SD) (n=16)</th>
<th>5th Grade Mean (SD) (n=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total years teaching</td>
<td>12.7 (10.6)</td>
<td>14.9 (9.3)</td>
</tr>
<tr>
<td>Years teaching at grade level</td>
<td>7.4 (7.8)</td>
<td>6.3 (5.6)</td>
</tr>
<tr>
<td>Percentage Master’s degrees</td>
<td>37.5</td>
<td>35.0</td>
</tr>
</tbody>
</table>
Rigorous recruiting efforts were particularly important because of the relatively small number of third- and fifth-grade teachers in the district. Sixty-four teachers met the study requirements: 11 teachers new to third grade, 23 teachers experienced at third grade, and 15 teachers each new to and experienced at fifth grade. Just over half (36 teachers or 56%) of all eligible teachers were interviewed: five teachers new to third grade, 12 teachers experienced at third grade, nine teachers new to fifth grade, and 11 teachers experienced at fifth grade. The small number of teachers in some cells, especially teachers new to third grade, precluded comparisons of teachers new and experienced at each grade level. Instead, data were aggregated across all classrooms.

Study participants worked at all 11 district elementary schools and taught students of similar ethnic and language backgrounds as the districtwide population. However, participating classrooms served fewer students with disabilities than is represented in the district population (Table 2).

Materials

Ninety-minute teacher interviews, focusing on instruction and assessment of two mathematics performance objectives, are the primary data source for the study. Teachers prepared for interviews by gathering the materials they used in teaching and assessing students. Teachers who forgot to bring assessments to interviews wrote out examples on blank sheets of paper.

During interviews, teachers were asked to provide information about their experience and education level, to describe their instruction and assessment practices, and
to review testlets of each performance objective. Of particular importance to the study were the amount of time devoted to teaching the objectives, key concepts teachers attempt to convey, types of activities used in instruction, instructional examples employed, and evidence required to deem that a student has “met” state standards.

Classroom assessment items helped teachers elucidate how they operationalized performance objectives. In particular, by discussing requirements for a correct response, teachers provided fine-grained detail about the subskills they deemed relevant to each objective. For example, classroom assessment items were used to determine not only the types of graphs teachers addressed when covering the graphing objective, but also helped identify the types of questions students must answer about graphs—whether students were required to make comparisons, make predictions, or simply to pull information from the graph.

Teachers were also asked how much time they devote to, or emphasis they place on, the objectives and to describe their lessons in detail. This helped to determine whether they require students to apply mathematical concepts in a variety of ways. This establishes not only whether teachers operationally defined the performance objectives narrowly, so that they specifically match AIMS items, but also whether they attempt to help students generalize content knowledge to novel situations. The interview protocol is included in Appendix A.

Performance objectives (Table 5) were selected for the study based on two criteria. First, they had to be represented by multiple AIMS items on previous tests (test blueprints for the 2005 exam were not available prior to data collection). Second, they
<table>
<thead>
<tr>
<th>Performance Objective</th>
<th>Grade 3</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective 1</td>
<td>Make a diagram to represent the number of combinations available when 1 item is selected from each of 3 sets of 2 items (e.g., 2 different shirts, 2 different hats, 2 different belts). <strong>2 AIMS Items (2005)</strong></td>
<td>Interpret graphical representations and data displays including bar graphs, circle graphs, frequency tables, three-set Venn diagrams, and line graphs that display continuous data. AND Answer questions based on graphical representations and data displays. <strong>4 AIMS Items (2005)</strong></td>
</tr>
<tr>
<td>Objective 2</td>
<td>Discriminate necessary information from unnecessary information in a given grade-level appropriate word problem. <strong>3 AIMS Items (2005)</strong></td>
<td>Describe the rule used in a simple grade-level appropriate function (e.g., T-chart, input-output model). <strong>4 AIMS Items (2005)</strong></td>
</tr>
</tbody>
</table>
had to be included in grade level standards from 2002-03 through 2004-05 (the year study data were collected). This time frame is important because the state standards changed in spring 2003, with most changes in math occurring at the third grade level. To increase the probability that teachers were familiar with the standards, objectives included in both the “old” and “new” standards were required. The spring 2005 version of AIMS was revised to reflect the new standards.

These selection criteria greatly narrowed the list of potential objectives. Few objectives were addressed more than once on AIMS. In fact, at the third grade level, the study objectives were the only objectives assessed more than one time. In addition, AIMS revisions included changes to the way that study objectives were assessed. Teachers who based their instruction of the study objectives on the “old” AIMS items would not be congruent with the spring 2005 test.

Classroom assessment examples (collected and discussed during interviews), AIMS items (from 2004 and 2005), publicly released items from other state’s assessments, AIMS 2005 test scores, Stanford Achievement Test, 9th Edition (Stanford-9) scores from the 2004 test administration, and student demographic data also contribute to analyses. Classroom assessments are largely comprised of items from the district-adopted text, Everyday Mathematics. Therefore, cross-classroom variation largely comes from requirements teachers place on students in completing assessments. Albeit, some teachers used assessments that they developed or borrowed from other sources.

2004 AIMS items and publicly-released items from other states’ assessments were used to generate testlets that assessed the performance objectives of interest (included in
Appendix B), which teachers reviewed after describing their teaching and assessment practices. While reviewing testlet items, teachers were asked to comment on the similarity between the items and the content of classroom assessments and also to judge whether they believed their students would be able to correctly answer the items. This became another source of information about teaching practice, as teachers often provided more detail about their instruction while reviewing testlets. To check whether alignment ratings were biased by testlet review, raters first generated alignment scores for each item without reading the testlet review section of transcripts and then scored a second time incorporating this information. Alignment ratings changed for only a handful of teachers (three fifth-grade and four third-grade teachers) after reviewing testlet comments, and each teacher’s alignment rating was changed for one item only.

The 2005 AIMS items were used to identify the subskills associated with each objective; the basis of rubrics used to judge instructional alignment with AIMS. Student-level 2005 AIMS scores, and responses to the items (five third-grade items and eight fifth-grade items) that assess the study performance objectives are used as dependent variables in analyses.

Although many of the items have remained the same from year to year, one challenge to working with AIMS scores is that the test format changed. In 2002-03, the mathematics test included both multiple choice and constructed response items. The test has only included multiple choice items since 2003-04. However, multiple choice items changed significantly between 2003-04 and 2004-05 when the test was redesigned to provide both criterion-referenced and norm-referenced scores. The 2004-05 test
incorporates items from the 2003-04 AIMS, newly-developed items, and items from the Terra Nova, an off-the-shelf, nationally-normed test. These changes in the testing program guaranteed that teachers would not be familiar with items currently in use.

Finally, student-level Stanford-9 scores collected in spring 2004 (when students were in second and fourth grade) and demographic characteristics (ethnicity, English learner status, grade level, and special education status) from the 2005 AIMS data file were all used as student-level covariates in analyses, as were classroom level measures of teacher qualifications, and participation in the free and reduced-priced lunch program (measured at the school level). Grade was used both to identify and control for any grade level differences that might exist, a likely occurrence since both the curriculum and test differ by grade level. Although student-level socio-economic status could also have been a useful covariate, prior-year test performance should correlate with this measure and account for much of the variation associated with poverty.

Procedure

Before interviewing participants, the researcher conducted two pilot interviews with teachers in a neighboring district to check equipment, assess the clarity and flow of interview questions, and ensure that the interview protocol was an appropriate length. Equipment consisted of a small audiotape recorder.

Interviews began by obtaining informed consent. The researcher explained the purpose of the study, outlined interview topics, reviewed the informed consent form, and reminded teachers that they could skip any questions or end the interview at any time. If
teachers agreed to participate in the study, the consent form was signed and the interview commenced.

Interviews were audiotaped and later transcribed. Transcripts were coded in several ways. First, two coders, a former elementary school teacher and a district mathematics specialist, reviewed 2005 AIMS items and created rubrics for each item that described the skills teachers would need to address to receive one of four scores: perfect alignment with the item (scored “3”), close alignment (scored “2”), some alignment (scored “1”), and not aligned (scored “0”). To create the rubrics, each coder independently listed the skills required to answer each item correctly and organized the skills into a proposed rubric. Coders compared rubrics and made adjustments to arrive at a final rubric, with the math specialist given final say because of her expertise.

An example of a scoring rubric is presented in Figure 3. This item assesses the objective, “Make a diagram to represent the number of combinations available when 1 item is selected from each of 3 sets of 2 items (e.g., 2 different shirts, 2 different hats, 2 different belts).” It presents the problem both in words and in a table and asks students to identify the tree diagram that displays all of the possible lunch combinations when selecting one of two possible drinks, sandwiches, and snacks. Therefore, for instruction to be “perfectly aligned” with the item, teachers must require students to interpret a table, both written and visual information, and a three set tree diagram. Instruction that is “closely aligned” with the item might require students to interpret a tree diagram or require students to deal with combinations of three sets of items using multiple approaches (that do not include a tree diagram). Instruction with “some alignment”
Figure 3. Scoring Rubric for One AIMS Item

<table>
<thead>
<tr>
<th>Perfect Alignment</th>
<th>Interprets a table</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Interprets visual and written information</td>
</tr>
<tr>
<td></td>
<td>Interprets a 3 set tree diagram</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Close Alignment</th>
<th>Combinations involve 3 sets of items AND multiple visual displays</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR students create a tree diagram</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Some Alignment</th>
<th>Introduces concept of combination:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Selects 1 item from each set</td>
</tr>
<tr>
<td></td>
<td>Represents combination in some way (list or diagram)</td>
</tr>
<tr>
<td></td>
<td>Uses relevant vocabulary (combination, diagram, different)</td>
</tr>
</tbody>
</table>

| Not Aligned       | Does not teach skill |

---

David is packing his lunch. The chart below shows the choices for his lunch.

<table>
<thead>
<tr>
<th></th>
<th>Sandwich</th>
<th>Drink</th>
<th>Snack</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Peanut butter and jelly</td>
<td>Juice</td>
<td>Cookie</td>
</tr>
<tr>
<td></td>
<td>Cheese</td>
<td>Milk</td>
<td>Cake</td>
</tr>
<tr>
<td>B</td>
<td>Peanut butter and jelly</td>
<td>Juice</td>
<td>Cookie</td>
</tr>
<tr>
<td></td>
<td>Cheese</td>
<td>Milk</td>
<td>Cake</td>
</tr>
<tr>
<td>C</td>
<td>Peanut butter and jelly</td>
<td>Juice</td>
<td>Cookie</td>
</tr>
<tr>
<td></td>
<td>Cheese</td>
<td>Juice</td>
<td>Cake</td>
</tr>
<tr>
<td>D</td>
<td>Peanut butter and jelly</td>
<td>Juice</td>
<td>Cookie</td>
</tr>
<tr>
<td></td>
<td>Cheese</td>
<td>Milk</td>
<td>Cookie</td>
</tr>
</tbody>
</table>

Which diagram shows all the different combinations of 1 kind of sandwich, 1 kind of drink, and 1 kind of snack that David can choose?
introduces the concept of combinations, but does not address tree diagrams or three sets of items and instruction that is “not aligned” does not deal with combinations at all. Similar scoring rubrics were generated for all 13 AIMS items that addressed the study performance objectives and are included in Appendix C.

Once rubrics were created, coders then independently reviewed interview transcripts and classroom assessment examples to generate holistic scores of the instructional alignment with each item. Coders scored teachers on each item twice, once without considering reactions to the testlet comprised of old AIMS items and items culled from other states assessments and once incorporating comments made during testlet review. Both scores contributed to alignment ratings for each performance objective and for overall grade level alignment, which were generated by summing ratings.

For example, Figure 3 presents some classroom assessment information provided by one third-grade teacher. It includes drawings the teacher made while explaining how she assesses the combinations objective. As she drew, she explained that she presents the problem to students in tabular and in written form (represented by the squiggles and box with pictures in it at the top of the page). She also requires her students to think through combinations of three sets of items, to construct a picture or tree diagram to represent the combination, and to write a number sentence that could be used to determine how many unique combinations are possible. In some ways this assignment is more complex than the objective; the number of items in one set is greater than two, and students must also generate a number sentence to symbolize the problem. However, the coders decided to score this problem as having “close alignment” rather than “perfect alignment” because
Figure 3. Classroom assignment example

Draw a picture or a tree graph.

<table>
<thead>
<tr>
<th>PANTS</th>
<th>SHOES</th>
</tr>
</thead>
<tbody>
<tr>
<td>red shirt</td>
<td>black - blue</td>
</tr>
<tr>
<td></td>
<td>black - black</td>
</tr>
<tr>
<td></td>
<td>blue - blue</td>
</tr>
<tr>
<td></td>
<td>blue - black</td>
</tr>
</tbody>
</table>

green

yellow

number sentence -
the AIMS item requires students to interpret a tree diagram instead of construct one. Students were not asked to interpret tree diagrams in this classroom.

