JUDGING ALIGNMENT OF CURRICULUM-BASED MEASURES IN MATHEMATICS AND COMMON CORE STANDARDS

by

CHRISTOPHER MORTON

A DISSERTATION

Presented to the Department of Educational
Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Education

December 2013

DISSERTATION APPROVAL PAGE

Student: Christopher Morton

Title: Judging Alignment of Curriculum-Based Measures in Mathematics and Common Core Standards

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Education degree in the Department of Educational Methodology, Policy, and Leadership by:

Dr. Gerald Tindal	Chair
Dr. Keith Hollenbeck	Core Member
Dr. Yong Zhao	Core Member
Dr. Laura Lee McIntyre	Institutional Representative

and

Kimberly Andrews Espy	Vice President for Research & Innovation/Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2013

DISSERTATION ABSTRACT

Christopher Morton

Department of Educational Methodology, Policy, and Leadership

December 2013

Title: Judging Alignment of Curriculum-Based Measures in Mathematics and Common Core Standards

Measurement literature supports the utility of alignment models for application with state standards and large-scale assessments. However, the literature is lacking in the application of these models to curriculum-based measures (CBMs) and common core standards. In this study, I investigate the alignment of CBMs and standards, with specific reference to methodologies historically applied to large-scale assessments and state standards: expertise of judgments, specific training, and rating values. The data are from items developed for the new easyCBM middle school math measures at 6th grade and the 6th grade math portion of the Common Core State Standards (CCSS). Analyses document the degree of reliability between judges. Interclass correlation coefficients reflect moderate reliability and an adequate Index of Agreement with 72% of the items rated as aligned to CCSSs by all judges and 92% by at least two-thirds of the judges. Furthermore, 13 of 15 math items not aligned to CCSSs by at least two judges nevertheless reflect requisite skills required by the standards. Finally, using a two-way ANOVA on two individual judge triads indicate differences in harshness. Future research addresses practical implications regarding the role of CBMs in a comprehensive assessment plan.

CURRICULUM VITAE

NAME OF AUTHOR: Christopher Morton

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Pacific University, Eugene, Oregon
Oregon Graduate Institute of Science and Technology, Tanasbourne, Oregon
Southern Oregon State University, Ashland, Oregon

DEGREES AWARDED:

Doctor of Education, 2013, University of Oregon
Master of Arts in Teaching, 1999, Pacific University
Master of Science, Environmental Science and Technology, 1996, Oregon Graduate Institute of Science and Technology
Bachelor of Science, Geology, 1995, University of Oregon

AREAS OF SPECIAL INTEREST:

Formative assessment
Alignment of assessments, standards, and instruction

PROFESSIONAL EXPERIENCE:

Assistant Director of School Improvement, Redmond School District, Redmond, Oregon, 2012-Present

Academic Achievement Coordinator, Redmond School District, Redmond, Oregon, 2011-2012

Teacher (4th/5th grade), Tom McCall Elementary School, Redmond School District, Redmond, Oregon 2006-2011

Teacher (4th grade), Evergreen Elementary School, Redmond School District, Redmond, Oregon, 2004-2006

Teacher (5th grade), Culver Elementary/Middle School, Culver School District, Culver, Oregon, 2001-2004

Teacher (5th grade), Meadow View School K–8, Bethel School District, Eugene, Oregon, 2000-2001

Teacher (5th grade), Crest Drive Elementary School, Eugene 4J School District, Eugene, Oregon, March 2000-June 2000

ACKNOWLEDGMENTS

I am grateful to my parents for their endless encouragement throughout this process and for instilling in me the value of education. I give a special thank you to the members of my doctoral cohort for their limitless support and friendship. I would also like to sincerely thank all the incredibly professional colleagues from whom I have learned and who have supported my efforts in pursuit of this degree. Additionally, I am grateful for the learning that was imparted to me by my advisor, Dr. Gerald Tindal, and the dedicated and supportive professors in the Department of Educational Methodology, Policy, and Management within the College of Education.

I dedicate this work to my amazing wife Nishka and daughters Emma and Bella. Nishka is the love of my life and the one to whom I attribute my accomplishments both personally and professionally. Bella and Emma are loving, sweet, creative, and talented. I am so proud of my girls and I know in my heart that they will accomplish everything they set their minds to in life.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

In the past, educational reform focused on the implementation and application of state-by-state standards and assessments to measure the academic performance of students across the nation (NCLB, 2001). The utility of these high-stakes assessments for determining students' content area proficiency had potential consequences for students, teachers, schools (Jiban & Deno, 2007), and districts. More recent, A Blueprint for Reform; The Reauthorization of the Elementary and Secondary Education Act (USDOE, 2010), supported renewed reform of the education system with a focus on the following:

> 1) Improving teacher and principal effectiveness to ensure that every classroom has a great teacher and every school has a great leader; (2) Providing information to families to help them valuate and improve their children's schools, and to educators to help them improve their students' learning; (3) Implementing college- and career-ready standards and developing improved assessments aligned with those standards; and (4) Improving student learning and achievement in America's lowest-performing schools by providing intensive support and effective interventions. (p. 3)

To meet these rising educational expectations and to meet the needs of all students, teachers must be able to efficiently and effectively access and interpret student performance data on a continuous basis. Curriculum-based measures (CBM) have provided educators with a formative tool for monitoring student

performance overtime and for making informed instructional decisions in the classroom.

My study addressed alignment of CBMs in mathematics to the Common Core State Standards (CCSS) and the degree to which items that did not strongly align could then serve as requisite skills to the CCSS at the sixth grade level. Methodological components utilized in this dissertation have previous application with large-scale assessments and standards. My research fills a significant void in the measurement literature given that alignment of standards and assessments is common practice for large-scale tests but few CBMs have undergone alignment with standards to date. In the end, a CBM aligned to state or common core standards would allow educators to predict student outcomes on large-scale assessments as well as inform instructional decisions in the classroom.

Alignment Models

In a review by La Marca (2001), an argument of validity was framed in the context of aligning standards and assessment. Similar to Messick's (1989) view, the basis of validity focuses on teacher inferences of assessment results aligned with curriculum and instruction (driven by standards) and not on the quality of the assessment. "Therefore, alignment is a key issue in as much as it provides one avenue for establishing evidence of score interpretation" (La Marca, 2001, p. 5). La Marca stressed the importance of alignment within an accountability system (e.g., standards and assessment) in terms of methodological procedures and ethical necessity. In other words, to judge a stakeholder (e.g., student, teacher, etc.) using

an assessment inadequately aligned to the standards students are striving to achieve, is unacceptable.

Various alignment models utilized similar methodologies to differing degrees. Four alignment models appeared throughout the literature (Case, Jorgensen, & Zucker, 2004; Council of Chief State School Officers, 2002; Bhola, J., Impara, J., and Buckendahl, C., 2003; Tindal, 2005): (a) Webb model, (b) Achieve model, (c) Surveys of Enacted Curriculum model (SEC), and (d) Council for Basic Education model (CBE). Bhola et al. (2003) differentiated these models by level of complexity (i.e., low, medium, or high). The lowest level of complexity forms the baseline for the design of the more complex models. The complex models included an increased number of dimensions investigated and thus a higher degree of alignment determined. Bhola et al. categorized the SEC and CBE models in the moderate complexity category and the Webb and Achieve models in the high complexity category. This literature synthesis is limited in focus to these two models.

Achieve model. Achieve Inc. developed an alignment model for use with its educational alignment service (Achieve Inc., 2012). The model utilized expert judges to rate five dimensions of alignment involving both qualitative and quantitative analysis (Tindal, 2005).

Judgment dimensions. The five dimensions include (a) content centrality, (b) performance centrality, (c) challenge, (d) balance, and (e) range (Council of Chief State School Officers, 2002; Resnick, Rothman, & Slattery, 2003-2004; Tindal, 2005).

1. *Content Centrality* is the degree to which each item on the assessment is a match with the related content standard. This is a measure of degree and quality of match (Resnick et al., 2003-2004).

2. *Performance Centrality* is the degree to which the performance expected from each item on the assessment matches the expectation of performance on the standard.

3. *Challenge* is the degree to which the level of cognitive demand required by an assessment item compares to the level necessary in order to meet the expectations of the standard.

4. *Balance* is the level of emphasis given to items on the assessment in comparison to the standards. In other words, balance is the degree to which items on the assessment appropriately reflect the emphasis of the standards (Council of Chief State School Officers).

5. *Range* is "the proportion of objectives explicating a standard that are assessed by at least one item" (Tindal, 2005, p. 38).

Studies located. Achieve Inc. conducted numerous studies (Achieve Inc., 2012) that have focused on (a) alignment of state standards and assessments, (b) alignment of college readiness standards to local placement tests, and (c) alignment of assessment anchors and tests in reading and mathematics. Of these, 12 have been state reports on alignment of standards and assessment. One peer-reviewed article was located summarizing the results of a five-state alignment study (Resnick et al., 2003-2004). However, none of the articles analyzed alignment between state standards and CBMs.

Resnick et al. (2003-2004) interpreted studies conducted by Achieve Inc. on alignment of state assessments to the respective standards for five states. The states' locations represented three U.S. regions: (a) Northwest region, (b) upper-Midwest region, and (c) mid-Atlantic region. One state utilized a commercially developed test, one state was an early adopter and had well-developed standards and assessments in place, and the remaining three states had recently developed assessments. Resnick et al. reviewed and analyzed the findings to determine the adequacy with which these states' assessments aligned to standards. The results showed that the five states had selected assessment items to reflect the content in their respective state standards. The high degree of alignment within the content centrality and performance centrality criterion supported this.

While the five states had high content alignment, the overall results indicated inadequate alignment of tests and standards based on rater judgment for the remaining dimensions of alignment. In general, the range and balance of items on the assessments insufficiently represented the standards and objectives as some standards were underrepresented in the assessment, if assessed at all (Resnick et al., 2003-2004). The results in the challenge dimension indicated that the standards requiring higher order cognitive skills were "often omitted in favor of much simpler cognitive processes – low- or noninference questions in reading, routine calculations in mathematics, for example" (Resnick et al., 2003-2004, p. 25).

The outcomes of the study conducted by Resnick et al. (2003-2004), supported the importance of aligning large-scale assessments to standards during test development. The study did not, however, reference judge experience or judge

training. This was likely due to the confidential nature of the state reports that Resnick et al. interpreted:

> The reports to states that Achieve makes are confidential. For this reason we are not able to show here details of the analysis state by state. For this article, we use results from five states whose data are most complete and were made available for this report. (p. 5)

Case et al. (2004), however, indicated that the Achieve Model utilized a panel of trained judges. Furthermore, the Achieve Model relied on judges with content area expertise when rating the alignment of test items and standards (Council of Chief State School Officers, 2002).

Webb model. The Webb model was both qualitative and quantitative in nature. The model consisted of five sequentially ordered alignment categories focused on (a) content, (b) articulation across grades and ages, (c) equity and fairness, (d) pedagogical implications, and (e) system applicability (Webb, 1997a, 1997b). First, trained raters make judgments on the depth of knowledge (DOK) required by each benchmark and underlying objectives within the content area standard (Council of Chief State School Officers, 2002; Tindal, 2005). Next, "reviewers determine the objective or benchmark represented by each item or task on the state assessment being reviewed, and they rate the level of knowledge necessary for a student to successfully complete the item or task" (Tindal, 2005, p.3). The results of these qualitative judgments undergo statistical analysis for (a) categorical concurrence, (b) depth of knowledge consistency, (c) range of knowledge correspondence, and (d) balance of representation. A fifth alignment

dimension, source of challenge, is sometimes utilized for helping determine whether a question could be answered correctly without adequate content knowledge or incorrectly in spite of adequate knowledge (Webb, 2007a).

Judgment dimensions. *Categorical Concurrence* is a broad indicator of how well the assessment and the standards represent the same content (Webb, 2007a). In other words, "at least some element of the content of the standard appears in the assessment" (Bhola et al., 2003, p.23). Webb indicated that adequate categorical concurrence requires a minimum of six-assessment items associate with each standard. This is a measure of the number of matches between an assessment item and an underlying objective within a standard. Although a one-to-one ratio is the norm, an assessment item might link to as many as three objectives within the same standard (Webb, 2007a).

Depth of Knowledge Consistency (DOK) is adequate when 50% of the items on that assessment require a minimum level of cognitive complexity that is at or above the expectation set within the standard (Webb, 2002). A prerequisite to determining DOK consistency is having assigned a DOK rating to each assessment item and each objective underlying the standards (Webb, 2007a). Reviewers assign one of four levels to rate the DOK of each item: (a) recall, (b) skill/concept, (c) strategic thinking, and (d) extended thinking (Tindal, 2005).

Range of Knowledge Correspondence is the breadth of knowledge required by an assessment item and the associated standards or objectives to which it corresponds (Webb, 2007a). An adequate number of items on the assessment must represent the goals and objectives of the standards in order for reasonable

alignment to occur (Bhola et al., 2003). Webb (2007a) suggested that in order to achieve an adequate range of knowledge correspondence at least 50% of the objectives within a standard must align to an assessment item.

Balance of Representation is the degree of distribution of test items among benchmark objectives. Webb (2007a) suggested that 50% of the objectives underlying a standard require an associated test item in order for adequate balance of representation.

In addition to these dimensions of alignment, La Marca (2001) also emphasized the need for "sound standards and assessment development activities" (p. 3). He suggested that viewing all phases of standards and assessment development through the lens of measuring expected student achievement was critical. The development process should include (a) development of test specifications/blueprints, (b) assessment items designed and aligned to measure specific standards and underlying objectives, and (c) a post hoc review of alignment following test creation (La Marca, 2001). Furthermore, following a change in cut scores or assessment modification, an alignment review is necessary.