This approach to coding represents one of many approaches that can be used to “quantitize” qualitative data (Miles & Huberman, 1994; Tashakkori & Teddie, 2003). One common approach is to count the frequency with which certain themes are mentioned during interviews. This is not useful for this study as the indicators of interest (emphasis placed on each objective and degree of alignment between instruction and AIMS items) can not be determined simply from the number of times teachers discuss the objectives.

Rather, this study employed data transformation methods that help expand the breadth of inquiry beyond what is possible with solely quantitative or qualitative approaches (Caracelli & Greene, 1993). Teachers were interviewed to achieve deep understanding of the instruction and assessment that occurred; the ways assessments were applied and interpreted instead of the number of times certain practices were used. Quantitative analyses explore the relationship between these assessment practices and student achievement.

It can be challenging, however, to achieve sufficient levels of inter-rater reliability when scorers are asked to judge a large quantity of information using a holistic rubric. In this case, teachers discussed one performance objective for as long as 30 minutes and often provided multiple assessment examples.

Independently generated scores were compared to determine the degree of inter-rater reliability (Table 6) using three measures, Cohen’s kappa, kappa divided by the
Table 6. Inter-rater Agreement on Overall Alignment Ratings

<table>
<thead>
<tr>
<th></th>
<th>Maximum Possible Kappa</th>
<th>K/Max(K) Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kappa (K)</td>
<td>(Max(K))</td>
</tr>
<tr>
<td>Third grade</td>
<td>.06</td>
<td>.37</td>
</tr>
<tr>
<td>Combinations</td>
<td>.19</td>
<td>.33</td>
</tr>
<tr>
<td>Necessary/unnecessary information</td>
<td>.16</td>
<td>.36</td>
</tr>
<tr>
<td>Fifth grade</td>
<td>.16</td>
<td>.53</td>
</tr>
<tr>
<td>Graphing</td>
<td>.24</td>
<td>.51</td>
</tr>
<tr>
<td>Functions/t-charts</td>
<td>.22</td>
<td>.30</td>
</tr>
</tbody>
</table>

Note: Although raters coded the degree of alignment with individual AIMS items, ratings were aggregated (averaged) to the objective level and also across grade levels. Kappa coefficients are generated from these aggregate (overall) scores.
maximum possible kappa, and agreement rate. All three statistics are based on teacher-level alignment ratings, calculated by aggregating alignment scores generated for each item. Agreement rate is simply the number of times that raters generated the same score divided by the total number of scores. Kappa is a measure of agreement rate that has been adjusted to account for chance agreement based on the marginal distributions of ratings (Cohen, 1960). Kappa ranges from zero to one, with a score of one indicating perfect agreement. However, the maximum possible value of kappa is not always one. It can be much less, depending on the number of categories available to raters and on the marginal distribution of scores. In this study, the maximum possible kappa varied greatly because of the differing numbers of items scored per objective. Therefore, dividing kappa by the maximum possible kappa allows for comparison between subtests. Although coding occurred on an item-by-item basis, all three figures are based on the mean of item-level ratings which were averaged within performance objectives and also within grade levels to arrive at overall scores.

As shown in Table 6, raters only achieved moderate to low agreement rates on all measures, with the exception of the fifth-grade functions subtests which were rated only once because all four items were quite similar. The large range of possible scores and the vast amount of information judged likely played a role in this outcome. Therefore, the study adopted an approach for arriving at consensus on each rating.

When codes diverged, raters were asked to review transcripts a second time and to re-code while blind to prior codes. If codes still diverged, raters discussed their rationale for scoring and came up with an agreed upon score. In the vast majority of cases,
disagreements were quickly resolved and were usually caused because one coder had missed some relevant section of the (quite lengthy) transcripts. Final item-level scores were summed for each subtest and standardized to generate the alignment measure.

Alignment is one of two main effects examined in relation to student mathematics performance using Hierarchical Linear Modeling (HLM). The second main effect examined, degree of emphasis placed on the objectives, was much more straightforward to code. Therefore, the researcher coded it alone. Teachers were asked how much time they devoted to each objective. The mean number of days devoted to the two objectives was calculated from their responses. It was converted to an ordinal scale: “0” not taught; “1” once or twice a year; “2” one week per year, “3” approximately 10 days per year, “4” every other week, “5” weekly, and “6” daily. This scale was then standardized relative to all classrooms (third and fifth grade) in the sample. The interaction between alignment and emphasis was also computed and standardized. All variables included in analyses are presented in Table 7.

There are several benefits of using HLM (Raudenbush & Bryk, 2002). First, it allows for analyses that take into account the nested structure of data. In this study, students are nested within classrooms. HLM simultaneously accounts for both student-level and classroom-level effects. For example, students enter school with differing mathematical knowledge, socioeconomic status, English proficiency, and learning challenges. Therefore, they may experience the same teaching quite differently. HLM models both the average effects of classroom characteristics and accounts for how classroom effects affect individual students.
Table 7. Analysis Variables

**Dependent Variables**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIMS</td>
<td>Performance on the 2005 AIMS mathematics test, standardized relative to all students in Arizona</td>
</tr>
<tr>
<td>Thirdtest</td>
<td>Performance on the five 2005 AIMS items that assess the 3\textsuperscript{rd} grade study objectives, standardized relative to all students in Arizona. Coefficient alpha=.49</td>
</tr>
<tr>
<td>Fifthtest</td>
<td>Performance on the eight 2005 AIMS items that assess the study objectives, standardized relative to all 5\textsuperscript{th} grade students in Arizona. Coefficient alpha=.68</td>
</tr>
</tbody>
</table>

**Independent Variables**

**Student-level covariates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAT9</td>
<td>Performance on the mathematics 2004 Stanford Achievement Test, 9\textsuperscript{th} Edition, standardized relative to the study sample</td>
</tr>
<tr>
<td>Minority*</td>
<td>Ethnicity, coded “1” for minority and “0” for non-minority</td>
</tr>
<tr>
<td>ELL*</td>
<td>English language learner status, coded “1” for English language learner and “0” for non-English language learner</td>
</tr>
<tr>
<td>SPED*</td>
<td>Special education status, coded “1” for special education students and “0” for non-special education student</td>
</tr>
<tr>
<td>Grade*</td>
<td>Student grade level, coded “1” for fifth and “0” for third</td>
</tr>
</tbody>
</table>

**Classroom-level covariates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freelunch</td>
<td>School percentage of students free- or reduced-price lunch eligible in 2005, standardized relative to the study sample</td>
</tr>
<tr>
<td>Tqual</td>
<td>Teacher qualifications composite, created through principal components analysis using the following factor weights (measures): .544 (overall experience in years), .475 (experience at their current grade level), and .241 (educational attainment). The factor captured 58% of the total variance (eigenvalue=1.726). Coefficient alpha=.57</td>
</tr>
</tbody>
</table>

**Classroom-level predictors**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alignment</td>
<td>Sum of alignment ratings for each grade level, standardized relative to the study sample</td>
</tr>
<tr>
<td>Emphasis</td>
<td>Mean of objective-level emphasis reports, categorized on a 7-point scale and standardized relative to the study sample</td>
</tr>
<tr>
<td>Alignment XEmphasis</td>
<td>The product of alignment and emphasis ratings, standardized relative to the study sample</td>
</tr>
</tbody>
</table>

*Measures based on data entered in 2005 AIMS datafiles.
One challenge of classroom level analyses is that the number of cases within classrooms can be quite small and only a small number of students with particular background characteristics (such as special education students) can be so small as to make it impossible to accurately estimate their effect. In addition, biased estimates can result from aggregating data; one method of increasing the number of cases with rare characteristics. This occurs because selection artifacts may exist that cause some classrooms to serve these special groups, which aggregation fails to take into account, HLM improves parameter estimates by weighting them to include information for a particular classroom and for all classrooms.

HLM also allows cross-level effects to be modeled. For example, in this study, the relationship between classroom-level measures of instructional alignment with AIMS and emphasis on study performance objectives and student-level AIMS performance is evaluated.

Finally, HLM can be used to partition the variance and covariance components between levels for unbalanced, nested data. For example, the proportion of variance in AIMS performance that exists between- and within-classrooms is examined (See Raudenbush & Bryk, 2002 for more information about HLM).

Multilevel models, such as HLM, have been used in other instructional validity studies. Yoon and Resnick (1998) used student-level measures of minority status, parent education and attendance in a Renaissance (intervention) classroom and classroom-level measures of participation in the Renaissance program, staff development, and instructional approach (traditional and reform-oriented) to predict student achievement.
The authors used classroom-level measures to predict mean classroom achievement (the intercept) only, the same approach used in this study. They found that use of the reform-oriented instructional techniques advocated by the intervention (the Renaissance program) was associated with improved student achievement. Other analyses indicated that participants in the Renaissance program more frequently applied these strategies and that staff development increased their implementation.

This study also employs a model similar to one used by Lee, Croninger, and Smith (1994) in their analysis of the relationship between family demographics, school resources, and opinions of school choice in greater Detroit, Michigan. That study also predicted mean Level-2 scores using Level-1 covariates and Level-2 main effects, holding all predictors as fixed effects. Other than the variables studied, the main difference between the two approaches is that that study used HLM to predict a dichotomous outcome, adults being for or against school choice, while the research discussed here uses an interval measure as the dependent variable.

Lee et al. (1994) investigated whether the predilection for school choice is associated with the following individual-level measures: ethnicity, social class (a composite of income and education), having school-age children, rating of local schools, opinion of the effect of choice on achievement, and experience residing in the city of Detroit. The second level of the HLM model estimated the main effect of district resources (respondents were linked to their school district of residence), a composite of percentage of students receiving free lunch, minority enrollment, graduation rate, and taxable property wealth per student enrolled.
They found that resource differences between districts affect individual’s opinions of school choice after individual demographics are taken into account. They also ascertained that less advantaged individuals held more favorable views of school choice and that district ratings are associated with opinions of choice.

HLM analyses inform this study in multiple ways. First, the amount of variation between- and within-classrooms is examined to determine if there is enough variation between classrooms to warrant investigation of instructional practice. If AIMS performance does not vary between classrooms, then examining teacher practices will not be helpful as there is no difference in student performance to explain. A fully unconditional HLM model—a one-way ANOVA analysis with random effects generated by the HLM software—accomplishes this. The equations representing this analysis follow the general model:

\[
\begin{align*}
(\text{Student level}) & \quad Y_{ij} = \beta_{0j} + r_{ij} \\
(\text{Classroom level}) & \quad \beta_{0j} = \gamma_{00} + u_{0j} \\
(\text{Combined}) & \quad Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}
\end{align*}
\]

where

- \(Y_{ij}\) is the mathematics performance of student \(i\) in classroom \(j\), expressed in units of standard deviation from the state mean
- \(\beta_{0j}\) is the average mathematics performance of classroom \(j\), expressed in units of standard deviation from the state mean
\( \gamma_{00} \) is the average mathematics performance of all classrooms in the sample, expressed in units of standard deviation from the state mean.

\( u_{0j} \) is the random effect associated with classroom \( j \).

\( r_{ij} \) is the random effect associated with student \( i \) in classroom \( j \).

The analysis was run five times, with \( Y_{ij} \) representing (1) performance on the mathematics portion of the 2005 AIMS test for both third and fifth grade, (2) performance on the 2005 AIMS mathematics test for both third grade only, (3) performance on the 2005 AIMS mathematics test for fifth grade only, (4) performance on the eight fifth-grade 2005 AIMS mathematics test items of relevance to this study, and (5) performance on the five third-grade items. For each analysis, the student level (within-classrooms) variance is an estimate of the true variation in student performance around the grand mean (e.g., mean of districtwide AIMS scores), denoted \( \hat{\text{Var}}(r_{ij}) \). The classroom level (between-classrooms) variance is an estimate of the true variation in classroom means around the grand mean, denoted \( \hat{\tau}_{00} \).

Next, student-level and classroom-level covariates are examined in relation to student achievement. Analyzing the relationship between control variables and student achievement before incorporating the study’s predictor variables serves three purposes. First, because students are not assigned to classrooms at random, failure to control for student characteristics may bias estimates of classroom effects. Second, if the student-level covariates are strongly related to student achievement, controlling for them will increase the power to find an effect by increasing the precision of estimates and reducing
the amount of unexplained within-classrooms variance. Finally, incorporating teacher qualifications and school poverty into the model helps to ensure that predictor parameter estimates are not confounded with these measures (Raudenbush & Bryk, 2002).

Statistically significant ($p<0.05$) covariates were retained in preparation for the final analysis, which incorporates instructional validity measures into the model. The model that includes all covariates is expressed as follows:

(Student level) \[ Y_{ij} = \beta_{0j} + \beta_{1j}(\text{SAT9}) + \beta_{2j}(\text{Minority}) + \beta_{3j}(\text{ELL}) + \beta_{4j}(\text{SPED}) + \beta_{5j}(\text{Grade}) + r_{ij} \]  

(School level) \[ \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Freelunch}) + \gamma_{02}(\text{Tqual}) + u_{0j} \]

\[
\begin{align*}
\beta_{1j} &= \gamma_{10}, \\
\beta_{2j} &= \gamma_{20}, \\
\beta_{3j} &= \gamma_{30}, \\
\beta_{4j} &= \gamma_{40}, \\
\beta_{5j} &= \gamma_{50}.
\end{align*}
\]

(Combined) \[ Y_{ij} = \gamma_{00} + \gamma_{01}(\text{Freelunch}) + \gamma_{02}(\text{Tqual}) + \gamma_{10}(\text{SAT9}) + \gamma_{20}(\text{Minority}) + \gamma_{30}(\text{ELL}) + \gamma_{40}(\text{SPED}) + \gamma_{50}(\text{Grade}) + u_{0j} + r_{ij} \]

where

$Y_{ij}$ is the mathematics performance of student $i$ in classroom $j$, expressed in units of standard deviation from the state mean

$\beta_{0j}$ is the average performance of classroom $j$, adjusted for differences in student demographics
\( \beta_{1ij} \) is the effect of student achievement on the 2004 mathematics portion of the SAT-9 (grand mean centered) on 2005 AIMS mathematics performance

\( \gamma_{10} \) is the pooled within-group regression coefficient for prior achievement on the SAT-9

\( \beta_{2ij} \) is the effect of minority status on 2005 mathematics performance

\( \gamma_{20} \) is the pooled within-group regression coefficient of minority status on 2005 mathematics performance

\( \beta_{3ij} \) is the effect of ELL status (grand mean centered) on 2005 mathematics performance

\( \gamma_{30} \) is the pooled within-group regression coefficient of ELL status on 2005 mathematics performance

\( \beta_{4ij} \) is the effect of special education status (grand mean centered) on 2005 mathematics performance

\( \gamma_{40} \) is the pooled within-group regression coefficient of special education status on 2005 mathematics performance

\( \beta_{5ij} \) is the effect of grade level (grand mean centered) on 2005 mathematics performance

\( \gamma_{50} \) is the pooled within-group regression coefficient of grade level on 2005 mathematics performance

\( \gamma_{01} \) is the effect of proportion of students eligible for free- or reduced-priced lunch (grand mean centered) on mean classroom 2005 mathematics performance
\( \gamma_{02} \) is the effect of teacher qualifications composite (grand mean centered) on mean classroom 2005 mathematics performance

\( u_{0j} \) is the random effect associated with classroom \( j \)

\( r_{ij} \) is the random effect associated with student \( i \) in classroom \( j \)

As with Equations 1.1-1.3, the model is identical for all five dependent measures (AIMS overall, third-grade AIMS, fifth-grade AIMS, Thirdtest, and Fifthtest).

Finally, the classroom-level predictors, alignment, emphasis, and the alignment X emphasis interaction, are introduced into the model to determine if they are associated with mathematics performance after adjusting for variation related to the covariates. As with the covariates, these predictors are entered as fixed effects. That is, the study assumes that the effect of alignment, emphasis, and the alignmentXemphasis interaction is the same across all classrooms. Although it would be possible to model the differential effect of these measures on student achievement across a variety of student level characteristics, there is no literature to suggest which factors might contribute to this differential effect and it is beyond the scope of this study to explore such a relationship. Therefore, the study did not apply any other models.

The equations representing the final HLM models are presented below. Parameter definitions are presented only for those parameters estimated for the first time in the final model. As with the other analyses, the analysis is conducted five times, for overall 2005 AIMS mathematics performance across grade levels, for AIMS performance examined one grade level at a time, and for testlet results relating to the objectives of interest at
third and fifth grade. Grade is included only in the overall AIMS analysis.

(Student level) \[ Y_{ij} = \beta_{0j} + \beta_{1j} (\text{SAT9}) + \beta_{2j} (\text{Minority}) + \beta_{3j} (\text{ELL}) + \beta_{4j} (\text{SPED}) + \beta_{5j} (\text{Grade}) + r_{ij} \] (3.1)

(School level) \[ \beta_{0j} = \gamma_{00} + \gamma_{01} (\text{Freelunch}) + \gamma_{02} (\text{Tqual}) + \gamma_{03} (\text{Alignment}) + \gamma_{04} (\text{Emphasis}) + \gamma_{05} (\text{AlignmentXEmphasis}) + u_{0j} \]

\[ \beta_{1j} = \gamma_{10} \]
\[ \beta_{2j} = \gamma_{20} \]
\[ \beta_{3j} = \gamma_{30} \]
\[ \beta_{4j} = \gamma_{40} \]
\[ \beta_{5j} = \gamma_{50} \]

(Combined) \[ Y_{ij} = \gamma_{00} + \gamma_{01} (\text{Freelunch}) + \gamma_{02} (\text{Tqual}) + \gamma_{03} (\text{Alignment}) + \gamma_{04} (\text{Emphasis}) + \gamma_{05} (\text{AlignmentXEmphasis}) + \gamma_{10} (\text{SAT9}) + \gamma_{20} (\text{Minority}) + \gamma_{30} (\text{ELL}) + \gamma_{40} (\text{SPED}) + \gamma_{50} (\text{Grade}) + u_{0j} + r_{ij} \] (3.3)

where

\( \gamma_{03} \) is the effect of alignment rating (grand mean centered) on mean classroom 2005 mathematics performance

\( \gamma_{04} \) is the effect of emphasis rating (grand mean centered) on mean classroom 2005 mathematics performance

\( \gamma_{05} \) is the effect of alignment-emphasis interaction (grand mean centered) on mean classroom 2005 mathematics performance
CHAPTER 4
RESULTS

This chapter presents the results of five analyses conducted with two-level HLM models. It begins by discussing differences in the third- and fifth-grade curriculum that may affect analyses. Next, descriptive statistics and intercorrelations for all measures are addressed. It is important to examine these statistics prior to running multilevel analyses to determine whether enough variation exists in dependent measures and to gauge the extent to which multicollinearity might diffuse estimates.

The five multilevel models follow the same structure, predicting mathematics performance with measures of instructional alignment with AIMS, degree of emphasis on study performance objectives, and the interaction between alignment and emphasis. Each analysis controls for student-level performance in the prior year, as well as including student demographic characteristics and classroom-level measures of teacher qualifications and student poverty as covariates. Several outcomes are studied, all based on the 2005 AIMS test. First overall performance, investigated across grade levels and at third and fifth grade individually, is examined. Then, performance on AIMS items relating to the study objectives are aggregated to create testlet scores that are analyzed one grade level at a time.

Differences between Grade Levels

Differences between grade levels in the newly-adopted district curriculum may
affect the degree of alignment between instruction and the AIMS test and also the amount of emphasis placed on study objectives. All of the third grade teachers interviewed reported that the curriculum does not address many of the topics covered by AIMS, while all fifth-grade teachers reported that the text addressed study objectives at least to some degree.

This presented a bit of a quandary for the third-grade teachers as the district directed them to teach the text exactly as written, in the exact order presented without skipping any activities or adding in new topics. Teachers had implemented the curriculum for nearly two years at the time of interviews. One third-grade teacher claimed that their text only addressed seven (of 84) AIMS items. Another reported that a teacher challenged the policy of following the curriculum lock-step during a professional development session and their principal was brought into the meeting to ensure compliance.

Some teachers responded by supplementing the curriculum despite district mandates, reporting that it was their ethical duty to provide instruction on state standards. Others, including fifth-grade teachers, supplemented because they thought the text confused students, did not emphasize certain skills enough, or presented skills too late in the school year to ensure mastery in time for state testing. Twelve (of 17) third grade teachers reported creating their own assessments to address state standards. Others simply taught the text as written. One teacher shared that she did so in an attempt to prove to the district that this approach would harm AIMS performance.

With regard to the study objectives, a review of the district text revealed that one
portion of one lesson addressed the third-grade combinations skill and that, while word problems are predominant in the text (one chapter is devoted to word problems and subsequent chapters regularly include word problems), problems with either too little or too much information are not used. However, several activities require students to write their own word problem, which could result in students grappling with that concept. Two fifth-grade chapters are almost entirely devoted to graphing concepts and graphs are also regularly included in other chapters. However, three set Venn diagrams are absent entirely (although the teacher’s edition of the text recommends that teachers compare and contrast geometric shapes using a two set Venn diagram in one activity). The text does require students to practice functions on a near-weekly basis in its workbook, although the concept of functions is not addressed formally in a lesson until the very end of the book, where multi-step functions are introduced. Functions are apparently introduced in fourth grade and, therefore, students are expected to already understand this concept.

This finding is important for several reasons. First, between-grade level differences in AIMS may provide some support for the notion that the test is instructionally sensitive given differences in the curriculum used. Second, a grade level flag was added to the model to test for such differences. Adding this measure is also important because the AIMS test is different at third and fifth grade. Although it would be preferable to run analyses separately by grade level, the limited number of teachers interviewed suggests that overall mathematics performance should be examined across grade levels, in addition to one grade level at a time. This allows for comparison of the reliability of predicted scores for AIMS overall and for grade level models which have a
small number of classroom-level observations. It also helps to address the potential for underpowered grade level analyses by also including a model with a greater number of observations. Despite the limited sample size, testlet scores are run separately by grade level (and the Grade variable is not used) because they reflect performance on very distinct skills. Descriptive statistics are examined by grade level and in the aggregate to identify other between-grade level differences.

Descriptive Results

Descriptive statistics for all measures, aggregated across third and fifth grade, are presented in Table 8. In preparation for analysis, variables were converted to z-scores, or standardized. That is, the difference from the mean was computed and divided by the standard deviation of that measure. AIMS scores (AIMS, Thirdttest, and Fifhttest) are standardized relative to statewide performance at each grade level. Therefore, means are interpreted in terms of standard deviation units from the state average. All other measures are standardized relative to the study sample and are interpreted in terms of standard deviation units from the sample mean. For ease of interpretation, the descriptive statistics for dichotomous variables were calculated before being standardized. The same approach was used in calculating descriptive statistics in Table 9.

As is true for the district overall, the sample outperforms the state on the mathematics portion of the AIMS exam; scoring over a quarter of a standard deviation higher than the state on average ($M=0.29$, $SD=0.86$). There is less variation in scores than is experienced by the state as a whole, reflected by a standard deviation less than 1.0.
Table 8. Descriptive Statistics for Variables Included in HLM Analyses

<table>
<thead>
<tr>
<th>Measure</th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIMS</td>
<td>784</td>
<td>0.29</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thirdtest</td>
<td>319</td>
<td>0.30</td>
<td>0.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fifthtest</td>
<td>465</td>
<td>0.59</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-9</td>
<td>784</td>
<td>0.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minority*</td>
<td>784</td>
<td>0.00</td>
<td>1.00</td>
<td>0.30</td>
<td>--</td>
</tr>
<tr>
<td>ELL*</td>
<td>784</td>
<td>0.00</td>
<td>1.00</td>
<td>0.03</td>
<td>--</td>
</tr>
<tr>
<td>SPED*</td>
<td>784</td>
<td>0.00</td>
<td>1.00</td>
<td>0.09</td>
<td>--</td>
</tr>
<tr>
<td>Grade*</td>
<td>784</td>
<td>0.00</td>
<td>1.00</td>
<td>0.59</td>
<td>--</td>
</tr>
<tr>
<td>Freelunch</td>
<td>36</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Teacherqual</td>
<td>36</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Alignment</td>
<td>36</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Emphasis</td>
<td>36</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>AlignmentXEmphasis</td>
<td>36</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Outcome measures (AIMS, Thirdtest, and Fifthtest) are standardized relative to statewide performance. SAT-9, Freelunch, Teacherqual, Alignment, Emphasis, and AlignmentXEmphasis are standardized relative to the study sample.