Studies located. Searches conducted of databases (i.e., ERIC, PsycINFO, and Academic Search Premier), web sites (i.e., Wisconsin Center for Educational Research, Adding Value to the Mathematics and Science Partnership Evaluations, and Norman Lott Webb), and the Internet resulted in numerous alignment studies that utilized the Webb model or a subset of its components. The Webb Model was utilized to generate at least 10 state reports on the alignment of state standards and assessments (Council of Chief State School Officers, 2002). The Wisconsin Center

for Education Research (WCER) indicated, "Some 25 states have used WCER's online Web Alignment Tool (WAT) to guide and automate the [alignment] process" (Webb, 2007b). The WAT is an online version of the Webb model. The utility of the WAT by numerous states further supports the application of the methodological components (e.g., judges, training, and values) utilized within the Webb model for aligning standards and assessments. In a study (i.e., technical document) conducted by Nese, Lai, Anderson, Park, Tindal, and Alonzo (2010) methodological components based on the Webb model were employed to judge the alignment of easyCBM math measures and a set of curricular standards (i.e., National Council of Teachers in Mathematics (NCTM) Focal Point Standards). One peer-reviewed article (Roach, Elliott, & Webb, 2005) was located on the topic of aligning state standards with an alternate assessment utilizing the Webb model. In addition to alignment studies, articles that focused specifically on the Webb model included (a) one book chapter, (b) two research monographs, (c) two working papers, and (d) three papers presented.

Webb (1999) conducted an alignment study of science and math standards with large-scale assessments in four states. Raters judged the degree of alignment between assessment items and standards in the four primary dimensions of the Webb model. The goals of the study included (a) improving upon the procedures for aligning assessments and standards leading to reliable and valid results, (b) determining the level of training raters required in order to make adequate expert judgments regarding alignment, and (c) refining the specificity of the DOK levels. During a four-day Alignment Analysis Institute in 1989, raters received limited

training on application of the alignment dimensions and focused primarily on improving the process for its utility under more formalized alignment conditions in future studies.

As part of the training on the four DOK ratings, judges identified key indicators within the individual levels. Webb (1999) indicated that the reviewer's effectiveness decreased overtime because of rating multiple states in rapid succession. The raters confused the wording within the DOK levels resulting in inaccurate ratings. In particular, the judges "did experience some interference in their thinking in trying to recall and locate objectives that matched assessment items" (p. 20). Webb stressed the importance of recalibrating overtime and questioned whether to limit judges to rating individual states. Furthermore, judges interpreted a score of two as encompassing a broader range of items and therefore assigned the score more often then other values during the alignment process. Conversely, judges interpreted the level one score narrowly, resulting in infrequent assignment of this level to assessment items and standards (Webb, 1999).

An additional limitation resulted from inadequate measures for addressing items that partially aligned to the standards. In other words, the judges "found exact matches for about 10% to 20% of the items, a near match for about 60% to 70% of the items, and no match for the remainder of items" (Webb, 1999, p. 25). He also reported on a number of findings specific to the alignment process:

1. Training and calibration of raters is necessary prior to making expert judgments.
2. Raters require training and understanding in the DOK levels.

3. How raters interpreted the context of assessment items had an impact on which standard was matched.

4. The DOK Level 2 did not differentiate between a substantial number of the items and standards and was therefore utilized most frequently due to its broad interpretation.

5. The DOK Level 1 was under utilized due to the raters' narrow interpretation of its definition.

6. "A clear procedure is needed for coding assessment activities that do not match any of the objectives or the category of expectations being compared with assessment activities" (Webb, 1999, p. 24).

7. At least three raters are required to calculate adequate results on the degree of alignment between assessment items and standards.

8. One rater suggestion was to develop a procedure for indicating near matches as well as exact matches.

In Webb's (1999) study, math problems composed of multiple parts also posed issues in the study regarding judge interpretation of the standards. In some cases, each individual part of a math problem equated to a Level 2 rating. However, when combined into one test item, the judges often coded the score with higher-level rating (e.g., Level 3 or Level 4). Webb indicated that this issue results from the verbiage written within the standards.

Roach et al. (2005) studied alignment between the Wisconsin Alternate Assessment (WAA) and state standards. The WAA is an assessment designed for students with significant cognitive disabilities who are unable to perform

adequately on the Wisconsin state assessment (Wisconsin Knowledge and Concepts Examinations) with accommodations (Roach et al., 2005). Twelve judges underwent training in the alignment process over a two-day period in 2002. In order to calibrate, the judges reached consensus on the depth-of-knowledge rating for both the standards and assessment items. Results of the study indicated that "the performance of the WAA on the four criteria that constitute Web's (1997) alignment model met or exceeded the performance of many states' general education assessment using the same alignment method" (Roach et al., 2005, p. 227-228).

Not only did this support the utility of the Webb model and its underlying methodological components for aligning state standards and assessments, but it also illustrated the degree of disparity in the alignment of standards and high stakes assessments across states. A movement towards a consistent set of standards and assessments among the states may remedy the issue of test alignment from state to state. The CCSS adopted by most states illustrates a trend in this direction. Furthermore, two consortiums comprised of multiple states, the Smarter Balances Assessment Consortium (SBAC) and the Partnership for Assessment of Readiness for College and Career (PARCC), are creating assessments aligned to the CCSS. However, there are no formative assessments aligned to the CCSS to date.

Important Methodological Issues

Although the Webb and Achieve models differed in some aspects, both models utilized common methodological components for conducting alignment analysis. The components included (a) expert judges, (b) training, and (c) rating

values. These methodological variables require consideration when conducting alignment studies and developing new alignment models.

Expert judges. Webb (1997b) stressed the importance of selecting a team of specialists within the specific content area of the study. The rational was that "complex distinctions need to be made requiring a level of sophistication far exceeding general lay knowledge in understanding how students learn" (Webb, 1997b, p. 10). In the area of mathematics, content area specialists should have experience teaching within the same grade level band in which they are reviewing alignment items and standards. Content area specialists might include (a) general education teachers, (b) special education teachers, (c) building- or district-level math coaches, and (d) intervention specialists.

In an alignment study of four state's math assessments and standards at various grade levels, six judges were selected from a trained panel of 16 people comprised of content-area specialists, state assessment consultants, content experts, and researchers (Webb, 1999). In another study, four judges that consisted of state assessment consultants, content experts, and researchers judged the alignment of math standards and assessments for select grade levels within three states (Webb, 2002).

The Achieve Model also relied on judges with content area expertise when rating the alignment of test items and standards (Council of Chief State School Officers, 2002). Resnick et al. (2003-2004) applied a process utilized by Achieve Inc., where judges were selected to represent diverse viewpoints that included content

area experts, teachers, and curriculum specialists. Furthermore, some judges were experienced with standards development and/or large-scale assessments.

Specific training. To ensure accurate results and an adequate level of reliability, rater training is critical. La Marca (2001) emphasized the importance of training in order to normalize scoring and develop consistency between judges. Judges underwent training in order to develop inter-rater reliability and an understanding of the alignment criteria and respective coding procedures. In particular, training focused on the procedures for assigning a DOK rating for each assessment item, standard, and underlying objective (Webb, 2007a). Training also addressed development of a common understanding of the four-score rating scale including examples reflecting each of the specific scores within the scale (Webb, 1999 & 2007a). First, the judges applied these ratings to the individual assessment items and then to the standards and objectives followed by a debriefing designed to help raters refine the process (Webb, 1999). The intent of this consensus process was to engage the raters in a detailed analysis of the standards and objectives (Webb, 1999) and "to determine the DOK levels of the state's objectives" (Webb, 2007a, p. 9).

Case et al. (2004) indicated that the Achieve Model utilized judges that had been trained. In addition, the model included time for judges to discuss their ratings of anchor items during the training process (Resnick et al., 2003-2004). This was likely to move participants towards consensus prior to rating the actual items thus facilitating calibration between judges.

One limitation to the procedure of training judges in person in all of these studies is that it precludes the inclusion of judges from other regions across the nation from participating in the alignment process thus introducing the possibility of local bias. One solution is to consider the use of technology in the selection, training, and score reporting process in order to broaden rater representation from various geographical regions.

Rating values. Raters make expert judgments by assigning rating values for specific dimensions of alignment. These dimensions are pre-established and depend on the alignment model used (e.g., Webb or Achieve).

In the Achieve model, judges assign values for two scoring dimensions. First, the content centrality is rated by determining the degree to which each assessment item matches the expectation stated within a particular standard (Resnick et al, 2003-2004). The four content centrality values include (a) 2=clearly consistent, 1A=not specific enough, (c) 1B=somewhat consistent, and (c) 0=inconsistent.

After making an expert judgment on the content centrality match for a particular item and standard, the judge determines the performance centrality for the pair. The performance centrality measures the degree to which the cognitive demand required by the item compares to the cognitive demand expected by the standard (Resnick et al. 2003-2004). The values assigned for the levels of match include (a) 2=clearly consistent, 1A=not specific enough, (c) 1B=somewhat consistent, and (c) 0=inconsistent.

Using the Webb model, raters judge the categorical concurrence between assessment items and standards. That is, raters match assessment items and

standards that consist of the same content or content categories (Webb, 2007a) referred to as a *hit*.

Raters then review the depth-of-knowledge (DOK) required for a student to complete each assessment item and the DOK required by the matching standard (CCSSO, 2002), based on the links established when judging categorical concurrence. The DOK levels include (a) Level 1 (recall), (b) Level 2 (skill/concept), (c) Level 3 (strategic thinking), and (d) Level 4 (extended thinking) (Webb 2007a).

Application of Alignment Models to Math CBMs

It is critical that state standards and assessments align in order to ensure that the content educators teach reflects the items presented on the test. Considering the high-stakes nature of assessments for all stakeholders (e.g., students, teachers, principals, and districts), the degree of alignment between state standards and assessments is paramount. In addition, aligning formative assessments to state standards provides educators the opportunity to screen student performance periodically and thus allow adequate time to make instructional changes prior to the administration of high-stakes summative testing. Knowledge of the requisite skills that students are lacking can provide important information for differentiating or individualizing instruction in order to increase student progress towards meeting specific standards.

Applications of the Webb model and Achieve model. Research supports the utility of either the Achieve Model or Webb Model for determining the degree of alignment between state standards and large-scale assessment. Although all of the state-base alignment reports conducted by Achieve are confidential (Resnick et al.,

2003-2004), the Achieve web site includes numerous reports, from various geographic locations including: (a) Montgomery County (2003), (b) New Jersey (2002), (c) Oklahoma (2002), (d) Massachusetts (2001), (e) Oregon (2000), (f) Indiana (2000), and (g) Michigan.

Webb also conducted numerous alignment studies of states standards and assessments including at least 10 state reports (Council of Chief State School Officers, 2002) in addition to utility of the WAT online tool by at least 25 states (Webb, 2007b). No reports that utilized the Achieve or Webb model were discovered that studied alignment of CBMs and the CCSS, which is critical given that most states have adopted the CCSS to replace state assessments in the 2014-15 school year. Additionally most states have voluntarily joined one of two state lead consortiums (i.e., SBAC and PARCC) with the intent of developing assessments aligned to these new standards.

Curriculum-based measures versus summative assessments. Although researchers have addressed the alignment of large-scale assessments and standards, considerably less research has been conducted on alignment of CBMs and standards. Researchers have primarily focused on the design and utility of CBMs for over thirty years for benchmarking students at risk or formative evaluation of instruction (Alonzo, Ketterlin-Geller, & Tindal, 2006; Deno, 2003; Fuchs, 2004; Tindal & Nese, 2011) as opposed to the alignment of CBMs to standards. Additionally, CBM research and evaluation focused on technical adequacy, time series data, and the decision-making utility of the measures (Foegen, Jiban, & Deno, 2007). The early days of CBM research focused on frequently administered short-duration measures

for monitoring student progress and instructional effectiveness (Tindal, Nese, & Alonzo, 2009). Alonzo et al. (2006) also noted that one significant feature of CBMs has been the time series data generated from alternate form measures intended for the utility of monitoring student progress over time. Overall, the information collected has enhanced the ability of educators to make informed instructional, curricular, and pedagogical decisions (Alonzo et al., 2006). That is:

> CBMs, which sample skills related to the curriculum material covered in a given year of instruction, provide teachers with a snapshot for their students' current level of performance as well as a mechanism for tracking the progress students make in gaining desired academic skills. (Tindal and Nese, 2011, p. 34)

Given the extensive research and development of CBMs, they have been recommended for use by educators to inform decision-making in a variety of ways (Alonzo et al., 2006): (a) as screening tools, (b) as diagnostic tools, (c) as progress monitoring tools, (d) for instructional intervention guidance, and (e) for school-wide accountability purposes.

In contrast, summative assessments measure student achievement at the end of a school year, term or semester, or instructional unit. The utility of summative assessments is for determining whether students have learned the intended content. This is in direct contrast to CBMs that are utilized to help inform instructional decision-making and provide ongoing progress monitoring data in the classroom. A CBM aligned to state standards provides ongoing feedback as to whether students are making adequate progress toward meeting the grade level expectation as

measured by a large-scale summative assessment (e.g., state or national assessment). Additionally, the alignment of CBMs to state or common core standards provides educators information to adjust instruction midstream in order to prepare students for success on large-scale summative assessments. Because the two consortia assessments (i.e., SBAC and PARCC) under development have been targeted to be summative measures of student achievement by year-end, educators could benefit from a series of alternate form CBMs also aligned to the CCSS.

Math CBM domains. The majority of CBMs research has focused on reading measures with limited studies conducted in mathematics (Alonzo et al., 2006; Tindal & Nese, 2011). There are, however, two common purposes for utilizing math CBMs in the classroom: (a) to monitor student progress within curricular instruction, and (b) to predict student outcomes on large-scale assessments. These measures are useful for monitoring the effects of interventions as well as planning, differentiating, and modifying instruction for students (Christ, Scullin, Tolbize, & Liban, 2008; Clarke & Shinn, 2004; Thurber, Shinn, & Smolkowski, 2002). Furthermore, CBMs aligned to the same standards as the state or national assessments would help inform educators of student progress towards meeting the expectations of these large-scale assessments.

Math concept and application CBMs sample a years worth of curriculum and assess multiple math domains including numeration, applied computation, word problems, geometry, charts and graphs, and measurement (Fuchs, Fuchs, & Zumeta, 2008). CBMs that sample and reflect a year's worth of curriculum are known as general outcome measures (GOMs). General outcome measures are stronger

predictive indicators of student performance on standardized assessments than computation-based CBMs (Helwig, Anderson, & Tindal, 2002). Two reasons are given: (a) problem solving situations require conceptual understanding to appropriately apply math procedures and operations, and (b) the interconnectivity of mathematical domains. That is, the "conceptual understanding within one domain has direct implications for facility within other domains as well" (Helwig, et al., 2002; p. 104).

A study conducted by Helwig et al. (2002) examined the relation between concept-based CBMs (i.e., GOMs) and student performance on statewide math assessments for middle school students with and without learning disabilities. One hundred ninety-nine eighth-grade students from five schools and four districts participated in the study. About half of the students had individual education plans (IEPs) in at least one content area. The researchers designed CBMs to assess students on the curricular content represented on statewide assessments. In particular, the design embedded math problems that required common mathematical procedures into word problems (Helwig et al., 2002). The researcher's purpose was to assess students' conceptual understanding as opposed to computation-based rote memorization. The students completed a computer adaptive test (CAT) of math achievement designed to simulate a statewide assessment for comparison as the criterion measure. Helwig et al. employed a discriminate function analysis (DFA) to measure the effectiveness of the CBM for predicting student performance on the CAT. The results indicated that the CBM predicted with 87% accuracy whether students would meet the standards on the

state assessment as measured by the CAT. The researchers concluded that the results of the study supported their hypothesis. That is, the more successful students were on the CBM, the more developed their mathematical knowledge would be, thus demonstrating the interconnectivity between mathematical domains. These mathematical domains have a complex interconnectivity and a relation to students' overall conceptual understanding (Helwig et al., 2002). Therefore, alignment of assessments to standards is important, particularly for those that are formative.

Empirical results of CBM systems. The three formative assessment systems utilized within the education system are: (a) easyCBM, (b) Dynamic Indicators of Basic Early Literacy Skills (DIBELS), and (c) AIMSWeb. Of these three systems, only one (easyCBM) has undergone alignment studies to a set of standards. At the time of this literature review there were numerous studies addressing alignment of state standards and assessments. However, limited research was available involving alignment between common core standards and math CBMs.

Nese et al. (2010) studied the alignment of the easyCBM benchmark and progress monitoring math measures (1st grade and 3rd through 8th grade) and the National Council of Teacher of Mathematics (NCTM) Focal Point Standards. States commonly utilize the Focal Point Standards during content standards development (Nese et al., 2010). This marked the "first attempt to align a CBM system with modified state curriculum standards" (Nese et al., 2010 p. 15). The term *modified* likely referred to reviewing alignment of easyCBM math measures to the NCTM Focal Points upon which many states' math standards are developed.

Thirteen raters underwent a 1.5 to 2-hour training using a live online training format (Nese et al., 2010). All raters had experience with easyCBM math and had teaching certifications with the exception of one district curriculum specialist. During the training, participants rated practice items and participated in a follow up discussion to justify their ratings. Following the training, judges completed the alignment process independently over a four-week period. Unlike the Webb model, raters judged alignment independently without the opportunity to review and discuss judgments as a group in order build consensus therefore limiting the calibration of the DOK ratings (Nese et al., 2010).

The Webb model was the basis for the alignment methodology: (a) depth of knowledge of focal point objectives, (b) depth of knowledge of items, (c) and alignment between items and the focal point objectives. Additionally, an intraclass correlation coefficient (ICC) helped determine the reliability for the raters' depth of knowledge judgments.

In general, the results of the study indicated that alignment of the easyCBM math items to NCTM Focal Point Standards was strong (Nese et al., 2010). However, one focal point at 8th grade varied in alignment (i.e., 65% to 100%). Additionally, the ICC ranged from .78 to 1.0 indicating moderately high to high inter-rater reliability on the alignment of items and standards. Nese et al. (2010) determined that "the Webb model can provide meaningful information when applied to formative assessments" (p. 15). The use of an online training format allowed raters from across the nation to participate, thus limiting local rater bias.

Anderson, Irvin, Alonzo, and Tindal (2012) studied the alignment of 2012 easyCBM math items and the CCSS for grades six through eight. Additionally, items not aligned to a standard were investigated to determine whether the items assessed a requisite skill required by the standard. Judges consisted of fifteen teachers with experience teaching middle school math. Each item received a rating by three judges.

A many-facets Rasch model (MFRM) was utilized to model and control for variances (e.g., leniency/severity) in raters (Anderson, Irvin, Alonzo et al., 2012). Rater leniency/severity was a concern in that it may have had an effect on how judges rated the alignment of items and the CCSS. The MFRM model allows for an adjusted alignment rating to "be computed that provides an estimate of what the rating on the item would have been had it been rated by a judge with the estimated average leniency/severity" (Anderson, Irvin, Alonzo et al., 2012, p. 9). The first MFRM analysis (i.e., primary analysis) was conducted using a fully crossed design and was intended to improve the accuracy of the alignment results by controlling for rater severity. The purpose of the second MFRM analysis (i.e., exploratory analysis) was to examine whether item ratings differed by math domain or grade level. This analysis utilized a nested design and therefore required greater caution when interpreting the results (Anderson, Irvin, Alonzo et al., 2012).

The results of the primary analysis indicated that 87% of the sixth through eighth grade items aligned to the CCSS when controlling for judge leniency/severity (Anderson, Irvin, Alonzo et al., 2012). This number resulted after the MFRM model adjusted specific ratings from not aligned to aligned or aligned to not aligned based

on judge leniency/severity. Additionally, 99.6% of the items aligned to either a standard or a pre-requisite skill required by the standard. This percentage was calculated by including all items that were rated as either aligned to a standard or as assessing a requisite skill required by the standard by at least two out of three judges. The effect of grade level and math domain on the variance of judge ratings was 0.25% and 5%, respectively. However, the results of the exploratory analysis should be interpreted with caution due to the nested study design (Anderson, Irvin, Alonzo et al., 2012).

They concluded that "modeling the rater variance as an additional parameter in the model likely produced more accurate results overall, as threats of systematic rater variance are minimized" (p. 15).

Irvin, Park, Alonzo, and Tindal (2012) conducted an alignment study of the easyCBM benchmark math items for fall, winter, and spring and the CCSS for grades six through eight. The study was conducted in two phases and involved reviewers with varying experience including general education teaching, special education teaching, and math coaching.

In Phase 1 of the study, one reviewer per grade level (i.e., sixth through eighth) was selected to align the 135 benchmark items at each grade level with the on- and prior-grade CCSS (Irvin et al., 2012). The prior-grade CCSS were also used to account for the inclusion of items with a broad range of difficulty within the easyCBM benchmark assessments allowing for the sensitivity of the assessment to capture students at and below grade level.

Phase 2 required the selection of four additional reviewers per grade level with the requirement to complete a 45-minute webinar training as part of this phase (Irvin et al., 2012). Reviewers utilized an online data distribution and collection tool, Distributed Item Review (DIR), to deliver the math items and collect reviewer ratings. The reviewers accessed the items and the associated CCSS, as determined in Phase 1, and used a rating scale (i.e., 0-3) to rate the degree of alignment. If an item was rated as not aligned to an on-grade level or prior-grade level standard (i.e., 0), the reviewer determined whether the item aligned to a prerequisite skill required for success of on-grade level content.

Irvin et al. (2012) indicated that about 99% of the items at sixth grade, 93% of the items at seventh grade, and 96% of the items at eighth grade aligned to on-grade or prior-grade CCSS. It is likely that some standards are over represented and others are under represented within the individual benchmark assessments (Irvin et al., 2010). However, from a practical standpoint, the study does "provide clear guidance into areas within the CCSS for which the current easyCBM assessment are insufficiently aligned" (Irvin et al., 2010, p. 28).

Summary of Alignment Models, Methods, and Applications to CBMs

Literature supports the design and application of alignment models (e.g., Achieve and Webb) for use with large-scale assessments. Additionally, there are defined methodological components of these alignment studies that include the use of (a) expert judgments, (b) specific training, and (c) rating values. The literature, however, is lacking in the application of alignment models to math CBMs. Yet, educators utilize CBMs to track student progress on a specific skill or set of skills;

therefore, such alignment is critical. In particular, if CBMs and large-scale assessments are aligned to standards, CBMs can not only provide educators with information for monitoring student growth and making instructional decisions, but also with information about meeting the standards as defined by state assessments.

My study uses the same methodological components previously associating state tests and standards by investigating the application of these components to the 2012 6th grade easyCBM middle school math items and the CCSS. The logic behind the study is that if qualified expert judges are selected and adequately trained, then judge ratings are likely to be reliable. If the ratings are reliable, then the degree to which the 2012 6th grade easyCBM middle school math measures predict the CCSS can be determined. In a follow-up of the item analysis and for those items not aligned to standards, a secondary purpose is to determine if they can serve as important requisite skills of the standards. In a follow-up on judges, their harshness/leniency is investigated. Specifically, the four questions addressed in this descriptive study include:

1. What is the degree of reliability between judges?
2. What is the degree of alignment between 2012 6th grade easyCBM math items and the CCSS?
3. If items are not aligned, do they reflect pre-requisite skills required by the standards?
4. Are judges differentially harsh or lenient?

CHAPTER II

METHODOLOGY

This study involved analysis of an extant data set that included judge ratings on the alignment of the 2012 easyCBM middle school math items at 6th grade with the CCSS. The following section includes a description of the (a) subjects, (b) settings, (c) measures, (d) items, and (e) analysis conducted.

Subjects and Setting

Subjects and setting of the study have been described by Anderson, Irvin, Alonzo, et al. (2012) in a technical report published with the University of Oregon research unit, Behavior Research and Teaching (BRT). Subjects responded to an open call (see Appendix A) for educators interested in participating in a research project posted on the BRT web site. Qualifications of the selected judges included one or more of the following (a) experience teaching math, (b) district math coaching, (c) special education, (d) math endorsement, and/or (e) experience with the CCSS. Appendix B lists the specific qualifications of each judge selected for participation. The convenience sample selected included 15 educators from across the United States. Of the 15, six were involved in scoring the 6th grade math items utilized in this dissertation (see Appendix B). Participants received $300 in compensation based on a rate of $25.00 per hour for an estimated 12 hours work. Each judge rated the alignment of 270 easyCBM math items and the CCSS. If an item did not align, judges rated whether the item aligned to a requisite skill of the standard.

An online training was repeated on two separate occasions in order to accommodate the rater's schedules (Anderson, Irvin, Alonzo, et al., 2012). An online conferencing service was utilized to deliver the 30- to 60-minute webinar (see Appendix C) on various topics including (a) summative versus formative assessment, (b) easyCBM, (c) universal design, (d) the purpose of the study, (e) alignment procedures, (f) the four-point alignment scale, and (g) use of the Distributive Item Review (DIR) tool, an online tool utilized for disseminating the easyCBM math items to raters and for the collection of expert judgments by the raters from across the country.

Measures and Items

The math items utilized in this dissertation were developed under the guidance of researchers at BRT and published in a technical document titled, *The Development and Scaling of Middle School Mathematics Progress-Monitoring Measures* (Anderson, Irvin, Patarapichayatham, Alonzo, and Tindal, 2012). Criteria considered in the development of the items included population invariance (e.g., universal design) and alignment to standards. Five lead teachers underwent training and instruction to further recruit and train 18 item writers in the development of math items in alignment with the domains of the CCSS. Each writer was commissioned to write 150 items. All lead teachers and item writers had experience teaching middle school math. For specifics regarding qualifications, see Anderson, Irvin, Patarapichayatham et al. (2012).

The lead teachers attended a one day training in December 2010 covering issues associated with the project: (a) item writing recommendations, (b) Universal

Design for Assessment, and (c) a practical and collaborative item writing activity. Prior to item readiness for alignment in the proposed study, (a) a pilot plan was developed, (b) anchor items were chosen, (c) pilot forms were created, (d) subjects were select for piloting the items, and (e) all items were calibrated to a common scale (Anderson, Irvin, Patarapichayatham, et al., 2012).

In collection of the data utilized in this dissertation, judges were trained to use a rating scale for determining the degree of alignment between formative assessment items and the CCSS (Anderson, Irvin, Alonzo et al., 2012). The values and descriptions for the rating scale include (a) 0=no alignment, (b) 1=vague alignment, (c) 2=somewhat alignment, and (d) 3=direct alignment. Judges underwent training for determining whether items (a) aligned to a standard, (b) aligned to a requisite skill of a standard, or (c) aligned to neither a standard nor a requisite skill. A rating of 2 (somewhat aligned to the standard) or 3 (direct alignment to the standard) indicated that an item aligned to a standard (see Appendix C). In contrast, a rating of 0 (not aligned to a standard and does not address a requisite skill) or 1 (vaguely aligned to the standard, but does address a requisite skill) indicated that the item did not align to a standard (see Appendix C).