* For dichotomous measures, the mean represents the proportion of students in the group of interest. For Grade, the mean represents the proportion of fifth-graders.
Table 9. Descriptive Statistics for Variables Included in HLM Analyses, by Grade Level

<table>
<thead>
<tr>
<th>Measure</th>
<th>Third Grade</th>
<th></th>
<th></th>
<th>Fifth Grade</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
<td>N</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>AIMS</td>
<td>319</td>
<td>0.44*</td>
<td>0.90</td>
<td>465</td>
<td>0.19*</td>
<td>0.82</td>
</tr>
<tr>
<td>Thirddtest</td>
<td>319</td>
<td>0.03</td>
<td>0.84</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Fifthtest</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>465</td>
<td>-0.23</td>
<td>0.59</td>
</tr>
<tr>
<td>SAT-9</td>
<td>319</td>
<td>0.02</td>
<td>1.02</td>
<td>465</td>
<td>-0.02</td>
<td>0.99</td>
</tr>
<tr>
<td>Minority*</td>
<td>319</td>
<td>0.32</td>
<td>--</td>
<td>465</td>
<td>0.29</td>
<td>0.45</td>
</tr>
<tr>
<td>ELL*</td>
<td>319</td>
<td>0.03</td>
<td>--</td>
<td>465</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>SPED*</td>
<td>319</td>
<td>0.11</td>
<td>--</td>
<td>465</td>
<td>0.08</td>
<td>0.27</td>
</tr>
<tr>
<td>Freelunch</td>
<td>16</td>
<td>-0.07</td>
<td>1.05</td>
<td>20</td>
<td>0.06</td>
<td>0.98</td>
</tr>
<tr>
<td>Teacherqual</td>
<td>16</td>
<td>0.05</td>
<td>1.10</td>
<td>20</td>
<td>-0.04</td>
<td>0.94</td>
</tr>
<tr>
<td>Alignment</td>
<td>16</td>
<td>0.07</td>
<td>1.00</td>
<td>20</td>
<td>-0.06</td>
<td>1.02</td>
</tr>
<tr>
<td>Emphasis</td>
<td>16</td>
<td>-0.43*</td>
<td>1.20</td>
<td>20</td>
<td>0.35*</td>
<td>0.65</td>
</tr>
<tr>
<td>AlignmentXEmphasis</td>
<td>16</td>
<td>0.14</td>
<td>1.30</td>
<td>20</td>
<td>-0.11</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Note: Outcome measures (AIMS, Thirddtest, and Fifthtest) are standardized relative to statewide performance. SAT-9, Freelunch, Teacherqual, Alignment, Emphasis, and AlignmentXEmphasis are standardized relative to the study sample.

For dichotomous measures, the mean represents the proportion of students in the group of interest.

Indicates statistically significant difference between grade levels ($p<0.05$).
However, some students in the sample do score at the extremes, more than two standard deviations above and below the state mean.

Performance on the five-item third-grade testlet (Thirdtest) and eight-item fifth-grade testlet (Fifthtest) is quite different than AIMS scores overall. Third-grade performance was quite close to the state average ($M=0.03, SD=0.84$) while fifth-graders scored considerably worse than the state overall ($M=-0.23, SD=0.59$). There was also much less variation in Fifthtest scores than in Thirddtest scores. As reported in Table 7, testlet scores are not very reliable. For the third grade testlet, just under half of the variation in scores can be attributed to true score differences, or students’ true understanding of the performance objectives (Coefficient alpha=0.49). At fifth grade, just over two thirds of the variation in scores can be attributed to true differences in student understanding (Coefficient alpha=0.68). Therefore, analysis of overall AIMS performance is much more meaningful than testlet results.

Since SAT-9 scores were standardized relative to the study sample, the mean and standard deviation are meaningless when examined alone ($M=0.00, SD=1.00$). Taken in conjunction with a maximum score of 1.95 and minimum score of -2.97, however, the statistics provide more information. To achieve a mean of 0.00 with a minimum score much smaller than the maximum score, the sample must have more students performing above the mean than below it. Additional analyses confirmed this assumption; a histogram of SAT-9 scores revealed a slightly negative skew, $Mdn=0.07$ and mode=1.05. Given the high-achieving nature of the district, this is unavoidable. Standardizing scores does not change the shape of their distribution, and students in this district outperform
state and national norms on this assessment.

The study sample is somewhat consistent with the district in terms of student demographics; 70% of the sample is Caucasian and 3% are English language learners. However, substantially fewer students have disabilities (9% in the sample and 22% in the district), as shown in Table 2. Still, it is not the goal of this study to draw inferences about the district from the study sample. Therefore, statistical comparisons between district and sample characteristics are not needed.

The distributions of classroom-level measures were first examined by reviewing histograms for each measure. This analysis seems to indicate that the proportion of students eligible for free and reduced lunch is positively skewed. Frequencies confirm this interpretation. Eleven of the 36 cases scored approximately one standard deviation below the district mean on this measure, seven scored approximately half a standard deviation below the mean, four scored approximately one standard deviation above the mean, four approximately one and a half standard deviations above the mean, and two scored approximately two standard deviations above the mean. This indicates that teachers in the district’s poorer schools participated in the study to a greater extent than teachers in more wealthy schools. Other classroom-level measures appear to approximate a normal distribution.

Third- and fifth-grade means on all measures are presented in Table 9. Only two statistically significant differences exist; third grade students ($M=0.44, SD=0.90$) performed better than fifth-graders ($M=0.19, SD=0.82$) on AIMS, $t(782)=4.132, p<0.01, d=0.29$ and fifth grade teachers ($M=0.35, SD=0.65$) emphasized the study objectives
more than third grade teachers ($M=-0.43$, $SD=1.20$), Welch’s $t(27) =-2.34$, $p<0.05$, $g=-0.81$. No other statistically significant differences were detected; third grade students performed similarly to fifth grade students on the SAT-9, the study measure of prior math achievement, and were similar in terms of their demographic characteristics. This also holds true for the classroom-level measures, the teacher quality and free lunch covariates and the instructional validity measures (alignment, emphasis, or alignmentXemphasis). However, the number of classroom level observations is quite small.

The study also compared the distribution of instructional validity measures. Figure 5 presents the distribution of alignment codes by grade level, Figure 6 displays emphasis scores. Third and fifth grade appear to be somewhat similar in terms of their alignment ratings in Figure 5. Both have two classrooms with ratings below “some alignment” and no classrooms with ratings of “not aligned.” The majority of classrooms score between “some alignment” and “close alignment,” with most scores in the middle or high end of this range. At fifth grade, one teacher received a score between “close alignment” and “perfect alignment” while no third grade teachers had similar scores. Student’s $t$-tests also failed to reject the null hypothesis that the mean alignment ratings for each grade level are equal, $t(34)=0.40$, $p=0.69$, indicating that alignment ratings were similar across grade levels.

However, differences appear across grade levels in emphasis ratings, as presented in Figure 6. Fifth-grade teachers cover the study objectives more frequently and third grade teachers vary more widely in their emphasis ratings. Fifteen out of 20 fifth-grade teachers taught the study objectives at least weekly and five did so for five or ten days a
Figure 5. Distribution of Classroom Alignment Ratings, by Grade

Figure 6. Distribution of Classroom Emphasis Ratings, by Grade
year. In contrast, five out of 16 third-grade teachers addressed the objectives at least weekly, with the remaining teachers doing so for ten days a year or less. As reported earlier, a Welch’s $t$-test also revealed a statistically significantly difference in grade level mean emphasis scores, $t(19.15) = -3.87, p<0.01, d = -1.77$. Since third-grade teachers reported that study objectives are completely omitted from the district text, it is not surprising that fifth-grade teachers would emphasize them more. To teach the objectives at all took a great deal of preparation at third grade, whereas materials were readily available at fifth grade.

**Intercorrelations**

Intercorrelations are important to this study for two reasons. First, they reveal the degree of relationship between dependent and independent variables before controlling for other measures. Second, they help to determine if independent variables are so intercorrelated as to distort parameter estimates.

Intercorrelations are presented in Tables 10a and 10b. The greatest relationship between any of the study variables occurred among test scores. Overall AIMS mathematics performance is positively correlated with Thirdtest, $r(319) = 0.71, p<0.01$, and Fifthtest, $r(465) = 0.79, p<0.01$, both comprised of items from AIMS. It is also positively correlated with prior performance on the SAT-9, $r(784) = 0.77, p<0.01$. The magnitude of the correlations between AIMS and other assessments is large and quite similar. In contrast, SAT-9 performance is more highly correlated with overall AIMS performance than with Thirdtest, $r(319) = 0.55, p<0.01$, or Fifthtest, $r(465) = 0.62, p<0.01$. 


<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. AIMS (n=784)</td>
<td>1.00</td>
<td>0.71*</td>
<td>0.79*</td>
<td>0.77*</td>
<td>-0.23*</td>
<td>-0.14*</td>
<td>-0.31*</td>
<td>-0.15*</td>
</tr>
<tr>
<td>2. Thirdtest (n=319)</td>
<td>1.00</td>
<td>--</td>
<td>0.55*</td>
<td>-0.10</td>
<td>-0.09</td>
<td>-0.19*</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>3. Fifthtest (n=465)</td>
<td>1.00</td>
<td>0.62*</td>
<td>-0.18*</td>
<td>-0.16*</td>
<td>-0.29*</td>
<td>--</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. SAT-9 (n=784)</td>
<td>1.00</td>
<td>-0.23*</td>
<td>-0.18*</td>
<td>-0.30*</td>
<td>-0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Minority (n=784)</td>
<td>1.00</td>
<td>0.19*</td>
<td>0.05</td>
<td>-0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. ELL (n=784)</td>
<td>1.00</td>
<td>0.01</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. SPED (n=784)</td>
<td>1.00</td>
<td>-0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Grade (n=784)</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* indicates statistically significant relationship at \( p < 0.05 \).

Table 10b. Zero-order Correlations Among Variables Used in Analysis (Continued)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. AIMS (n=784)</td>
<td>1.00</td>
<td>0.71*</td>
<td>0.79*</td>
<td>-0.26*</td>
<td>0.15*</td>
<td>0.09*</td>
<td>0.11*</td>
<td>-0.09*</td>
</tr>
<tr>
<td>2. Thirdtest (n=319)</td>
<td>1.00</td>
<td>--</td>
<td>-0.23*</td>
<td>0.10*</td>
<td>0.25*</td>
<td>0.13*</td>
<td>-0.26*</td>
<td></td>
</tr>
<tr>
<td>3. Fifthtest (n=465)</td>
<td>1.00</td>
<td>-0.17*</td>
<td>0.17*</td>
<td>-0.02</td>
<td>0.17*</td>
<td>0.12*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Freelunch (n=36)</td>
<td>1.00</td>
<td>-0.14</td>
<td>0.14</td>
<td>-0.24</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Teacherqual (n=36)</td>
<td>1.00</td>
<td>-0.02</td>
<td>0.03</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Alignment (n=36)</td>
<td>1.00</td>
<td>0.20</td>
<td>-0.42*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Emphasis (n=36)</td>
<td>1.00</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. AlignmentXEmphasis (n=36)</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* indicates statistically significant relationship at \( p < 0.05 \).
It is difficult to draw conclusions based on these results because the magnitude of a correlation between two measures is attenuated when the measures are not very reliable, as is the case with the Thirdtest and Fifthtest. Therefore, correlation coefficients involving these measures do not reveal much. Lengthening these tests would likely increase their reliability; Spearman Brown Prophecy formula estimates that an 80-item version of Thirdtest (comparable to the length of the full AIMS test) would have a reliability coefficient of 0.94 and an 80-item version of Fifthtest would have a reliability coefficient of 0.96 (Nunnally & Bernstein, 1994).

Although many of the other zero order correlations are statistically significant at the $p<0.05$ level, the magnitude of $r$ is small, suggesting that differences were found due to the large sample sizes involved. There are a few exceptions to this finding. Moderate (and negative) correlations exist between special education status and overall AIMS performance, $r(784)=-0.31$, $p<0.01$, and between special education status and SAT-9 performance, $r(784)=-0.30$, $p<0.01$. Alignment and the interaction between alignment and emphasis are also moderately and negatively correlated, $r(36)=-0.42$, $p=0.01$. All other measures had small intercorrelations or were not correlated.

These results revealed no problems with multicollinearity. In addition, the small number of sizeable correlations makes it doubtful that a statistically significant relationship will be found between the instructional validity measures and mathematics achievement. However, the correlational analyses apply classroom-level instructional validity scores to each student in a classroom. Since the enrollment in each classroom varied, analyses are weighted by class size with larger classes having a greater impact on
findings. HLM corrects for this by weighting its estimates.

Multilevel Analyses

Unconditional Model

The multilevel analyses presented here partitioned the variance in test scores between- and within-classrooms and determined how much of the between-classroom variation could be accounted for by the three instructional validity measures: alignment, emphasis, and the alignmentXemphasis interaction. In order for analyses to yield useful results, performance must vary between classrooms. That is, if student achievement is the same in every classroom it will not be possible to identify the effect of any classroom-level measure.

Therefore, the first step taken was to run an unconditional model to estimate the amount of between-classroom variation. This analysis is visually represented in Figures 7a-d, which present the 95% confidence interval around each classroom mean for the outcome measures. AIMS scores are presented separately for third and fifth grade.

These figures reveal relatively little between-classroom variation in AIMS, Thirdtest, and Fifthtest; most vertical lines representing each classroom overlap with many other lines. They also show that between-classroom variation in AIMS scores was greater for third grade than fifth grade; the lines representing each classroom are more spread out. The pattern is quite similar for Thirdtest and Fifthtest. Classrooms also maintained similar positions on the testlets as on AIMS, which is expected given the strong correlations between AIMS, Thirdtest, and Fifthtest. Finally, the variation within
95% Confidence Interval around Mean Classroom AIMS, Grade 3

95% Confidence Interval around Mean Classroom AIMS, Grade 5

95% Confidence Interval around Mean Classroom Thirdtest

95% Confidence Interval around Mean Classroom Fifthtest
classrooms appear greater at third grade than fifth-grade; the lines are longer. Therefore, third grade teachers taught more heterogeneous classrooms than fifth-grade teachers. Within-classroom variation also appears smaller on Fifthtest than on fifth-grade AIMS, and larger on Thirdtest than third-grade AIMS.

The quantitative results from the unconditional model are presented in Tables 12-16. Since the outcome measures are all expressed on the same scale, it is possible to compare the amount of variation in scores across HLM models. The total amount of variation in third-grade AIMS scores ($s^2=0.82$), all AIMS scores ($s^2=0.76$), in Thirdtest ($s^2=0.73$) scores, and in fifth-grade AIMS scores ($s^2=0.68$), before introducing any independent variables, was nearly twice as large as the variation in Fifthtest scores ($s^2=0.38$). In addition, the proportion of variation between classrooms was greatest for third-grade AIMS scores (18.3% of the variance lays between classrooms), followed by all AIMS scores (14.5%), Thirdtest (11.0%), fifth-grade AIMS scores (7.3%), and Fifthtest scores (7.2%).

Finally, variation in third- and fifth-grade scores were compared using $F$ tests for two population variances (Sheskin, 2004) for the study sample, districtwide and across the state of Arizona, with variance estimates taken from Tables 1, 12, and 13. The ratio of third to fifth grade AIMS variance did not reveal statistically significant differences between grade levels for the study sample, $F(464, 318)=1.21, p=0.07$, or for the district overall, $F(975, 987)=1.03, p=0.70$. However, at the state level differences were found, $F(76,718, 77,442)=1.16, p<0.01$. Although two-tailed tests were run, and the directionality of the relationship was not tested, descriptive results show that the variance
in fifth-grade scores statewide ($s^2=2948.49$) exceeded the variance in third-grade scores ($s^2=2530.09$).

**Covariate Analysis**

After examining the results from the unconditional model, student-level covariates were introduced and the amount of variation in scores reexamined, followed by classroom-level covariates. Covariates were retained in the model only if they were significant predictors of mathematics achievement.

The student-level covariates reduced the between-classroom variance in AIMS scores across grade levels by 69.0%, indicating that student ethnicity, special education status, grade level and 2004 SAT-9 performance accounted for over two-thirds of the initial differences between classrooms in 2005 AIMS math scores. Percentage of students eligible for free lunch was retained as a classroom-level covariate, and reduced the between-classroom variance an additional 11.7% more than the model including the Level-1 covariates. All covariates account for 72.6% of the initial between-classroom differences in math achievement. After entering the covariates, 10.4% of the total variance remained between classrooms. Although a sizeable decrease in between-classroom variance, chi square analyses revealed that a statistically significant amount of variation between classrooms remained to be modeled, $\chi^2(34, N=784)=112.14$, $p<0.01$ (Table 11).

For third-grade students, the student-level covariate SAT-9 reduced the between-classroom variance in AIMS scores by 73.3%. Percentage students eligible for free lunch
Table 11. Variation in AIMS Scores Explained by the HLM Models, Third- and Fifth-grade Combined

<table>
<thead>
<tr>
<th></th>
<th>Unconditional Model</th>
<th>Model with Student-level Covariates</th>
<th>Model with Classroom-level Covariates</th>
<th>Final Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance between classrooms, $u_{0j}$</td>
<td>0.11</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Variance within classrooms, $r_{ij}$</td>
<td>0.65</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>Total variance</td>
<td>0.76</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>14.5%</td>
<td>11.7%</td>
<td>10.4%</td>
<td>10.4%</td>
</tr>
<tr>
<td>Reliability of intercept</td>
<td>0.78</td>
<td>0.74</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>Percentage of between-classroom variance explained by the model</td>
<td>69.0%</td>
<td>72.6%</td>
<td>72.6%</td>
<td></td>
</tr>
<tr>
<td>$chi square$</td>
<td>155.43</td>
<td>126.94</td>
<td>112.14</td>
<td>112.14</td>
</tr>
<tr>
<td>$df$</td>
<td>35</td>
<td>35</td>
<td>34</td>
<td>34</td>
</tr>
</tbody>
</table>

Note: Final model figures match the model with classroom-level covariates because no instructional validity measures were retained; they were not statistically significant predictors of math achievement (at $p<0.05$) after controlling for student demographics and classroom characteristics.

Table 12. Variation in Third-grade AIMS Scores Explained by the HLM Models

<table>
<thead>
<tr>
<th></th>
<th>Unconditional Model</th>
<th>Model with Student-level Covariates</th>
<th>Model with Classroom-level Covariates</th>
<th>Final Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance between classrooms, $u_{0j}$</td>
<td>0.15</td>
<td>0.04</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Variance within classrooms, $r_{ij}$</td>
<td>0.67</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>Total variance</td>
<td>0.82</td>
<td>0.32</td>
<td>0.31</td>
<td>0.30</td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>18.3%</td>
<td>12.5%</td>
<td>10.2%</td>
<td>8.2%</td>
</tr>
<tr>
<td>Reliability of intercept</td>
<td>0.82</td>
<td>0.73</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>Percentage of between-classroom variance explained by the model</td>
<td>73.3%</td>
<td>79.1%</td>
<td>83.5%</td>
<td></td>
</tr>
<tr>
<td>$chi square$</td>
<td>82.45</td>
<td>54.57</td>
<td>45.19</td>
<td>40.22</td>
</tr>
<tr>
<td>$df$</td>
<td>15</td>
<td>15</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>
Table 13. Variation in Fifth-grade AIMS Scores Explained by the HLM Models

<table>
<thead>
<tr>
<th></th>
<th>Unconditional Model</th>
<th>Model with Student-level Covariates</th>
<th>Model with Classroom-level Covariates</th>
<th>Final Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance between classrooms, $u_{ij}$</td>
<td>0.05</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Variance within classrooms, $r_{ij}$</td>
<td>0.63</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>Total variance</td>
<td>0.68</td>
<td>0.28</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>7.3%</td>
<td>11.9%</td>
<td>9.6%</td>
<td>9.6%</td>
</tr>
<tr>
<td>Reliability of intercept</td>
<td>0.64</td>
<td>0.75</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Percentage of between-classroom variance explained by the model</td>
<td>33.9%</td>
<td>47.8%</td>
<td>47.8%</td>
<td>47.8%</td>
</tr>
<tr>
<td>$chi^2$</td>
<td>52.21</td>
<td>74.07</td>
<td>61.77</td>
<td>61.77</td>
</tr>
<tr>
<td>$df$</td>
<td>19</td>
<td>19</td>
<td>18</td>
<td>18</td>
</tr>
</tbody>
</table>

Note: Final model figures match the model with classroom-level covariates because no instructional validity measures were retained; they were not statistically significant predictors of math achievement (at $p<0.05$) after controlling for student demographics and classroom characteristics.

Table 14. Variation in Thirdtest Scores Explained by the HLM Models

<table>
<thead>
<tr>
<th></th>
<th>Unconditional Model</th>
<th>Model with Student-level Covariates</th>
<th>Model with Classroom-level Covariates</th>
<th>Final Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance between classrooms, $u_{ij}$</td>
<td>0.08</td>
<td>0.03</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Variance within classrooms, $r_{ij}$</td>
<td>0.64</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>Total variance</td>
<td>0.71</td>
<td>0.50</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>11.0%</td>
<td>5.4%</td>
<td>5.4%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Reliability of intercept</td>
<td>0.73</td>
<td>0.53</td>
<td>0.53</td>
<td>0.12</td>
</tr>
<tr>
<td>Percentage of between-classroom variance explained by the model</td>
<td>65.8%</td>
<td>65.8%</td>
<td>95.9%</td>
<td>95.9%</td>
</tr>
<tr>
<td>$chi^2$</td>
<td>55.14</td>
<td>31.90</td>
<td>31.90</td>
<td>16.67</td>
</tr>
<tr>
<td>$df$</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>14</td>
</tr>
</tbody>
</table>
Table 15. Variation in Fifthtest Scores Explained by the HLM Models

<table>
<thead>
<tr>
<th></th>
<th>Unconditional Model</th>
<th>Model with Student-level Covariates</th>
<th>Model with Classroom-level Covariates</th>
<th>Final Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variance between classrooms, $u_{ij}$</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Variance within classrooms, $r_{ij}$</td>
<td>0.35</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>Total variance</td>
<td>0.38</td>
<td>0.22</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>Intraclass correlation</td>
<td>7.2%</td>
<td>7.1%</td>
<td>4.2%</td>
<td>4.2%</td>
</tr>
<tr>
<td>Reliability of intercept</td>
<td>0.66</td>
<td>0.63</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>Percentage of between-classroom variance explained by the model</td>
<td>43.6%</td>
<td>67.8%</td>
<td>67.8%</td>
<td></td>
</tr>
<tr>
<td>chi square</td>
<td>54.46</td>
<td>50.43</td>
<td>36.35</td>
<td>36.35</td>
</tr>
<tr>
<td>df</td>
<td>19</td>
<td>19</td>
<td>18</td>
<td>18</td>
</tr>
</tbody>
</table>

Note: Final model figures match the model with classroom-level covariates because no instructional validity measures were retained; they were not statistically significant predictors of math achievement (at $p<0.05$) after controlling for student demographics and classroom characteristics.
was retained as a classroom-level covariate, and reduced the between-classroom variance an additional 21.5% more than the model including the Level-1 covariates. All covariates account for 79.1% of the initial between-classroom variance. After entering SAT-9 and free lunch eligibility, 10.2% of the total variance remained between classrooms. A statistically significant amount of variation between classrooms remained to be modeled, $\chi^2(14, N=316)= 45.19, p<0.01$ (Table 12).

For fifth-grade students, the student-level covariates SAT-9, SPED, and Minority reduced the between-classroom variance in AIMS scores by 33.9%. Teacher qualifications was retained as a classroom-level covariate, and reduced the between-classroom variance an additional 21.0% more than the model including the Level-1 covariates. All covariates account for 47.8% of the initial between-classroom variance. After entering the covariates, 9.6% of the total variance remained between classrooms. A statistically significant amount of variation between classrooms remained to be modeled, $\chi^2(18, N=460)= 61.77, p<0.01$ (Table 13).

For Thirdtest, the student-level covariate SAT-9 reduced the between-classroom variance by 65.8%. No classroom-level covariates significantly predicted Thirdtest achievement and therefore they were all omitted from the model. After entering SAT-9, 5.4% of the total variance remained between classrooms. A statistically significant amount of variation between classrooms remained to be modeled, $\chi^2(15, N=319)=31.90, p<0.01$ (Table 14).

The SAT-9 and SPED Student-level covariates reduced the between-classroom variance in Fifthtest scores by 43.6%. The teacher quality measure (Tqual) increased the
proportion of between-classroom variation explained by 42.9% more than the model including the Level-1 covariates. All covariates account for 67.8% of the initial between-classroom differences in math achievement. After entering the covariates, 4.2% of the total variance remained between classrooms. As with the other models, chi square analyses revealed that a statistically significant amount of variation between classrooms remained, $\chi^2(18, N=465)=36.35, p<0.01$ (Table 15).

Final Model

The final step in analysis involved introducing the three instructional validity measures: alignment, emphasis, and alignmentXemphasis into the model. As with the covariates, only statistically significant instructional validity measures were retained. This increased the precision of the parameter estimates for these variables, the key measures of interest in the study, and guarded against overspecification (an admittedly unlikely problem given the low intercorrelations among instructional validity measures). In addition, the relatively small number of classrooms limits the number of classroom-level predictors that can be included, and the measure with the greatest relationship would be selected. If the interaction term were significant (at $p<0.05$), all three instructional validity measures would have been retained. However, neither approach proved necessary.