Each judge rated three sets of 90 items equally covering all five math domains within the CCSS (Anderson, Irvin, Alonzo, et al., 2012): (a) ratios and proportions, (b) number systems, (c) expressions and equations, (d) geometry, and (e) statistics and probability. In addition, each item set received ratings by three different judges. Raters were provided approximately three to four weeks to complete the rating process. During that time, raters logged into the DIR system

using a unique identifier that displayed the three item sets specific to that rater. The system provided access to the (a) easyCBM math items, (b) underlying standard for each item, (c) the scale for collecting judge ratings, (d) training webinar PowerPoint for reference (see Appendix C), and (e) contact information for communication related to the study. Use of the DIR allowed for a broader cross-section of raters from across a large region to participate in the study thus minimizing local bias.

Analysis

This dissertation included analyses focused on inter-rater reliability and item alignment to the CCSS. Secondary analyses were conducted on requisite skill alignment to the CCSS and judge harshness/leniency.

To control for differences in reliability as a function of math items, all judgments were blocked by math domain (i.e., ratio and proportions, number systems, expressions and equations, geometry, and statistics and probability) and grade level (i.e., sixth grade). The analyses were conducted on two unique groups of judges and items. The first group included a set of 90 math items (18 items from each math domain) and a triad of judges. The second group consisted of a second set of items and a second triad of judges.

For the first analysis, an Intraclass Correlation Coefficient (ICC) was calculated to help determine inter-rater reliability as a function of the variance attributed to judges relative to the total variance. A two-way mixed model was selected based on the assumption that the judges in this dissertation did not represent a random sample of a broader population (Nichols, 1998). But instead, inferences made were specific to the sample population of judges selected.

Next, an Index of Agreement was generated (see Appendix D) in order to investigate (a) inter-rater reliability, (b) the degree of alignment between math items and the CCSS, and (c) whether items not aligned reflected requisite skills required by the standards. The index displaying how judges rated individual items compared to one another. Some possibilities included: (a) 100% agreement (i.e., all 0s, 1s, 2s, or 3s); (b) partial agreement, but off by one (e.g., 2, 2, 3), and (c) off by more than one (e.g., 2, 1, 0). Furthermore, items rated 0 or 1 by a judge underwent a second rating by the same judge to determine whether the item aligned to a pre-requisite skill required by a standard.

The final analysis entailed a two-way factorial ANOVA with two independent variables (i.e., factors) for determining judge leniency and harshness. The analysis provided information regarding the potential presence of main effects and interactions. Specifically, the analysis included a comparison of the marginal means in addition to conducting pair-wise comparisons as a result of an interaction in order to compare the means between judges and across domains.

CHAPTER III

RESULTS

Two judge triads were the focus of the analyses conducted for this study. Each triad was unique in that judges did not overlapped into the composition of both groups. Additionally, the math items in Triad 1 differ from the items in Triad 2. Both triads were analyzed independently for judge (i.e., rater) harshness and inter-rater reliability.

General Description

The first analysis focused on answering the question of reliability. An ICC statistic was calculated to investigate inter-rater reliability as a function of variance partitioned. Next, an Index of Agreement (see Appendix D) was generated and a descriptive analysis conducted focusing on judge agreements. The purpose of the index was to organize information in order to investigate inter-rater reliability, item alignment to the CCSS, and the presence of items that reflect requisite skills required by the standards. Categories included (a) 100% judge agreement, (b) judge agreement off by one, and (c) judge agreement off by more than one. Items rated 0 or 1 by a judge underwent a second rating by the same judge to determine whether the item was aligned to a requisite skill of the standard. Finally, a two-way analysis of variance (ANOVA) was conducted to determine whether math domain and judge had an effect on ratings and if the effect of math domain on ratings is dependent on the specific judge.

The results of each analysis are presented below and organized by judge triad. The results for Triad 1 are reported first followed by the results for Triad 2.

Judge Triad 1

The descriptive statistics for judge ratings by math domain (i.e., item type) are shown in Table 1 and include the number of math items (per domain), means, and standard deviations. Mean judge ratings are included for all five math domains; (a) equations and expressions (EE), (b) geometry (G), (c) number systems (NS), (d) ratios and proportions (RP), and (e) statistics and probability (SP). Judges rated 18 items from each of the five math domains for a total of 90 items rated per judge within each item set. All three judges within Triad 1 rated the same 90 items.

Table 1

Triad 1: Means, SD, and n for Rating as a Function of Judge and Item Type

	Judge 1			Judge 2			Judge 3			Total	
Type	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*M*	*SD*
EE	18	2.83	.383	18	2.89	.323	18	2.67	.767	2.80	.528
G	18	2.78	.428	18	2.44	.705	18	2.22	1.003	2.48	.771
NS	18	2.67	.686	18	3.00	.000	18	2.61	.778	2.76	.612
RP	18	2.72	.826	18	2.89	.323	18	2.44	.705	2.69	.668
SP	18	1.94	.938	18	2.28	.826	18	1.78	1.353	2.00	1.064
Total	90	2.59	7.48	90	2.70	.589	90	2.34	.985	2.54	.802

Inter-rater reliability. An Intraclass Correlation Coefficient (ICC) and Index of Agreement (see Appendix D) were utilized for determining inter-rater reliability between raters and across math domains. The ICC for all three judges across the five math domains was .635 (ρ=<.001) indicating moderate inter-rater reliability. That is, 63.5% of the variance in ratings was common among the three raters.

Additionally, results from the Index of Agreement (see Appendix D) indicated that the judges were in complete agreement on 52% and off by one 22% of the items. In other words, judges were in agreement or off by one on 74% of the items.

Alignment of easyCBM and CCSS. A four point rating scale was utilized for this study. A rating of 0 and 1 indicated that items did not align to a given CCSS. In contrast, ratings of 2 and 3 did align to a CCSS. Results from the Index of Agreement (see Appendix D) showed that 62 out of 90 items (i.e., 69%) were rated as aligned to a standard by all three raters. Additionally, 82 out of 90 items (i.e., 91%) were rated as aligned to a standard by at least two out of the three raters.

If a judge rated an item as 0 (no alignment) or 1 (vague alignment), the same judge rated the item a second time in order to determine whether it aligned to a requisite skill of the standard. Eight items were rated by at least two out of three judges as 0 or 1. Of the eight items, seven (i.e., 88%) were rated as aligned to a requisite skill by at least two of the judges.

Analysis of judge harshness. A two-way analysis of variance (ANOVA) was utilized to examine the effect of math domain and judge on item ratings. However, no significant interaction occurred between math domain and judge (ρ=.693). There was a significant main effect of math domain on ratings, $F(4, 255) = 10.513$, ρ<.001.

As a result of the main effect due to domain, a Bonferroni post hoc analysis was conducted, Table 2, indicating that the statistics and probability domain differed significantly from equations and expressions (ρ<.001), geometry (ρ=.009), number systems (ρ<.001), and ratios and proportions (ρ<.001).

Table 2

Item Type Multiple Comparisons Test

(I) Type	(J) Type	Mean Difference (I-J)	*ρ*	95% Confidence Interval	
				Lower Bound	Upper Bound
EE	G	.31	.286	-.09	.72
	NS	.04	1.000	-.37	.44
	RP	.11	1.000	-.29	.52
	SP	.80*	.000	.39	1.20
G	NS	-.28	.531	-.68	.13
	RP	-.20	1.000	-.61	.20
	SP	.48*	.009	.08	.89
NS	RP	.07	1.000	-.33	.48
	SP	.76*	.000	.35	1.16
RP	SP	.69*	.000	.28	1.09

*. The mean difference is significant

Results of the two-way ANOVA also indicated a significant main effect of judges on rating, $F(2, 255) = 5.396$, ρ=.005. A Bonferroni post hoc analysis (see Table 3) resulted in a statistically significant difference in the ratings between judge 2 and judge 3 (ρ=.004). In particular, judge 3 rated items lower (i.e., harsher) then judge 2 on average.

Table 3

Rater Multiple Comparisons Test

(I) Judge	(J) Judge	Mean Difference (I-J)	*ρ*	95% Confidence Interval	
				Lower Bound	Upper Bound
1	2	-.11	.950	-.38	.16
	3	.24	.085	-.02	.51
2	3	.36*	.004	.09	.62

*. The mean difference is significant

Judge Triad 2

Triad 2 ratings underwent identical analysis as Triad 1 in order to address inter-rater reliability, item alignment, and judge harshness. However, post hoc analysis differed for Triad 2 as a result of a significant interaction between the math domain and judge factors when investigating judge harshness/leniency. The descriptive statistics for judge ratings by math domain (i.e., item type) are shown in Table 4 and include the number of math items (per domain), means, and standard deviations. Judge ratings include all five of the math domains. Judges rated 18 items from each of the five math domains for a total of 90 items rated per judge. All three judges were unique to Triad 2 and rated a different set of 90 items than Triad 1.

Table 4

Triad 2: Means, SD, and n for Rating as a Function of Judge and Item Type

	Judge 1			Judge 2			Judge 3			Total	
Type	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*M*	*SD*
EE	18	2.17	.985	18	2.06	.998	18	3.00	.000	2.41	.901
G	18	2.28	.752	18	2.22	.647	18	2.83	.514	2.44	.691
NS	18	2.78	.548	18	2.83	.514	18	2.67	.686	2.76	.581
RP	18	2.78	.428	18	2.72	.575	18	2.89	.323	2.80	.451
SP	18	2.56	.705	18	2.17	.924	18	2.28	.895	2.33	.847
Total	90	2.51	.738	90	2.40	.804	90	2.73	.614	2.55	.734

Inter-rater reliability. An ICC statistic and Index of Agreement (see Appendix D) were utilized for determining inter-rater reliability between raters and across math domains for Triad 2. The ICC for all three judges across the five math domains was .567 (ρ=<.001) indicating moderate inter-rater reliability with 56.7% of the variance in ratings in common among the three raters. In addition, results from the Index of Agreement indicated that the judges were in complete agreement on 48% and off by one 32% of the items. In other words, judges were in complete agreement or off by one on 80% of the items.

Alignment of easyCBM and CCSS. Results from the Index of Agreement (see Appendix D) showed that 67 out of 90 items (i.e., 74%) were rated as aligned to a standard (i.e., 2 or 3) by all three judges. Furthermore, 83 of the 90 items (i.e., 92%) were rated as aligned to a standard by at least two out of the three judges. Seven of the items, however, received ratings of 0 (i.e., no alignment) or 1 (vague alignment) by at least two-thirds of the judges. Of the seven items, six (i.e., 86%) aligned to a requisite skill as determined by two out of three judges.

Analysis of judge harshness. A two-way ANOVA was conducted to examine the effect of math domain and judge on item ratings. A significant interaction was determined between the effects of math domain and judges on ratings $F(8, 255) = 2.800$, $\rho = .005$. In other words, the effect of judges on ratings depends on math domain. As a result of the interaction, the significant main effects of math domain on ratings $F(4, 255) = 5.275$, ρ=<.001 and judge on ratings $F(2, 255) = 2.593$, ρ=.004 were ignored.

In follow up to the interaction, a post hoc analysis was conducted to compare the simple main effects in order to explain the interaction between math domains and judges. Statistically significant differences are included in Table 5 and means within math domain across judges are displayed in Figure 1. Within the geometry and the equations and expressions domains, there was a statistically significant difference between judge 3 and judge 1 and between judge 3 and judge 2. Specifically, judge 3 rated items higher than judge 1 and higher than judge 2.

Table 5

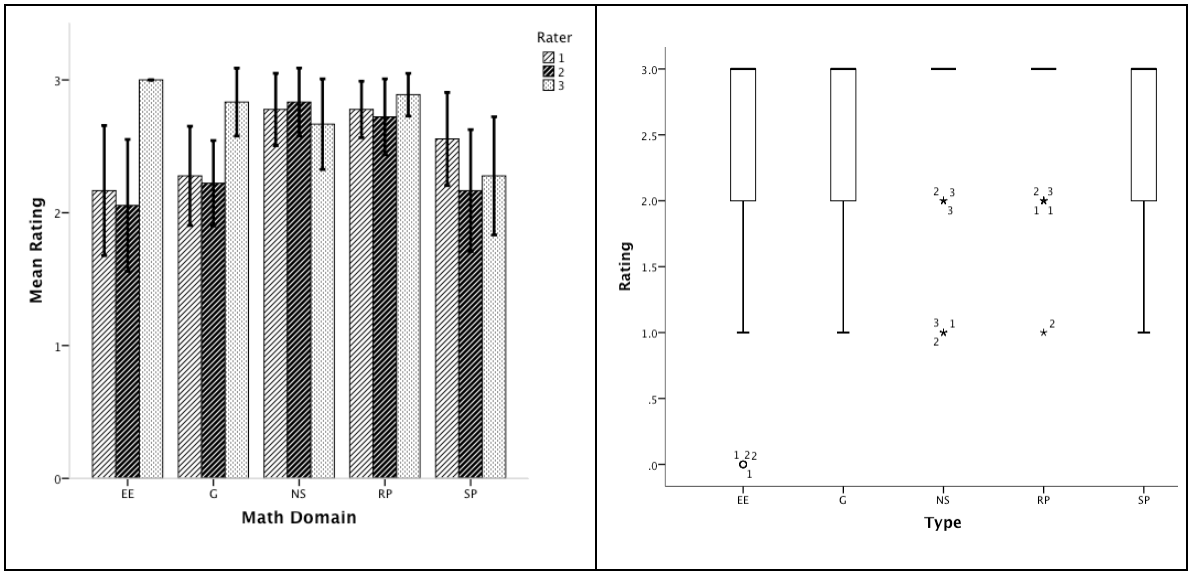
Rater Pairwise Comparisons by Item Type

Type	(I) Judge	(J) Judge	Mean Difference (I-J)	p	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
EE	1	2	.111	1.000	-.438	.661
		3	-.833*	*.001	-1.383	-.284
	2	3	-.944*	*<.001	-1.494	-.395
G	1	2	.056	1.000	-.494	.605
		3	-.556*	*.047	-1.105	-.006
	2	3	-.611*	*.024	-1.161	-.062
NS	1	2	-.056	1.000	-.605	.494
		3	.111	1.000	-.438	.661
	2	3	.167	1.000	-.383	.716
RP	1	2	.056	1.000	-.494	.605
		3	-.111	1.000	-.661	.438
	2	3	-.167	1.000	-.716	.383
SP	1	2	.389	.268	-.161	.938
		3	.278	.673	-.272	.827
	2	3	-.111	1.000	-.661	.438

*. The mean difference is significant
a. Adjustment for multiple comparisons: Bonferroni.

Figure 1

Means within math domain across judges for Triad 2



Statistically significant differences are included in Table 6 and means within judges across math domains are displayed in Figure 2. Judge 2 rated equations and expressions significantly lower than number systems and ratios and proportions. Furthermore, Judge 2 rated statistics and probability significantly lower than number systems. Judge 3 rated statistics and probability significantly lower from equations and expressions.

Table 6

Item Type Pairwise Comparisons by Rater

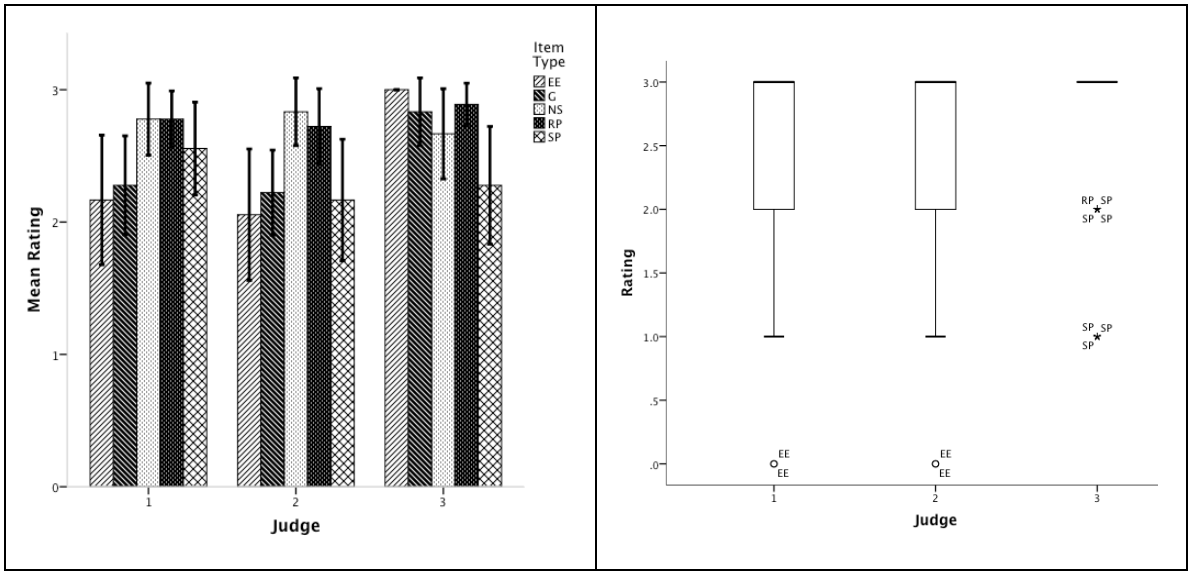
Judge	(I) Type	(J) Type	Mean Difference (I-J)	ρ	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	EE	G	-.111	1.000	-.757	.535
		NS	-.611	.078	-1.257	.035
		RP	-.611	.078	-1.257	.035
		SP	-.389	.893	-1.035	.257
	G	NS	-.500	.292	-1.146	.146
		RP	-.500	.292	-1.146	.146
		SP	-.278	1.000	-.923	.368
	NS	RP	.000	1.000	-.646	.646
		SP	.222	1.000	-.423	.868
	RP	SP	.222	1.000	-.423	.868
2	EE	G	-.167	1.000	-.812	.479
		NS	-.778*	.008	-1.423	-.132
		RP	-.667*	.038	-1.312	-.021
		SP	-.111	1.000	-.757	.535
	G	NS	-.611	.078	-1.257	.035
		RP	-.500	.292	-1.146	.146
		SP	.056	1.000	-.590	.701
	NS	RP	.111	1.000	-.535	.757
		*SP	.667*	.038	.021	1.312
	RP	SP	.556	.155	-.090	1.201
3	EE	G	.167	1.000	-.479	.812
		NS	.333	1.000	-.312	.979
		RP	.111	1.000	-.535	.757
		*SP	.722*	.017	.077	1.368
	G	NS	.167	1.000	-.479	.812
		RP	-.056	1.000	-.701	.590
		SP	.556	.155	-.090	1.201
	NS	RP	-.222	1.000	-.868	.423
		SP	.389	.893	-.257	1.035
	RP	SP	.611	.078	-.035	1.257

*. The mean difference is significant

a. Adjustment for multiple comparisons: Bonferroni

Figure 2

Means within judges across math domains for Triad 2



CHAPTER IV

DISCUSSION

The purpose of this study was to determine whether defined methodological components utilized in high complexity alignment models (i.e., Achieve and Webb) for use with large-scale assessments, could also apply to CBMs. It involved the application of these components with the 2012 easyCBM 6th grade middle school math items and the CCSS. The following section includes (a) a summary of the analyses and results, (b) limitations to the study, (c) interpretations of the findings, and (d) implications of the findings with respect to current application and future research.

Summary of Results

Overall reliability (consistency of judges relative to the total variance) of the judges was moderate as interpreted from the ICC statistic. This was further supported by a descriptive analysis of the Index of Agreement indicating that judges were in complete agreement or off by one on 77% of the items for both triads combined.

With respect to alignment, 72% of the items were rated as aligned to the CCSS by all judges when both triads were combined. Furthermore, in terms of a majority (two of three judges), about 92% of the items were aligned to the CCSS by at least two out of the three judges when combining triads. These findings suggest that the majority of math items and standards are aligned.

Of the eight items rated as not aligned for Triad 1, seven (i.e., 88%) were rated as aligned to a requisite skill reflecting a standard by at least two out of three

judges. Of the seven items rated as not aligned to the CCSS within Triad 2, 86% were rated as aligned to a requisite skill by two out of three judges. Overall, 98.9% of the 2012 6th grade easyCBM math items were rated as either aligned to a CCSS or a requisite skill reflecting a standard by at least two out of three judges.

Results from the ANOVA analysis indicated that some domains were rated significantly harsher than other domains. This may be attributed to judge or domain variance and may or may not reflect bias. Furthermore, the variance may be systematic and a reflection of the interaction between these two dimensions. That is, some judges may have rated specific domains differently than other domains regardless of whether or not it was a function of judge (e.g., expertise) or a function of domain (e.g., difficulty and content). This effect may necessitate the need for an increased focus on the selection of judges or judge training in future studies.

Limitations

The primary limitation in this study involved the difference in composition between the two triads. In particular, each triad was composed of different judges and different math items. Additional limitations included judge training and the recent shift from state to national standards.

Judge triads and item sets. This study utilized a partially nested study design allowing for a larger participant pool and a larger pool of math items therefore increasing the number of participants included in each analysis. Statistical differences were unique to each triad thus limiting the ability for generalizations across triads to be made. As a result of employing a nested study design that involved two independent judge triads and two unique items sets, caution is

recommended when interpreting the results (Anderson, Irvin, Alonzo, et al., 2012). For example, in Triad 1 statistics and probability received significantly harsher ratings than all other math domains while Triad 2 did not exhibit this issue. Judge differences (e.g., training and previous experience) or differences between the item sets might have accounted for the discrepancy between triads. Although judges and items were specific to each triad, analyzing and interpreting the results provided meaningful information regarding inter-rater reliability, item alignment, requisite skill alignment, and judge harshness for each triad individually.

Shift from state to national standards. The basic logic model for this study relied on judge qualifications and training. That is, with adequate qualifications and training, judge ratings would be reliable. Although qualified (see Appendix B) and trained (see Appendix C), the recent shift from state to national standards (i.e., CCSS) may have accounted for some inconsistency among judges within triads. Even though some judges had experience with standards, the degree to which they were (a) familiar with the CCSS, (b) able to generalize their knowledge across standard types, and/or (c) able to apply their understanding of the CCSS in practice might have effected judge ratings. While this limitation stems from each judge's prior experience with the CCSS, it is important to note that the specific standard was provided for each item rated within the DIR. In other words, judges did not need to select the standard to which an item aligned, but instead rate the degree to which an item aligned to a given standard within a given math domain.

Judge training. Unlike previous studies where judges underwent face-to-face training as a group (Resnick et al., 2003-2004; Webb, 1999), judges included in

this study participated in a live online training session. Furthermore, judges did not discuss their ratings on practice items as a group, thus limiting the opportunity for consensus building. Although providing training to judges focused on developing group consensus using sample items prior to rating the actual items might have increased the reliability between judges (i.e., rater calibration), group conformity may have become a confounding factor. That is, the possibility of individual members conforming to the group majority (Asch, 1956; Bond, 2005) as opposed to maintaining their independence and furthering the discussion towards consensus. As a result, the group might have reached consensus reflective of the majority therefore increasing reliability between judges while potentially decreasing validity of the results. For this dissertation, the training format limited potential issues related to group conformity. Furthermore, the online format allowed for judges to be selected from a broader geographic region thus limiting the potential for local bias.

Interpretations

Educators utilize CBMs to inform instructional decision-making with the intent of improving student learning. In particular, curriculum-base measures enhance the ability of educators to make data-driven decisions on instruction, curriculum, and student progress. La Marca (2001) stresses the importance of alignment (e.g., standards and assessment) within an accountability system in terms of methodological procedures and ethical necessity. In other words, to judge a stakeholder (e.g., student, teacher, etc.) using an assessment inadequately aligned to the standards students are striving to achieve, is unacceptable. Considering the high

stakes nature of testing in today's educational system with respect to students, teachers, schools, and districts, a comprehensive assessment system is critical. If CBMs were not only available for teachers to help inform instructional decisions, but also aligned to the same standards as state and national assessments, then the ability for teachers to predict student performance on large-scale assessments should increase.

Inter-rater reliability. The logic for this study was supported by the premise that if expert judges were selected and adequately trained, then judges and ratings would be reliable. Expert judges (i.e., content area specialists) make detailed distinctions about assessment items and standards that require specific knowledge about student learning (Webb, 1997a). Judges included in the extent data set utilized in my dissertation research had qualifications (see Appendix B) in one or more of the following: (a) experience teaching math, (b) district math coaching, (c) special education, (d) math endorsement, and/or (e) experience with the CCSS. Furthermore, three judges within each group rated the alignment of 90 items and standards, which is supported by Webb's (1999) finding that at least three raters are required to calculate adequate results on the degree of alignment between items and standards.

Judge training is critical in order to calibrate raters for inter-rater reliability and to develop understanding of rating values and rating procedures (La Marca, 2001). The judges included in this dissertation, participated in an online training format focused on (a) summative versus formative assessment, (b) easyCBM, (c) universal design, (d) the purpose of the study, (e) alignment procedures, (f) the

four-point alignment scale, and (g) use of the DIR tool. In the study conducted by Nese et al. (2010) judges also underwent online training. However, in addition, judges had the opportunity to discuss and justify their ratings on practice items during the training. This may account for the higher ICC statistic generated from the ratings of the three sixth grade judges in the Nese et al. research in comparison to the ICC statistics computed for the two judge triads within my study. In contrast, the judges included in my study had participated in a training format that controlled for group conformity by limiting judge interaction.

The study I conducted involving CBMs and the CCSS adds to the research literature in that it was conducted utilizing the same common methodological components previously applied to large-scale state assessments and standards. The findings suggest that the application of these components when studying the alignment of formative assessments and standards would likely yield reliable and meaningful results for practitioners in the field of education. In addition to the ICC results that indicated moderate inter-rater reliability, the Index of Agreement provided additional descriptive information regarding the reliability between raters in terms of judge agreements. These findings suggest that overall the judges were reliable.

Alignment of easyCBM math items and CCSS. The extant data set analyzed in this dissertation was a subset (i.e., 6th grade) of larger data set collected by Anderson, Irvin, Alonzo, et al. (2012). Furthermore, my analyses were conducted on a subset of the sixth grade data. The results of this dissertation add to the findings determined by Anderson, Irvin, Alonzo, et al. in which three grades were analyzed

using a MFRM model that adjusted for judge harshness and leniency. The researchers determined that 87% of the sixth through eighth grade items aligned to the CCSS when controlling for judge leniency and harshness within the model. Considering that my study did not adjust for judge harshness and only accounted for a portion of the sixth grade items, the findings show that 72% of the items judged within the two triads collectively were rated as aligned to standards by all three judges and 92% by at least two out of the three judges. These findings demonstrate that the majority of the items are aligned to the CCSS. Although beyond the scope of this dissertation and similar to the conclusions reached by Irvin et al. (2012), it is possible that some standards are over represented and others under represented by the number of items rated as aligned to each standard. This may be a consideration in the event that these items are utilized in future assessment development.