Parameter estimates for the final HLM models are presented in Tables 16-20. The instructional validity measures did not significantly predict mathematics performance except in two cases: alignment is associated with third-grade AIMS performance and
Table 16. Final Conditional HLM 2-Level Model Predicting Variation in AIMS, Third- and Fifth-grade Combined

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>coefficient</th>
<th>se</th>
<th>df</th>
<th>t ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model for mean classroom math achievement, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{00}$</td>
<td>0.29</td>
<td>0.03</td>
<td>34</td>
<td>8.71*</td>
</tr>
<tr>
<td>Freeland, $\gamma_{01}$</td>
<td>-0.07</td>
<td>0.03</td>
<td>34</td>
<td>-2.59*</td>
</tr>
<tr>
<td>Model for SAT9, $\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{10}$</td>
<td>0.60</td>
<td>0.03</td>
<td>778</td>
<td>21.25*</td>
</tr>
<tr>
<td>Model for Minority, $\beta_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{20}$</td>
<td>-0.12</td>
<td>0.04</td>
<td>778</td>
<td>-3.19*</td>
</tr>
<tr>
<td>Model for SPED, $\beta_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{40}$</td>
<td>-0.30</td>
<td>0.11</td>
<td>778</td>
<td>-2.75*</td>
</tr>
<tr>
<td>Model for Grade, $\beta_5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{50}$</td>
<td>-0.25</td>
<td>0.06</td>
<td>778</td>
<td>-3.86*</td>
</tr>
</tbody>
</table>

* indicates statistically significant relationship at $p<0.05$.

Table 17. Final Conditional HLM 2-Level Model Predicting Variation in Third Grade AIMS

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>coefficient</th>
<th>se</th>
<th>df</th>
<th>t ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model for mean classroom math achievement, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{00}$</td>
<td>0.45</td>
<td>0.04</td>
<td>13</td>
<td>10.12*</td>
</tr>
<tr>
<td>Freeland, $\gamma_{01}$</td>
<td>-0.07</td>
<td>0.04</td>
<td>13</td>
<td>-1.68</td>
</tr>
<tr>
<td>Alignment, $\gamma_{03}$</td>
<td>0.10</td>
<td>0.04</td>
<td>13</td>
<td>2.40*</td>
</tr>
<tr>
<td>Model for SAT9, $\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{10}$</td>
<td>0.68</td>
<td>0.04</td>
<td>315</td>
<td>17.01*</td>
</tr>
</tbody>
</table>

* indicates statistically significant relationship at $p<0.05$.

Table 18. Final Conditional HLM 2-Level Model Predicting Variation in Fifth Grade AIMS

<table>
<thead>
<tr>
<th>Fixed Effect</th>
<th>coefficient</th>
<th>se</th>
<th>df</th>
<th>t ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model for mean classroom math achievement, $\beta_0$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{00}$</td>
<td>0.18</td>
<td>0.04</td>
<td>18</td>
<td>4.34*</td>
</tr>
<tr>
<td>Tqual, $\gamma_{02}$</td>
<td>0.10</td>
<td>0.04</td>
<td>18</td>
<td>2.34*</td>
</tr>
<tr>
<td>Model for SAT9, $\beta_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{10}$</td>
<td>0.57</td>
<td>0.04</td>
<td>460</td>
<td>13.94*</td>
</tr>
<tr>
<td>Model for Minority, $\beta_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{20}$</td>
<td>-0.20</td>
<td>0.04</td>
<td>460</td>
<td>-4.63*</td>
</tr>
<tr>
<td>Model for SPED, $\beta_4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept, $\gamma_{40}$</td>
<td>-0.39</td>
<td>0.16</td>
<td>460</td>
<td>-2.48</td>
</tr>
</tbody>
</table>

* indicates statistically significant relationship at $p<0.05$. 
<table>
<thead>
<tr>
<th>Table 19. Final Conditional HLM 2-Level Model Predicting Variation in Thirdtest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effect</td>
</tr>
<tr>
<td>Model for mean classroom math achievement, $\beta_0$</td>
</tr>
<tr>
<td>Intercept, $\gamma_{00}$</td>
</tr>
<tr>
<td>Alignment, $\gamma_{03}$</td>
</tr>
<tr>
<td>Model for SAT9, $\beta_1$</td>
</tr>
<tr>
<td>Intercept, $\gamma_{10}$</td>
</tr>
<tr>
<td>* indicates statistically significant relationship at $p&lt;0.05$.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 20. Final Conditional HLM 2-Level Model Predicting Variation in Fifthtest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effect</td>
</tr>
<tr>
<td>Model for mean classroom math achievement, $\beta_0$</td>
</tr>
<tr>
<td>Intercept, $\gamma_{00}$</td>
</tr>
<tr>
<td>Tqual, $\gamma_{02}$</td>
</tr>
<tr>
<td>Model for SAT9, $\beta_1$</td>
</tr>
<tr>
<td>Intercept, $\gamma_{10}$</td>
</tr>
<tr>
<td>Model for SPED, $\beta_4$</td>
</tr>
<tr>
<td>Intercept, $\gamma_{40}$</td>
</tr>
<tr>
<td>* indicates statistically significant relationship at $p&lt;0.05$.</td>
</tr>
</tbody>
</table>
with Thirdtest performance. This finding is somewhat suspect for Thirdtest; the reliability of the intercept decreases from $\alpha=0.53$ to $\alpha=0.12$ when alignment is added to the model, indicating the accuracy of predicted classroom means has been jeopardized. Also, the reduction in between-classroom variance, from $u_{0j}=0.03$ with all covariates included to $u_{0j}=0.00$ when alignment is introduced, indicates that there are not enough level-two cases (classrooms) to support this model (Table 14). The direction of the relationship between all significant independent variables is discussed next.

For the model predicting AIMS achievement for both grade levels (Table 16), the classroom-level free lunch measure, and student-level minority status, special education status, grade level, and 2004 SAT-9 achievement are all significant predictors of 2005 AIMS performance ($p<0.05$). All relationships are in the expected direction; attending higher-poverty schools is associated with lower AIMS scores, as is being a minority, a disabled student, or a fifth-grader (recall that third grade students performed better overall, relative to the state, than fifth-graders in this district). Prior performance on the SAT-9 is positively correlated with AIMS achievement.

Third-grade AIMS achievement was associated with two independent variables, 2004 SAT-9 achievement and instructional alignment with AIMS (Table 17). Percent of students eligible for free lunch significantly predicted AIMS performance when only covariates were entered into the model. Once alignment was introduced, the free lunch measure was no longer statistically significant. However, it was retained in the full model because of the results from the covariate analysis. All relationships are in the expected direction; performing well on the SAT-9 is related to high performance on AIMS, as is
receiving instruction that is well-aligned with AIMS.

Fifth-grade performance on AIMS is associated with the following measures: the teacher quality composite, 2004 SAT-9 performance, minority status, and special education status. Fifth-grade students whose teachers had a greater combination of experience and education were more likely to perform well on AIMS, as were students with high scores on the SAT-9, while disabled student status and being a minority is negatively correlated with fifth-grade AIMS scores (Table 18).

Only one covariate significantly predicted Thirdtest performance: 2004 SAT-9 scores. As with the AIMS analysis, students with high SAT-9 scores are expected to also perform well on Thirdtest. As mentioned earlier, alignment was also a significant predictor of Thirdtest performance (as reflected in Table 19), but including this measure in the model seems to seriously jeopardize its accuracy.

Fifthtest performance is associated with the following measures: the teacher quality composite, 2004 SAT-9 performance, and special education status. Students whose teachers had a greater combination of experience and education were more likely to perform well on Fifthtest, as were students with high scores on the SAT-9, while disabled student status is negatively correlated with Fifthtest scores (Table 20).

Finally, because alignment significantly predicted third-grade AIMS scores, but not fifth-grade scores, two classroom-level variables, grade and the interaction of gradexalignment, were added to the model (excluding the student-level grade measure). While grade significantly predicted AIMS performance (and had a similar parameter estimate, standard error, and t ratio as the student-level variable), alignment and the
Gradexalignment interaction were not significantly associated with AIMS performance ($t_{31}=0.19$ and $t_{31}=0.25$, respectively).

Despite this finding, and because alignment significantly predicted AIMS performance at third grade, the approaches of third-grade teachers with high alignment scores and high-performing classrooms were compared with those experiencing moderate and low levels of alignment and with moderate- and low-performing classrooms. Teachers in the high-performing group reported using the district-adopted curriculum, but supplementing to address state standards. Teachers with low levels of alignment solely taught the district text and did so exactly as written.

For example, one highly aligned, high-performing teacher reported addressing the study objectives during morning “bellwork” to ensure that her students were exposed to state standards not addressed by the district curriculum. The examples of problems that she shared indicated she exposed students to multiple ways to approach a problem instead of presenting one or two solutions. She accomplished this by working with her grade level team, “I use a curriculum map and what we’re supposed to cover and I go through my curriculum guide and make sure I’ve covered everything and we went through the Everyday Math book and lined it up with the curriculum guide for third grade and [the grade-level team] wrote down all the things that were missing and just made sure that we covered those.”

One moderately aligned, moderate-performing teacher reported using the AIMS study guide to identify concepts that she needs to address that either aren’t included in the district text or won’t be addressed by the time AIMS is administered. She identified the
combinations objective as one such skill. To address it, she had students participate in hands-on activities making combinations of things (making ice cream sundaes, putting on different outfits, etc.), but seems to have focused more on the activity than on the skill of diagramming. Only one approach to diagramming was used, and it was consistent with an old version of AIMS not with current AIMS items. With regard to the second objective, she taught students approaches to solving word problems (such as underlining needed information and crossing out unneeded information), but focused on whether or not students solved the problem correctly in her assessments instead of on student facility with identifying necessary and unnecessary information.

In contrast, one of the not aligned, low-performing teachers reported teaching the district test exactly as written because of district requirements. She indicated that she did so to prove the district wrong; she’d teach the text just as written to show that it would result in poor AIMS performance. As a result, she told me that she did not address one objective at all and devoted little time (30 minutes per year) to the second objective, focusing on only one aspect of the skill (unnecessary information, ignoring necessary information).
CHAPTER 5
DISCUSSION

This dissertation discusses the evolution of thinking about instructional validity and defines instructional validity within the context of standards-based assessments. It argues that traditional measures of standards-based instruction suffer from construct-underrepresentation and presents an interview-based approach to gauging this concept. The method, applied in the current study to investigate how two mathematics performance objectives are taught in each of two grade levels, is described in detail. Results from analyses of the relationship between three measures of instructional validity and mathematics performance are presented. Between-grade level differences are also examined due to substantial variation in the alignment between the curriculum and the assessment in the grade levels of interest.

Measuring instructional alignment

Study results lead to several conclusions about generating alignment ratings from interview data. First, although interviews yield rich descriptions of instruction, it is difficult to generate alignment ratings with a high degree of inter-rater reliability. Several factors may affect the consistency of ratings. It can be difficult to code large volumes of information, such as that presented when discussing an objective that is taught over a period of weeks and in a variety of venues—formal mathematics lessons, computer labs, other content area lessons, test preparation activities, and independent practice during
early morning “bellwork.” Relevant information can be easily overlooked and balancing different pieces of information requires many judgment calls. Second, the complexity of objectives may affect coding. Despite using the skills assessed item-by-item to judge how alignment is measured, the large number of skills required to correctly answer each test item and addressed in instruction make rating alignment using interview transcripts a cognitively demanding task. Finally, although collecting assessment examples provided valuable information to the study and helped teachers to better explain how they operationalized each objective, their addition created even more information to process in coding and may have affected inter-rater reliability.

These challenges necessitate joint coding of all transcripts. Using only two coders, the process applied in this study, may not be sufficient to study the reliability of this coding procedure. However, determining the optimal number of scorers for cognitively complex coding is beyond the scope of this study. It may be an appropriate research topic if the study’s approach to quantifying instruction is more widely implemented. Another line of research might involve investigating the consistency of codes generated using a combination of assessment examples and interview transcripts, using transcripts independently of assessments, and only examining assessment examples. It would be beneficial, both in terms of coding time and in complexity of analysis, to use less information if the scores generated do not vary.

Differences between grade levels

Several factors lead the study to predict that findings would differ by grade level,
and indeed, grade is a significant predictor of AIMS performance. Discrepancies exist between third and fifth grade in the curricular alignment with AIMS, in the degree to which teachers participate in test preparation activities, in the amount of change in AIMS items between 2004 and 2005, and in the AIMS performance of all students in the district relative to the state. However, third and fifth grade performance on a nationally normed test of mathematics achievement in the previous year was remarkably similar, indicating that students entered their respective grade levels with comparable levels of background knowledge.

For the objectives of interest to this study, the degree of curricular alignment with AIMS varied greatly between grade levels. This might have resulted in differences in instructional alignment ratings, especially since the district requires teachers to teach the district-adopted text exactly as written. However, alignment ratings did not vary much. This speaks to teachers’ commitment to teach the state standards, especially at third grade where the greatest gaps exist.

In contrast, the amount of emphasis placed on objectives differed between grade levels, with third-grade teachers emphasizing the study objective less than fifth-grade teachers. Two factors may affect this finding. First, the fifth grade objectives may be considered “power standards” (Reeves, 2002) or skills central to success in fifth grade mathematics while the third grade objectives were much more tangential. This might cause fifth-grade teachers to devote a larger portion of instructional time to these skills. Second, since third-grade teachers had to look outside of the curriculum to teach the study objectives, it took greater effort to teach them. With fewer instructional materials
readily available, it would be difficult for third-grade teachers to address these skills with
great frequency.