Alignment of easyCBM math items and requisite skills. Unlike large-scale assessments, CBMs are designed to provide educators with the information necessary to monitor student progress overtime. Including items in CBMs that align to requisite skills required by standards extends the reach of these assessments providing educators with a broad range tool for capturing students that are functioning at or below grade level. In a study conducted by Webb (1999) 60% to 70% of the math items received a near match and only 10% to 20% of the items received an exact match. Webb referred to this as a limitation of the study resulting from the use of inadequate measures for addressing items that partially aligned to the standards. One rater in the study suggested having a procedure for indicating near matches in addition to exact matches. This reinforces the importance of

considering the inclusion of items that reflect pre-requisite skills required by the standards when developing formative assessments in the future.

Anderson, Irvin, Alonzo, et al. (2012) showed that 99.6% of the sixth through eighth grade items aligned to either a standard or a pre-requisite skill required by the standard. This percentage was calculated by including all items that were rated by at least two out of three judges as either aligned to a standard or as assessing a requisite skill required by the standard. When employing this calculation to the Index of Agreement within my study the result is a comparable 98.9% lending additional support to the probability of successfully aligning CBMs and CCSS utilizing common methodological components previously applied to large-scale assessments and standards.

Providing judges the opportunity to rate items that are not aligned to standards as either aligned or not aligned to requisite skills of standards, would increase the depth of information about an assessment. Furthermore, designing assessments to include items aligned to requisite skills in conjunction with a judging protocol for rating these items, as in this research and other studies (Anderson, Irvin, Alonzo, et al., 2012; Irvin et al., 2012), would extend the reach and practical utility of the assessment. From an instructional standpoint, this would provide educators with information regarding the requisite skills a student needs to develop in order to access the skill and content expectations of the standards.

Judge harshness and leniency. As a result of the ANOVA analysis, significant main effects for Triad 1 and an interaction for Triad 2 were determined indicating that some domains were rated significantly harsher than other domains.

In Triad 1, for example, statistics and probability received a harsher rating than all other domains by all three judges. This outcome maybe a function of (a) the complexity of the statistics and probability domain, (b) the specific items written for the domain within the CBM, (c) judge expertise, or (d) judge interpretation of the statistics and probability standards within the CCSS. The study design was not capable of determining the specific cause of this finding and future research is needed specifically targeted to document the cause of standards, items, and judgment differences.

This research involved the use of a partially nested study design in that each triad included unique judges and items. As in the research conducted by Anderson, Irvin, Alonzo, et al. (2012), the nested design limited the ability to make generalizations across triads considering that judges and items differed between groups. Subsequently, the results of this research support the use of a fully crossed study design in future studies when conducting similar analysis (i.e., ANOVA, ICC, and Index of Agreement) in order to enhance the ability for researchers and practitioners to make more generalizable inferences regarding judge ratings.

Implications

The United States Department of Education (2010) has recently demonstrated continued support for upgraded assessments:

> Improved assessments can be used to accurately measure student growth; to better measure how states, districts, schools, principals, and teachers are educating students; to help teachers adjust and focus their teaching; and to provide better information to students and their families. (p. 11)

This in conjunction with the enactment of new national standards (i.e., CCSS) and state consortiums (i.e., SBAC and PARCC) charged with the design of new assessments aligned to these standards further demonstrates the growing movement in support of standards-based reform in the United States educational system.

In this climate of education reform, it is critical that teachers measure student growth based on the same set of standards to which students are held accountable (i.e., state and CCSS). Furthermore, it is critical to provide all stake holders with the information necessary in order for (a) teachers to make effective instructional decisions regarding students' performance towards meeting the standards, (b) students to understand how they are progressing towards proficient understanding of the standards, and (c) parents to have the feedback necessary to support their child in meeting the expectation of the standards. In order to achieve this, a comprehensive assessment system should include the alignment of national standards (i.e., CCSS) to (a) national assessments (e.g., SBAC and PARCC), (b) CBMs, and (c) requisite skills of standards. The inclusion of CBMs within a comprehensive system would help inform teachers of student growth towards meeting the standards measured by large-scale national assessments. In other words, "if CBMs and the state achievement test are both developed to be aligned to the same content standards, the predictive and criterion validity of the CBMs will be enhanced, effectively increasing teachers' instructional decisions" (Tindal & Nese, 2011, p. 39).

My study fills a void in the research literature by investigating the utility of defined methodological components from complex alignment models in application

with CBMs and common core standards. A statistically significant difference across judges within domains and within judges across domains reinforced the need to adequately train judges before beginning the rating process. The judges included in this dissertation underwent training in an online format allowing judge selection to reach beyond the local region thus minimizing local bias. Furthermore, the format provided the opportunity to capture a larger applicant pool therefore increasing the potential for a greater number of qualified applicants to select from. Conversely, the format limited the opportunity for the selected judges to build consensus prior to making judgments on the alignment of items and standards therefore limiting the potential issue of group conformity. In order to ensure reliability among judges while maintaining valid ratings, future research should focus on training models designed to maximize judge calibration while limiting the potential for group conformity. Furthermore, improved training models and recruitment processes are recommended in order to minimize the issue of judge harness and leniency. In Triad 2, for example, judge 3 rated equations and expressions and geometry more leniently than judge 1 and judge 2. In fact, judge 3 had a mean rating of 3 and standard deviation of 0 in equations and expressions. This is in contrast to the mean ratings for judge 1 and judge 2 of 2.17 and 2.06, respectively. This may be explained by differences in judge expertise and may potentially be mitigated by an improved training model that includes a review of grade level and domain specific content and standards for all judges involved within the study.

Future research is recommended on defining the criteria for determining the degree (e.g., low, moderate, and strong) to which the total number of items and

standards within an assessment are aligned. That is, what percentage of items aligned to standards is required within an assessment to constitute a ranking of low, moderate, or strong alignment for the assessment overall? Such criteria would be useful for test developers as well as state and local educational agencies in the development and adoption of specific assessments. This effort would also assist educators in better understanding the holistic effects of instruction.

While the application of large-scale assessments for determining student achievement is common practice, CBMs are commonly applied on a broader scale to measure student growth over time in order to help inform instructional decision-making in the classroom. A CBM, including requisite skills reflecting standards, aligned to the same standards to which the large scale assessment is also aligned would not only increase the predictive validity of the CBM with respect to how students are likely to perform on the large scale assessment, but also broaden the reach of the CBM to capture students functioning below grade level. This more universal approach would provide educators with additional information about a student's content knowledge of the requisite skills necessary for success on the specific standards they have not yet met. For example, although a student may be not be meeting the expectations of a specific standard, results of the CBM may indicate that they are proficient on 2 out of the 3 requisite skills required by the standard. This student would likely require a different instructional intervention then a student who was not meeting the expectations of the same standard and, in addition, had not demonstrated proficiency on any of the requisite skills required by that standard. From an instructional decision-making standpoint inclusion of

requisite skills within CBMs would provide educators with deeper knowledge about how to best provide individualized instruction and intervention based on the specific needs of each student. Additionally, because inclusion of requisite skills within a CBM would increase the sensitivity of the assessment for measuring students across a broader skill base (i.e., at or below grade level), the assessment would inform educators on how to strategically serve (i.e., intervention or enrichment) a boarder population of students within their classrooms. To further this field of study, future research is recommended on the alignment of requisite skills to standards within CBMs.

In conclusion, literature supports the application of alignment models and an associated set of common methodological components for use with large-scale assessments and standards. My research investigated whether the same common methodological components utilized by alignment models with large-scale assessments and state standards could also apply when studying the alignment of CBMs and standards. The results of this research support the use of these components when determining alignment of CBMs and common core standards. Considering the analysis conducted, the primary issue with this study came down to the application of a partially nested study design. To further enhance the generalizability of study results, future alignment studies involving the analyses employed in this research should consider the application of similar methodological components with CBMs and common core standards within a fully crossed study design.

APPENDIX A

RATER RECRUITMENT ADVERTISEMENT

BRT
behavioral research & teaching

About Us Publications

Featured Web Project:
cbmtraining
Register and login for free access to training on interventions in reading and mathematics as well as middle school concept-based instruction.
http://slds.ziptrain.com

Current Research Projects

8. Middle School Math Measures Alignment

Date: April 2012

Purpose: In 2010, the national Common Core Standards for Math were released to provide a unified set of expectations for developing mathematical skills across grade levels. The purpose of this study is to determine the degree to which middle school mathematics measures align with the Common Core Standards. We are looking for middle school mathematics teachers to examine items that were written to align with specific common core standards across grades 6-8. Study results can help strengthen alignment, provide a basis for evaluating math domain representativeness among test items and test forms, and enhance the validity of data score interpretations (e.g., instructional decisions made in response to student performance).

Teacher-student samples: We need approximately 13 teachers to judge the alignment of 300 middle school math items with the Common Core standards.

Description of Logistics: April 2012
1) Teachers are trained on alignment ratings via a webinar
2) Teachers individually rate the alignment of a sample of items through an online tool
Benefits or payments: We estimate that it will take teachers approximately 12 hours to fully participate in this study. We will compensate teachers for their time at a rate of $25 per hour for a sum of $300 – to be paid after completion of item ratings. The 12 hour time allotment also includes a mandatory 90 minute webinar training.

Training webinar: TBA

Sign-Up

APPENDIX B

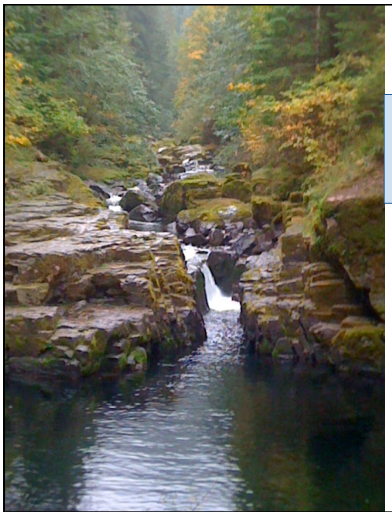
JUDGE QUALIFICATIONS

Triad	Judge	Qualifications
1	1	Bachelors in math, Masters in Curriculum Instruction Taught on, above, and below grade-level math in 7th and 8th grade. Facilitated online professional development for teachers in math. 5 years experience as district math coach
	2	Bachelors in economics 4 ½ years teaching experience at 8th grade Advanced Mathematics endorsement Involved in district alignment review of newly adopted curriculum
	3	Bachelors in Elementary Education, Masters in Special Education Taught 6th grade math for 2 years and 8th grade math for 6 years Experience with students with diverse learning needs (in SPED) Attended numerous workshops on CCSS
2	1	Bachelors in Elementary and Secondary Mathematics Elementary Math Specialist – provides PD for teachers Over 20 years teaching math experience, 7 years of PD experience Working to place CCSS math framework for district (5-8)
	2	Bachelors in business, Masters in Education Former community college math teacher District math coach, former math teacher in 8th and 9th grade Has run workshops on implementing CCSS
	3	BA w/concentrations in early child development, SPED, & Elementary Ed Instructs students w/IEPs as well as Gen Ed students Experience with academically & culturally diverse students Very familiar w/CCSS – attended district trainings

*Anderson, Irvin, Alonzo, & Tindal, 2012, pp. 18-19

APPENDIX C

TRAINING WEBINAR



BRT
behavioral research & teaching

easyCBM Middle School Math

Common Core State Standards Alignment Study

Big Picture

- Formative assessments are designed to inform teachers classroom decision-making.
- If the test items are not aligned with the content standards the teacher is instructing to, the results may lead to inaccurate decisions.
- **Alignment** – important for both summative *and* formative assessments, but needs to be re-conceptualized for formative assessments

RTI: A Complex system

A lot of moving parts have to come together to help teachers make appropriate decisions.

Reliability & Validity

- Scores on tests has to be sufficiently reproducible (i.e., a student shouldn't get a different score each time he or she takes it)
- Tests can't be biased for or against groups of students
- Alternate test forms have to be comparable in difficulty and content

And, why we're all here…

- Tests have to be **aligned** with the **instructional standards** and measure what they say they are measuring.
- We need your expert review to help us in developing these new items and ensure that they are aligned with the common core standards

Universal Design

How should we design a building so the widest range of populations possible can access it?

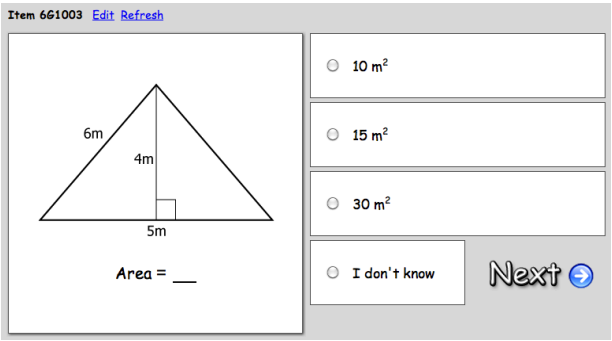
Universal Design for Assessment

- Considers **all** characteristics of test-takers.
- Precisely defined constructs.
- Accessible, non-biased items.
- Items amendable to accommodations.
- Simple, clear, and intuitive instructions and procedures.
- Maximum readability and comprehensibility.
- Maximum legibility of text, tables, figures, and illustrations

Thompson, Johnstone, and Thurlow (2002)

Example

Universal Design Features?