There are also between-grade level differences in the test itself. More fifth-grade
items relate to study objectives (eight items at fifth grade, five at third grade), while third-
graders outperformed fifth-graders relative to the rest of the state in terms of overall
mathematics performance and on items gauging performance on the objectives of interest
to the study. Finally, test scores varied to a greater extent between-classrooms at third
grade than at fifth grade. Analyses of the amount of variation between classrooms
revealed that a significant amount of variation in mathematics performance on all three
dependent measures remained to be modeled after accounting for covariates. However,
the amount of variation between classrooms was relatively small, requiring a large
sample size to detect effects.

Finally, the way objectives are assessed on AIMS changed dramatically between
2004 and 2005 for one third-grade objective and for both fifth-grade objectives. The fifth-
grade changes increased the consistency between the way problems are presented in the
district-adopted text and on AIMS. For the third-grade objective, the items changed to
better match the objective; however, the AIMS items are not consistent with the district
text.

These changes are important because if teachers attempted to present examples
that mirrored AIMS, they would have based them on the 2004 exam (they had not yet
seen the 2005 exam at the time of interviews). At fifth grade, a combination of teaching
the text and basing examples off of old AIMS items should have resulted in high levels of
instructional alignment with AIMS. At third grade, however, this approach would result in lower levels of alignment for one objective.

Alignment levels did not vary between grade levels. Based on interview transcripts, this seems to be due to the fact that many third grade teachers presented information in ways that extended beyond the text and 2004 AIMS and that fewer fifth grade teachers participated in test preparation, or based their test preparation on different sources of information. The district testing program may also have played a role. Many items on the district assessment mirrored the 2005 AIMS, including those that assessed objectives that the text did not fully address.

Although the fifth-grade curriculum is better aligned, third-grade students performed considerably better (relative to the state) than fifth-graders, both on tests of study objectives (Thirdtest and Fifthtest) and on AIMS. This seems to indicate that curricular alignment is not predictive of test performance, at least in this district. It could also indicate that the objectives selected for the study are atypical of mathematics instruction in the district. Low levels of internal consistency in subtest scores may also affect testlet results.

In addition, although the amount of variation in AIMS scores was similar across grade levels for the district and the study sample, the amount of variance statewide was greater at fifth grade than at third grade. This may indicate that the presence of a mandatory, districtwide curriculum implemented quite similarly across classrooms at fifth grade (where little supplementation was necessary) increased the uniformity of fifth-grade performance in the district. At third grade, where teachers more frequently
supplemented, variation may have also been increased, which would also minimize the rate of fifth-grade to third-grade variance. More evidence is needed, however, before any definitive statement can be made about the impact of variations in the curriculum on AIMS performance.

**Relationship between instructional validity measures and achievement**

After controlling for student demographics and classroom characteristics, the instructional validity measures: alignment, emphasis, and the alignmentXemphasis interaction did not significantly predict AIMS achievement for the sample as a whole. They also did not predict performance on items related to the fifth grade objectives considered by this study, or predict the overall AIMS performance of fifth-graders. However, alignment did predict performance on the five-item testlet addressing the third grade objectives and the overall AIMS performance of third-graders. The former relationship should be examined with skepticism at best because adding alignment to the model results in very unreliable predictions of classroom mean achievement (estimated Bayesian intercepts). The outcome measure (Thirdtest) also has low levels of internal consistency, due at least in part to the small number of items that comprise the subtest.

Third-grade results for overall AIMS performance are more meaningful since the outcome measure is much more reliable, as are the estimated Bayesian estimates predicted by the HLM model. At third grade, alignment is an important predictor of test performance. This finding is especially notable since teachers could only be aligned if they supplemented the district curriculum. In contrast, fifth-grade teachers who solely
used the district text would automatically be coded as aligned with AIMS. Nevertheless, third- and fifth-grade trajectories were not different enough to result in a significant gradeXalignment interaction. This finding could be affected by a limited sample size (as the analysis is based on only 36 classrooms), but also indicates that the alignment-achievement trajectories are somewhat similar across grade levels.

Several factors could account for a grade-level specific relationship. First, instructional alignment and emphasis on state standards may not be predictive of AIMS performance in this district at all grade levels. Although other studies have found these measures to be important predictors of achievement (see, Gamoran, Porter, Smithson, & White, 1997), this district may be anomalous. Enactment of standards-based report cards to encourage teachers to teach state standards might encourage all teachers to implement standards-based instruction to at least some extent, muting the ability to measure the effect of alignment (especially at those grades where the curriculum and test are pretty well aligned).

Second, the limited number of performance objectives studied may also result in unusual findings. There are approximately 90 mathematics objectives per grade level included in the state standards. Findings based on only two of these objectives may not be representative of the relationship between standards-based instruction and achievement in the district overall. This may be especially problematic because of all of the changes in the state standards and the AIMS test in the year preceding the study. Teachers who made good faith efforts to teach the standards were still in the process of relearning the standards documents at the time data were collected.
Third, the small number of classrooms involved in analysis, and especially the small number of classrooms studied per grade level, could have resulted in the study not having enough power to find an effect. This finding holds special weight given the large drop in reliability of intercepts when alignment is added to the HLM model predicting Thirdtest scores, and in the reduction in variance between classrooms to $u_{0j}=0.00$, an indicator that there are not enough level-2 predictors to support the model. However, it is impossible to say what effects might be found if more classrooms were included in analyses.

Fourth, the AIMS test may be an unusual standards-based assessment. Unlike the tests used in many states, the mathematics portion of AIMS is comprised only of multiple choice items and is used to generate both norm-referenced and criterion-referenced scores. Because of this, instructional alignment to the state standards may sometimes differ from alignment with AIMS. For example, the combinations objective used in this study requires that students “Make a diagram…” In many states, the state mathematics assessment includes a constructed response portion in which students could make a diagram. Because AIMS is comprised only of multiple choice options, the AIMS items require students to select the diagram that corresponds to a combinations problem instead of constructing one on their own.

Finally, AIMS may not be instructionally sensitive at all grade levels or may be of only moderate instructional sensitivity. Since participating teachers all taught at least one performance objective to some extent, the study can not ascertain if there is a difference in the performance between classrooms that did and did not address the objectives, an
approach that is recommended by many, including Popham (2007).

Suggestions for future research

A major limitation of this study is the limited sample involved; participants all came from the same school district, taught at only two grade levels, and discussed two objectives. Future work might expand this approach by including teachers from a wider range of school districts. In particular, it might be useful to include teachers who work in a variety of settings. For example, including districts that haven’t adopted districtwide texts might increase the variation in alignment and in emphasis ratings. Comparing districts using traditional and standards-based report cards might also increase the variation in instructional validity measures.

It might also be helpful to study a wider range of objectives within a grade level as this would help to ensure that results generalize across the full set of standards that teachers are expected to address. To make this more feasible (given the amount of time required to understand how objectives are operationalized), it might be helpful to replace teacher interviews with collection of assessments conducted over the course of the school year, with a teacher feedback form that provides information about the specific response required to receive full points on each assessment. Finally, expanding the data collection to more grade levels would be helpful, especially since this study found different results in third and fifth grade. Examining trends across four grades, for example, would help to determine if the third grade results or the fifth grade results were anomalous.
Conclusion

In conclusion, like all social science research, the present study has several limitations that must be kept in mind while interpreting results. Most prominently, the sample size is small in terms of the number of classrooms studied and the number of performance objectives addressed, and generalizations can not be made across grade levels or school districts since the study collected data from only two grade levels in one school district.

However, the study also more fully represents the construct of instructional alignment by capturing the full range of activities used to teach a set of performance objectives. It identifies, and attempts to address, many of the limitations of other attempts to gauge instructional validity, and adds to the literature by using mixed methods to explore issues of instructional sensitivity which have most often been addressed with solely quantitative techniques. As such, it deepens understanding of the complexity of all the factors affecting instructional alignment.
APPENDIX A: INTERVIEW PROTOCOL
Interview Protocol

Thank you for agreeing to meet with me. As I mentioned in my letter, I am studying how standards-based education plays out in classrooms in Arizona. Therefore, I would like to ask you some questions about how you teach and assess student performance in mathematics. The interview should last about one hour and participation in this study is completely voluntary. Do you have any questions? Do you agree to be interviewed?

[Assuming agrees to interview] Great. Please read over this subject consent form, indicating that you agree to participate in this interview. Just as a reminder, you are free to skip questions or to stop the interview altogether at any time.

Now let’s start with some basic background information:

1. How long have you been teaching?
2. How long have you taught 3rd/5th grade in Arizona?
3. What was your major in college?
4. What’s your highest level of education?
5. What subjects are you certified in?

For the rest of our conversation, I’d like you to walk me through how you teach and assess performance on two performance objectives in mathematics. Let’s start with the first performance objective and then we’ll repeat the process with the second performance objective (refer to table below).

<table>
<thead>
<tr>
<th>Performance Objective 1</th>
<th>Grade 3</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make a diagram to represent the number of combinations available when 1 item is selected from each of 3 sets of 2 items (e.g., 2 different shirts, 2 different hats, 2 different belts).</td>
<td>Interpret graphical representations and data displays including bar graphs, circle graphs, frequency tables, three-set Venn diagrams, and line graphs that display continuous data. AND Answer questions based on graphical representations and data displays.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance Objective 2</th>
<th>Grade 3</th>
<th>Grade 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discriminate necessary information from unnecessary information in a given grade-level appropriate word problem.</td>
<td>Describe the rule used in a simple grade-level appropriate function (e.g., T-chart, input-output model).</td>
<td></td>
</tr>
</tbody>
</table>
Okay. The following questions are designed to help me understand the experiences of your students as they learned this material. Pretend that I am a student in your class. Using the materials you have brought with you, please describe how you would help me learn this performance objective. In particular (Probe for):

6. When in the school year would I learn about this performance objective?

7. How much time would I spend learning it? (number of lessons over how many hours/weeks)

8. What materials would you use?

9. What would occur during these lessons?

10. What problems would you have me work through?

11. What do you want me to learn from these problems?

12. What products would you expect me to create to show that I have mastered these skills?

How would you assess my learning? Please show me the assessments you would use to determine whether I have learned this performance objective and explain your thinking in designing/selecting these assessments.

Additional questions/probes (some may be not applicable if teachers use end of chapter assessments out of the book)

13. What assessment tools or methods do you use?

14. What are the purposes of these assessments?

15. What do you hope students will get out of completing these assessments?

16. The performance objective covers a wide array of material, how do you decide which skills to assess?

17. What responses do you look for as evidence of learning?

18. What would you do with the assessment after it is graded? (probe for: return it to students, go over answers with class, discuss results with students or parents one on one, use results to refine instruction.)
19. (Present test items that assess the performance objective). Here’s how one test measures student proficiency on this objective. Do you think that this test accurately captures what you teach to students?

[Repeat 6-19 with second objective]

20. What do you expect students to do with the information they get from graded assessment?

21. In some cases, students may learn a performance objective one way in class and be asked a question in a different format on the state test. How do you concern yourself with this?

22. (If yes), do you purposefully teach students so that your instructional efforts are reflected in students’ AIMS scores?

23. (If yes), how do you ensure that your classroom instruction is aligned with AIMS?

24. How would you describe any changes you have made in your mathematics instruction since last year? (Not at all, A little, quite a lot, completely) If quite a lot or completely: Please describe the changes you have made.

25. Thank you for your time. Do you have any questions for me or anything you would like to add to what you have said?
APPENDIX B: TESTLETS REVIEWED BY TEACHERS
Third Grade Test: Discriminate necessary from unnecessary information in a given grade-level appropriate word problem.

1. Keith’s mom made cupcakes. Each cupcake had five candies on top. What else do you need to know to find the total number of candies Keith’s mom used for the tops of the cupcakes?
   A. The number of candies that come in a bag
   B. The number of cupcakes she made*
   C. The size of the candies
   D. The time it took to bake the cupcakes

2. There are 10 boys and 15 girls in Mr. Kendrick’s class. Mr. Kendrick bought 1 pencil for each student in his class. The pencils cost 25¢ each. The pencils are either red or blue. How much did the pencils cost all together?
   What information is not needed to solve the problem?
   A. The pencils cost 25¢ each.
   B. There are 10 boys and 15 girls in the class.
   C. Mr. Kendrick bought 1 pencil for each student.
   D. The color of the pencils.*

3. Jerry bought five cases of oil. He wants to know how many cans of oil he bought. What other information is needed?
   A. Where the oil was purchased
   B. How many cans of oil were in each case*
   C. How many cans of oil were needed
   D. How many trips were made to the store

4. Steven had $30. He bought a T-shirt, a hat, and a package of model cars at the store. He spent a total of $11 including tax on the T-shirt and the hat together. What other information is needed to find how much money Steven had left after leaving the store? Mark your answer.
   A. The price of the T-shirt
   B. The number of model cars in the package
   C. The price of the package of model cars*
   D. The price of each model car
Third Grade Test: Make a diagram to represent the number of combinations available when 1 item is selected from each of 3 sets of 2 items (e.g., 2 different shirts, 2 different hats, 2 different belts).