Why Common Core?

- easyCBM was designed to be accessible to the widest geographic range of students possible (i.e., national).
- With the majority of states committing to adopt the common core standards, we opted to develop a test aligned to those standards

Common Core Standards

Grade 6

Domains

- Ratios and Proportional Relationships
- Number System
- Expressions and Equations
- Geometry
- Statistics and Probability

29 standard divided among 5 objectives.

Common Core Standards

Grade 7

Domains

- Ratios and Proportional Relationships
- Number System
- Expressions and Equations
- Geometry
- Statistics and Probability

24 standards divided among 5 domains.

Common Core Standards

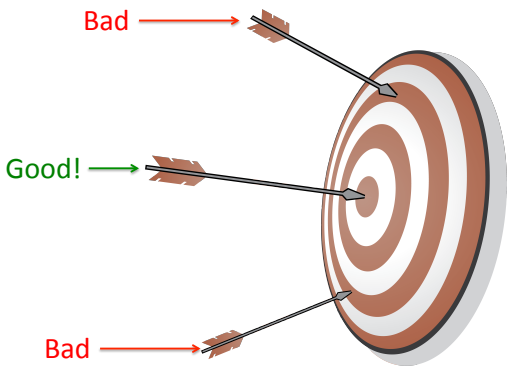
Grade 8

Domains

- Number System
- Expressions and Equations
- Functions
- Geometry
- Statistics and Probability

28 standards divided among 5 domains.

Importance of Alignment



Importance of Alignment

- Misaligned items create serious threats to the validity of interpretations and decisions made from the test

BUT

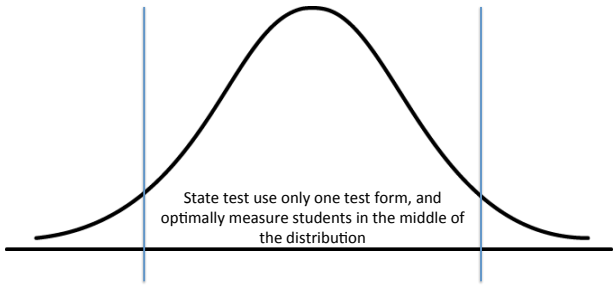
- Alignment of formative assessments needs to be conceptualized differently from large-scale summative assessments (e.g., state tests).

Formative versus Summative Alignment

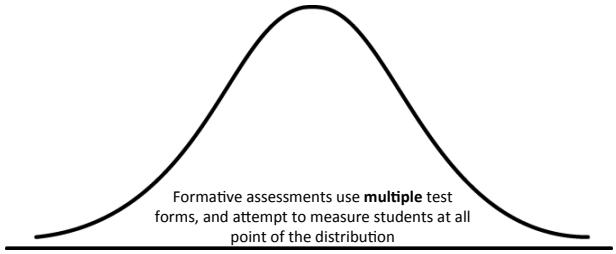
- Formative assessments need to reeeaaaaaaach.
- Often students performing below expectations are progress-monitored w/ formative assessments
- State tests are designed to determine "proficiency".

http://www.hahastop.com/pictures/Reaching_Monkey.htm

In other words…



In other words…



Formative versus Summative Alignment

Summative Assessments

- Generally viewed through Webb (1999)
- Important criteria include:
  - Categorical concurrence
  - Depth of knowledge
  - Range of knowledge
  - Balance of Representation

Formative Assessments

- Much less research overall, but still important
- Formative assessments are designed to have more "reach" to so low performing students can access the scale
- Generally have multiple forms

All of this is to say…

- Some items that were developed to reach those lower performing students may not directly align to a standard **BUT** may address a requisite skill TO the item.
- In other words, the item may address a skill that students must master prior to mastering the standard. This is good information to know!

In our alignment

- We will not only be asking you to judge whether an item aligns with a given standard, but also whether it addresses a requisite skill.
- If the item addresses the standard then you don't have to worry about whether it addresses a requisite skill or not (answer "no").
- Only if the item **does not address the standard** are we interested in whether it addresses a requisite skill.

Scales and Questions

- 4 point alignment scale

Aligned
- 3 = Item is **directly** aligned with the standard
- 2 = Item is **somewhat** aligned with the standard

Not Aligned
- 1 = Item is **vaguely** aligned with the standard
- 0 = Item has **no** alignment to the standard

Scales and Questions

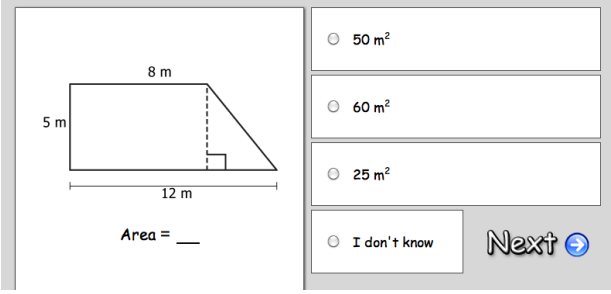
Dichotomous requisite skill item

- Does the item address and important requisite skill to the standard?
  - Yes
  - No

NOTE: If item is aligned, **always** choose "No" so you can finish with a "complete review". (More on this later)

Example: 3 (direct alignment)

Standard 6G1: Find the area of right triangles, other triangles, special quadrilaterals, and polygons by composing into rectangles or decomposing into triangles and other shapes; apply these techniques in the context of solving real-world and mathematical problems.

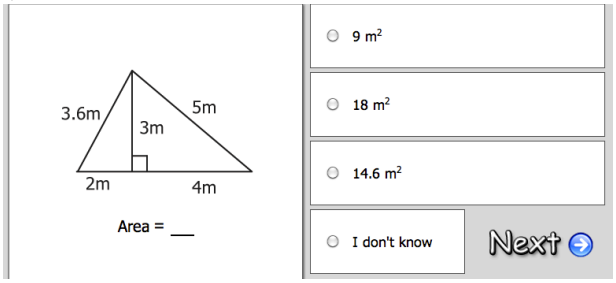


Addresses a Requisite Skill?

- No – because it addresses the standard.

Example: 2 (somewhat aligned)

Standard 6G1: Find the area of right triangles, other triangles, special quadrilaterals, and polygons by composing into rectangles or decomposing into triangles and other shapes; apply these techniques in the context of solving real-world and mathematical problems.

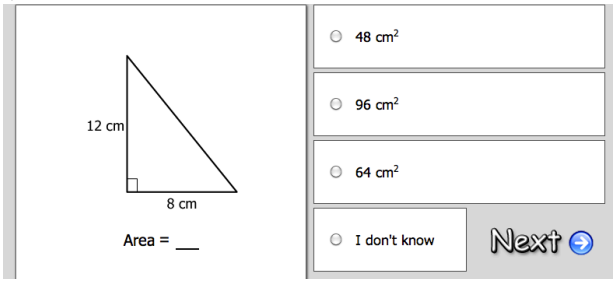


Addresses a Requisite Skill?

- No – because it addresses the standard.

Example: 1 (vaguely aligned)

Standard 6G1: Find the area of right triangles, other triangles, special quadrilaterals, and polygons by composing into rectangles or decomposing into triangles and other shapes; apply these techniques in the context of solving real-world and mathematical problems.

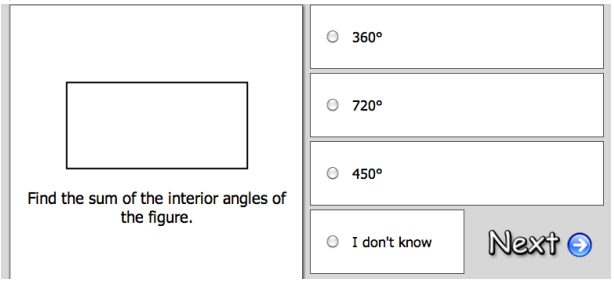


Addresses a Requisite Skill?

- Yes! – because you have to understand how to calculate the area of a triangle before you can find the area of complex shapes by decomposing it into quadrilaterals and rectangles.

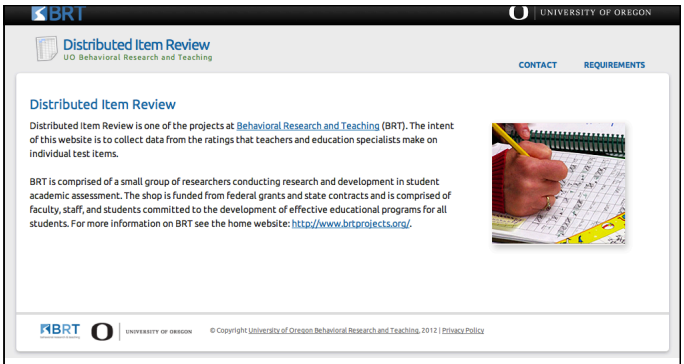
Example: 0 (no alignment)

Standard 6G1: Find the area of right triangles, other triangles, special quadrilaterals, and polygons by composing into rectangles or decomposing into triangles and other shapes; apply these techniques in the context of solving real-world and mathematical problems.



Addresses a Requisite Skill?

- No – because summing interior angles is unrelated to the standard.



Transition time

Web-based alignment tool

Web-based Alignment Tool

Distributed Item Review (DIR)

- A web-based system for presenting ***test items*** to ***experts*** across a ***broad geographic region*** so they can ***review*** them for important dimensions of ***bias, sensitivity, and alignment with standards***.

Your Role in the Study

1. Complete a short, 3-item proficiency training reviewing the DIR
2. Using the DIR, complete main review to determine the alignment of 270 grade-level easyCBM® math items to:

- *The corresponding* ***Common Core Standard***

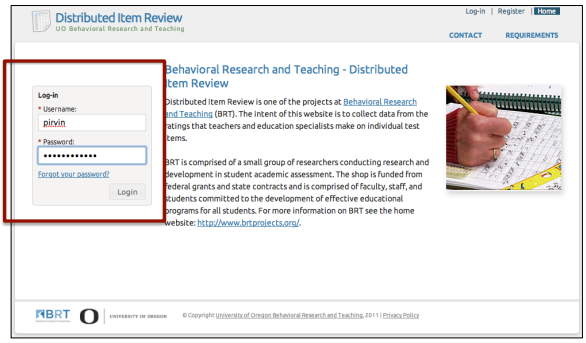
or

- ***Requisite skills*** necessary for mastery

Accessing the DIR

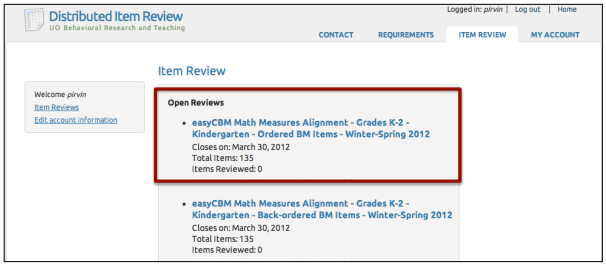
Log on to the DIR website:

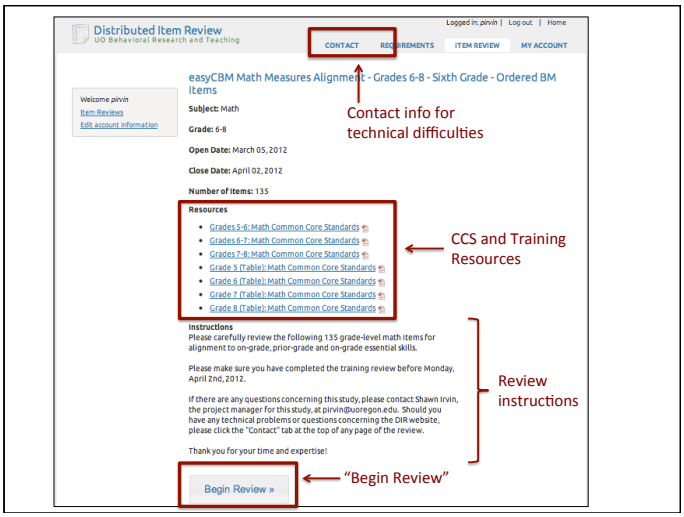
http://www.brtitemreview.com/shawndir



Accessing the DIR

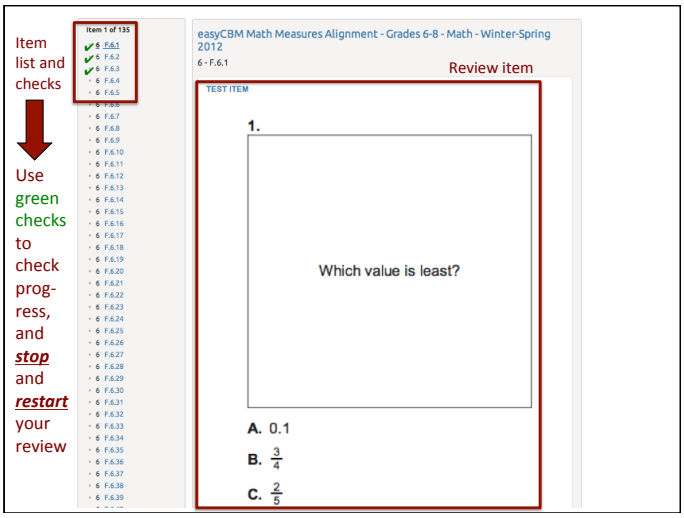
Access an open review by clicking on the title of your first review