1. George is planning to decorate his bedroom. He will choose 1 color of paint, 1 color of carpet, and 1 pattern for wallpaper. The colors and patterns he will choose from are shown below.

How many different combinations of 1 paint color, 1 carpet color and 1 wallpaper pattern are possible?

A. 7
B. 8
C. 10
D. 12

2. Jay’s father took 3 shirts and 2 pairs of pants on a trip. Which of the following pictures shows how many outfits he can make?
3. The table below shows the different colors and kinds of shirts that a club member can choose from.

<table>
<thead>
<tr>
<th>Shirt Choices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
</tr>
<tr>
<td>Green</td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Which of the following lists all the different ways to combine 1 color and 1 kind of shirt?

A. Green, T-shirt *
   Green, Tank top
   Green, Sweatshirt
   White, T-shirt
   White, Tank top
   White, Sweatshirt

B. Green, T-shirt
   Green, Tank top
   White, Tank top
   White, Sweatshirt

C. Green, Tank top
   Green, Sweatshirt
   White, T-shirt
   White, Tank top
   White, Sweatshirt

D. Green, T-shirt
   Green, Sweatshirt
   White, Tank top
   White, Sweatshirt

4. Shelley is ordering a skirt from a catalog. She can choose one of two lengths: a short skirt or a long skirt. Then she can choose one of the three fabric patterns: stripes, plaid or flowers.

How many different skirts could Shelley order choosing a length and a fabric pattern?

A. 2
B. 3
C. 5
D. 6*

5. The table below shows the ice cream flavors and toppings available for Mrs. Johnson’s class party.

<table>
<thead>
<tr>
<th>Ice Cream</th>
<th>Topping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chocolate</td>
<td>Nuts</td>
</tr>
<tr>
<td>Vanilla</td>
<td>Chocolate Chips</td>
</tr>
</tbody>
</table>

How many different ice cream cones can be made with one scoop of ice cream and one topping?

A. 4*
B. 5
C. 6
D. 7
Fifth Grade Test: Describe the rule used in a simple grade-level appropriate function (e.g., T-chart, input-output model).

1. A function machine triples a number. What number would you get if you put a y into the function machine?
   A. 3 + y
   B. 3 - y
   C. 3 ÷ y
   D. 3 x y*

2. Which of the following is a possible rule for the input-output table shown below?
<table>
<thead>
<tr>
<th>Input</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
<td>11</td>
</tr>
</tbody>
</table>

   A. The output is 2 plus the input.
   B. The output is 3 times the input.
   C. The output is 0.5 less than half the input.
   D. The output is 1 more than twice the input.*

3. The number machine shown below uses a rule to change each number that is put into it to a different number. The same rule is used every time.

   The table below shows what happened when three different numbers went into the same machine.

<table>
<thead>
<tr>
<th>In</th>
<th>Out</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>12</td>
<td>28</td>
</tr>
</tbody>
</table>

   Which of the following could be the rule used by this number machine?
   A. Add 12
   B. Add 24
   C. Multiply by 2, then add 4 to the result*
   D. Multiply by 3, then subtract 4 from the result
4. The table below shows the number of paddles Mr. Watson must order for different numbers of canoes.

<table>
<thead>
<tr>
<th>Number of Canoes</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Paddles</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>?</td>
</tr>
</tbody>
</table>

If the pattern in the table continues, how many paddles must be ordered for 10 canoes?

A. 17  
B. 18  
C. 20* 
D. 23

5. The number sentences below are true for a certain value of N.

2 N = 4  
3 N = 6  
4 N = 8

If the value of N does not change, what number goes inside the box below to make the number sentence true?

5 N =

A. 4  
B. 5  
C. 10* 
D. 12

6. If a 2 goes into the function machine, what number would come out?

A. 5  
B. 6* 
C. 7  
D. 9
Fifth Grade Test: Interpret graphical representations and data displays including bar graphs, circle graphs, frequency tables, three-set Venn diagrams, and line graphs that display continuous data. AND Answer questions based on graphical representations and data displays.

1. The graph below shows the number of hours Barbara studied each day for her final exam during a five-day period. How many hours did Barbara study for her exam in all?
   A. 3  
   B. 6  
   C. 9*  
   D. 10

2. Elm school had a book sale. The graph below shows the number of books sold during each day of the sale. What is the difference between the GREATEST number of books sold and the LEAST number of books sold?
   A. 20  
   B. 30  
   C. 40*  
   D. 50
3. Paula timed how long some traffic lights stayed red before changing to green. The graph below shows how many seconds the traffic light at each cross street stayed red.

At which cross street is the traffic light red more than 10 seconds, but less than 20 seconds?

A. Pine St.
B. Elm St.
C. Market St.
D. Cayuga St.*

4. There are 28 students in Mrs. Garcia’s class. According to the graph, how many students in her class did not participate in the survey?

A. 1 student
B. 3 students*
C. 4 students
D. 5 students
5. The Venn diagram below shows the number of students in three different activities at Jonah’s school.

What is the total number of students in more than 1 activity?

A. 4
B. 39
C. 43*
D. 79

6. Connie recorded her height on her birthday for the 6 years shown in the table below.

<table>
<thead>
<tr>
<th>Year</th>
<th>Height (inches)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1993</td>
<td>43.5</td>
</tr>
<tr>
<td>1994</td>
<td>47</td>
</tr>
<tr>
<td>1995</td>
<td>50</td>
</tr>
<tr>
<td>1996</td>
<td>53.25</td>
</tr>
<tr>
<td>1997</td>
<td>57.5</td>
</tr>
<tr>
<td>1998</td>
<td>60</td>
</tr>
</tbody>
</table>

Between which two consecutive years did Connie’s height change the least?

A. between 1993 and 1994
B. between 1995 and 1996
C. between 1996 and 1997
D. between 1997 and 1998*
7. The graph below shows that the amount of yearly rainfall for Smallville remains fairly constant each year.

Based on the graph, which of the following is the most reasonable estimate for the amount of rainfall that Smallville received in 2002?

A. 5 inches  
B. 12 inches  
C. 20 inches*  
D. 34 inches

8. Use the Venn diagram below to answer the following question.

If 56 students have dogs for pets, how many students own cats, dogs and fish at the same time?

A. 9  
B. 11*  
C. 23  
D. 32
APPENDIX C: ALIGNMENT RUBRICS
Rubrics for Third Grade Objective: Discriminate necessary from unnecessary information in a given grade-level appropriate word problem.

<table>
<thead>
<tr>
<th>Score</th>
<th>Item 72</th>
<th>Item 10</th>
<th>Item 37</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect Alignment (3)</td>
<td>Selects missing information from a list of possible options</td>
<td>Indicates which item in a list is consistent with student response</td>
<td>Selects unnecessary information from a list of possible options</td>
</tr>
<tr>
<td>Close Alignment (2)</td>
<td>Identifies what’s missing to solve problems</td>
<td>Unnecessary information fits with content of word problem</td>
<td>Unnecessary information fits with content of word problem</td>
</tr>
<tr>
<td></td>
<td>Generates a number sentence to represent a word problem</td>
<td>Responds to open-ended questions about unnecessary information</td>
<td>Students indicate (yes or no) whether specific information is unnecessary</td>
</tr>
<tr>
<td>Some Alignment (1)</td>
<td>Requires students to identify needed information</td>
<td>Determines that the problem requires students to identify unnecessary information</td>
<td>Determines that the problem requires students to identify unnecessary information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Diagrams word problems to emphasize needed/unneeded information</td>
<td>Diagrams word problems to emphasize needed/unneeded information</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unnecessary information is clearly unrelated to this problem</td>
<td>Unnecessary information is clearly unrelated to this problem</td>
</tr>
<tr>
<td>Not Aligned (0)</td>
<td>Does not teach</td>
<td>Does not teach</td>
<td>Does not teach</td>
</tr>
</tbody>
</table>
Rubrics for Third Grade Objective: Make a diagram to represent the number of combinations available when 1 item is selected from each of 3 sets of 2 items (e.g., 2 different shirts, 2 different hats, 2 different belts).

<table>
<thead>
<tr>
<th>Score</th>
<th>Item 70</th>
<th>Item 58</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect Alignment (3)</td>
<td>Interprets table</td>
<td>Interprets table</td>
</tr>
<tr>
<td></td>
<td>Interprets visual and written information</td>
<td>Interprets visual and written information</td>
</tr>
<tr>
<td></td>
<td>Interprets a 3 set tree diagram</td>
<td>Interprets a 3 set picture list</td>
</tr>
<tr>
<td>Close Alignment (2)</td>
<td>Combinations involve 3 sets of items AND multiple visual displays</td>
<td>Combinations involve 3 sets of items AND multiple visual displays</td>
</tr>
<tr>
<td></td>
<td>OR students create a tree diagram</td>
<td>OR students create a picture list</td>
</tr>
<tr>
<td>Some Alignment (1)</td>
<td>Introduces concept of combination</td>
<td>Introduces concept of combination</td>
</tr>
<tr>
<td></td>
<td>Selects 1 item from each set</td>
<td>Selects 1 item from each set</td>
</tr>
<tr>
<td></td>
<td>Represents combinations in some way (list or diagram)</td>
<td>Represents combinations in some way (list or diagram)</td>
</tr>
<tr>
<td></td>
<td>Uses relevant vocabulary (combination, diagram, different)</td>
<td>Uses relevant vocabulary (combination, diagram, different)</td>
</tr>
<tr>
<td>Not Aligned (0)</td>
<td>Does not teach skill</td>
<td>Does not teach skill</td>
</tr>
</tbody>
</table>
Rubric for Fifth Grade Objective: Describe the rule used in a simple grade-level appropriate function (e.g., T-chart, input-output model).

*Note: The same rubric was used with all four items as they were essentially the same question with new numbers/operations inserted.*

<table>
<thead>
<tr>
<th>Score</th>
<th>One Rubric for All Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect Alignment (3)</td>
<td>Selects function rule from a list of possibilities</td>
</tr>
<tr>
<td>Close Alignment (2)</td>
<td>Creates rule to describe relationship between input and output</td>
</tr>
<tr>
<td></td>
<td>Demonstrates that addition/multiplication will increase output and subtraction/division will decrease output</td>
</tr>
<tr>
<td>Some Alignment (1)</td>
<td>Communicates what a function is; that there is a one-to-one correspondence between an input and an output that applies to all ordered pairs</td>
</tr>
<tr>
<td></td>
<td>Explains concept of ordered pair</td>
</tr>
<tr>
<td>Not Aligned (0)</td>
<td>Does not teach</td>
</tr>
</tbody>
</table>
Rubrics for Fifth Grade Objective: Interpret graphical representations and data displays including bar graphs, circle graphs, frequency tables, three-set Venn diagrams, and line graphs that display continuous data. AND Answer questions based on graphical representations and data displays.

<table>
<thead>
<tr>
<th>Score</th>
<th>Item 53</th>
<th>Item 79</th>
<th>Item 80</th>
<th>Item 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perfect Alignment (3)</td>
<td>Selects the correct answer from a list of plausible responses</td>
<td>Selects true/false statement from a list of statements about a bar graph</td>
<td>Selects true/false statement from a list of statements about a circle graph</td>
<td>Selects true/false statement from a list of statements about frequency tables</td>
</tr>
<tr>
<td>Close Alignment (2)</td>
<td>Finds information in a table and uses it to compute an answer</td>
<td>Draws conclusions by comparing information contained in a bar graph</td>
<td>Draws conclusions by comparing information contained in a circle graph</td>
<td>Compares data compared in two separate frequency tables</td>
</tr>
<tr>
<td></td>
<td>Compares information in order to answer a question</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some Alignment (1)</td>
<td>Extracts facts from or constructs frequency tables</td>
<td>Extracts facts from or constructs a bar graph</td>
<td>Extracts facts from or constructs a circle graph</td>
<td>Extracts facts from or constructs frequency tables</td>
</tr>
<tr>
<td>Not Aligned (0)</td>
<td>Does not teach</td>
<td>Does not teach</td>
<td>Does not teach</td>
<td>Does not teach</td>
</tr>
</tbody>
</table>
REFERENCES


*Arizona Instrument to Measure Standards 2005 district results* [Data file]. Phoenix, AZ: Arizona Department of Education.