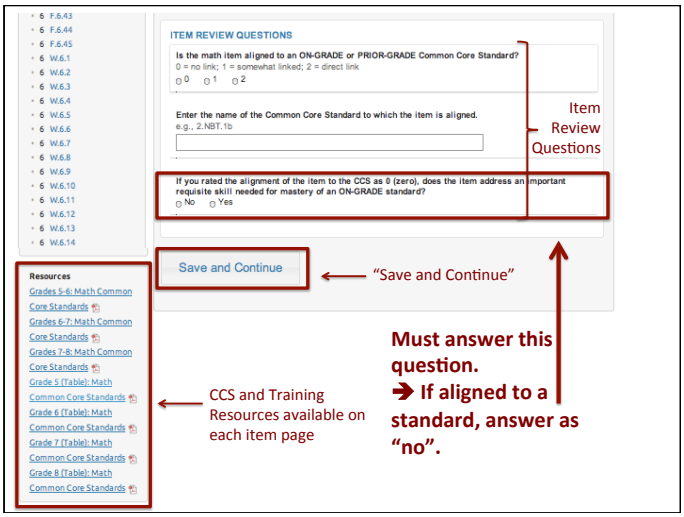




Reviewing Items on the DIR

1. Answer/complete ***all questions appropriate*** for a given item
2. Resources still accessible on item pages
3. ***CRITICAL!*** Click "Save and Continue" to save your responses and move to the next item
4. Check your progress, and stop and restart a review using "green checks"

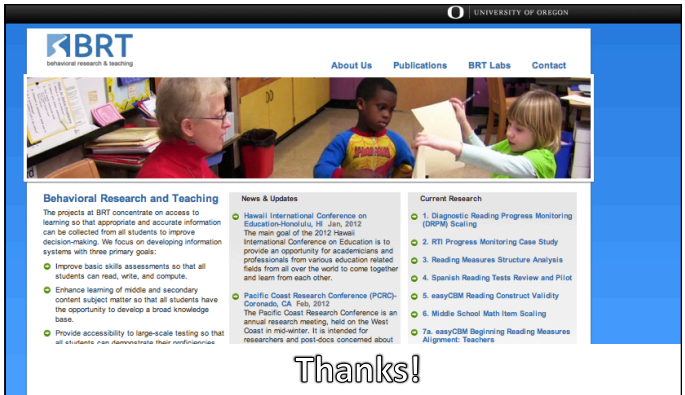

Item list and checks
Use green checks to check prog-ress, and ***stop*** and ***restart*** your review
Item 1 of 135
easyCBM Math Measures Alignment - Grades 6-8 - Math - Winter-Spring 2012
6 - F.6.1
Review item
TEST ITEM
1.
Which value is least?
A. 0.1
B. 3/4
C. 2/5


ITEM REVIEW QUESTIONS
Is the math item aligned to an ON-GRADE or PRIOR-GRADE Common Core Standard?
0 = no link; 1 = somewhat linked; 2 = direct link
Enter the name of the Common Core Standard to which the item is aligned.
e.g., 2.NBT.1b
If you rated the alignment of the item to the CCS as 0 (zero), does the item address an important requisite skill needed for mastery of an ON-GRADE standard?
No
Yes
Item Review Questions
Save and Continue
"Save and Continue"
Must answer this question.
➔ If aligned to a standard, answer as "no".
Resources
Grades 5-6: Math Common Core Standards
Grades 6-7: Math Common Core Standards
Grades 7-8: Math Common Core Standards
Grade 5 (Table): Math Common Core Standards
Grade 6 (Table): Math Common Core Standards
Grade 7 (Table): Math Common Core Standards
Grade 8 (Table): Math Common Core Standards
CCS and Training Resources available on each item page

Conclusions

- Each participant will use the DIR to rate the alignment of 270 middle school math items
  - If the item is not aligned, we also want to know if it addresses a requisite skill
- This work is critical to helping us develop an sound measurement system to inform teacher decision-making
  - Nearly 2.5 million students are in the easyCBM system! Your work will directly impact them!

Questions?



Thanks!

Contact info

DIR help: Shawn Irvin (pirvin@uoregon.edu)

Anything else: Daniel Anderson (daniela@uoregon.edu)

*Anderson, Irvin, Alonzo, & Tindal, 2012, pp. 34-55, 2012

APPENDIX D

INDEX OF AGREEMENT

For all strands, the number in each cell represents the number of judges who assigned the rating value specific to the associated column.

Strand: Expressions and Equations
Raters: 1, 2, 3

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6EE1004				3										
6EE1011				3										
6EE1024				3										
6EE2008				3										
6EE2019				3										
6EE3001				3										
6EE3016				3										
6EE3024				3										
6EE4010				3										
6EE5003										1	1	1	1	
6EE5011										1		2		1
6EE6005				3										
6EE6014				3										
6EE7005							1	2						
6EE7013							2	1						
6EE8008				3										
6EE8017				3										
6EE9009										1	1	1		1

Strand: Geometry
Raters: 1, 2, 3

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6G1007				3										
6G1015				3										
6G1023							2	1						
6G1037									1		1	1	1	
6G1038							1	2						
6G2003						1	2							1
6G2017				3										
6G2028						1	2							1
6G2043							2	1						
6G3007							2	1						
6G3013										1		2		1
6G3027				3										
6G3034										1		2		1
6G3046										1		2		1
6G4008				3										
6G4017				3										
6G4028										1	1	1		
6G4038				3										

Strand: Number Systems
Raters: 1, 2, 3

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6NS1001										1		2		1
6NS1012				3										
6NS1021							1	2						
6NS2006				3										
6NS2018				3										
6NS3006				3										
6NS3013				3										
6NS4002				3										
6NS4013				3										
6NS4021				3										
6NS5009										1		2	1	
6NS5017										1		2	1	
6NS6007							1	2						
6NS6013							1	2						
6NS7001										1		2		1
6NS7012										1		2		1
6NS8001				3										
6NS8012				3										

Strand: Ratios and Proportional Relationships
Raters: 1, 2, 3

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6RP1002				3										
6RP1013							1	2						
6RP1029							1	2						
6RP1037							1	2						
6RP1050							1	2						
6RP1056							1	2						
6RP2008							1	2						
6RP2014				3										
6RP2028				3										
6RP2036									1		1	1		
6RP2045				3										
6RP2058				3										
6RP3010				3										
6RP3014										1	1	1		1
6RP3025				3										
6RP3040				3										
6RP3049				3										
6RP3053										2		1		2

Strand: Statistics and Probability
Raters: 1, 2, 3

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6SP1002				3										
6SP1007				3										
6SP1022				3										
6SP1032				3										
6SP2005									1	1	1			2
6SP2017									1	2			1	2
6SP2027									1	1	1			2
6SP2034							2	1						
6SP3008		3												3
6SP3022		3												3
6SP3029						2	1							2
6SP4002										1	1	1	1	
6SP4016									1	1	1		1	1
6SP4023							2	1						
6SP4036										1		2	1	
6SP5007									1			2		1
6SP5013				3										
6SP5028				3										

Strand: Expressions and Equations
Raters: 4, 5, 6

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6EE1005				3										
6EE1020				3										
6EE2002							2	1						
6EE2016							1	2						
6EE2022							1	2						
6EE3011				3										
6EE3020							1	2						
6EE4005				3										
6EE4016										1	1	1		1
6EE5007				3										
6EE5018							2	1						
6EE6011										1	1	1		1
6EE7001									2			1		2
6EE7012									2			1		2
6EE8006										1	1	1		1
6EE8014											2	1		
6EE9006				3										
6EE9014											2	1		

Strand: Geometry
Raters: 4, 5, 6

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6G1008						1	2							
6G1022							2	1						
6G1026			3											
6G1043							1	2						
6G2001							1	2						
6G2010										1	1	1		1
6G2031				3										
6G2036				3										
6G2039				3										
6G3009							2	1						
6G3020										2		1		2
6G3025										1	1	1		1
6G3042				3										
6G4007				3										
6G4009				3										
6G4022							2	1						
6G4030										1	1	1		1
6G4046							2	1						

Strand: Number Systems
Raters: 4, 5, 6

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6NS1008				3										
6NS1019				3										
6NS2002						1		2						1
6NS2015				3										
6NS3001				3										
6NS3010				3										
6NS3021				3										
6NS4010				3										
6NS4017				3										
6NS5008				3										
6NS5013							1	2						
6NS6001							1	2						
6NS6009				3										
6NS6024						2	1							2
6NS7008										1		2		1
6NS7019				3										
6NS8007				3										
6NS8019							2	1						

Strand: Ratios and Proportional Relationships
Raters: 4, 5, 6

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned			Aligned	Not Aligned			Aligned		
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6RP1007							2	1						
6RP1016				3										
6RP1022							2	1						
6RP1036				3										
6RP1043				3										
6RP1055				3										
6RP2001				3										
6RP2011				3										
6RP2024				3										
6RP2039				3										
6RP2042				3										
6RP2060							1	2						
6RP3007							1	2						
6RP3019				3										
6RP3027				3										
6RP3031							2	1						
6RP3041				3										
6RP3059						1	2							1

Strand: Statistics and Probability
Raters: 4, 5, 6

	100% Agreement				Alignment to Standard (Off by 1)				Alignment to Standard (Off by > 1)				Aligned to Requisite Skill	
					Not Aligned		Aligned		Not Aligned		Aligned			
Item ID	0	1	2	3	0	1	2	3	0	1	2	3	No	Yes
6SP1010							1	2						
6SP1018							2	1						
6SP1030										2		1	1	1
6SP2004										1	1	1		1
6SP2010										1		2		1
6SP2019										1	1	1		1
6SP2035										2		1		2
6SP3005				3										
6SP3016										1		2		1
6SP3025				3										
6SP3032										1		2		1
6SP4010				3										
6SP4024							2	1						
6SP4027				3										
6SP5006							1	2						
6SP5017				3										
6SP5023						1	2							1
6SP5032		3											1	2

REFERENCES CITED

Achieve Inc. (2012). Retrieved January 2011, from http://www.achieve.org/publications/state_report_view

Alonzo, J., Ketterlin-Geller, L.R., & Tindal, G. (2006). Curriculum-based measurement in reading and math: providing rigorous outcomes to support learning. In L. Florian (Ed.), *The Sage Handbook of Special Education* (pp. 307-318). Thousand Oaks, CA: Sage.

Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. (2012). The Alignment of the easyCBM Middle School Mathematics CCSS Measures to the Common Core State Standards (Technical Report No. 1208). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Anderson, D., Irvin, P. S., Patarapichayatham, C., Alonzo, J., & Tindal, G. (2012). The Development and Scaling of the easyCBM CCSS Middle School Mathematics Measures (Technical Report No. 1207). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70*(9), 1-70. doi: 10.1037/h0093718

Bhola, D. S., Impara, J. C., Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, 22(3), 21-29.

Bond, R. (2005). Group size and conformity. *Group Processes Intergroup Relations, 8*(4), 331-354. doi: 10.1177/1368430205056464

Case, B., Jorgensen, M., Zucker, S. (2004). *Alignment in Educational Assessment*. Assessment Report. Pearson Education , Inc.

Christ, T. J., Scullin, S., Tolbize, A., & Liban, C. L. (2008). Implications of recent research: curriculum-based measurement of math computation. *Assessment of Effective Intervention, 33*, 198-205. doi: 10.1177/1534508407313480

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review, 33*, 234-248.

Council of Chief State School Officers (CCSSO). (2002, September). *Models for alignment analysis and assistance to states.* Washington, DC: Author.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184-192. doi: 10.1177/00224669030370030801

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics. *The Journal of Special Education, 41*(2), 121-139. doi: 10.1177/00224669070410020101

Fuchs, L. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*, 188-192. Retrieved from web.ebscohost.com

Fuchs, L. S., Fuchs, D., & Zumeta, R. O. (2008). A curricular-sampling approach to progress monitoring: Mathematics concepts and applications. *Assessment for Effective Intervention, 33*, 225-233. doi: 10.1177/1534508407313484

Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *The Journal of Special Education, 36*(2), 102-112. Retrieved from http://sed.sagepub.com

Irvin, P. S., Park, B. J., Alonzo, J., & Tindal, G. (2012). The Alignment of the easyCBM Grades 6-8 Math Measures to the Common Core Standards (Technical Report No. 1230). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Jiban, C. L., & Deno, S. L. (2007). Using math and reading curriculum-based measurements to predict state mathematics test performance: Are simple one-minute measures technically adequate? *Assessment for Effective Intervention, 32*(2), 78-89. doi: 10.1177/15345084070320020501

La Marca, Paul M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7(21). Retrieved January 13, 2012 from http://PAREonline.net/getvn.asp?v=7&n=21.

Messick, S. (1989). *Validity*. In R. L. Linn (Editor), Educational Measurement (3rd Edition). New York: American Council on Education – Macmillan Publishing Company.

Nese, J. F. T., Lai, C. F., Anderson, D., Park, B. J., Tindal, G., & Alonzo, J. (2010). *The alignment of easyCBM math measures to curriculum standards* (Technical Report No. 1002). Eugene, OR: Behavioral Research and Teaching, University of Oregon. Retrieved from http://www.brtprojects.org/publications/technical-reports

Nichols, D. P. (1998). Choosing an intraclass correlation coefficient. *SPSS keywords*, *67*.

No child left behind act of 2001, P.L. 107-110, 115 Stat. 1425.

Resnick, L. B., Rothman, R., Slattery, J. B., & Vranek, J. L. (2003-2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, 9(1 & 2), 1-27).

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin Alternate Assessment. *The Journal of Special Education, 38*(4), 218-231.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity curriculum-based mathematics measures. *School Psychology Review, 31*, 498-513.

Tindal, G. (2005). Alignment of alternate assessments using the Webb system, in *Aligning Assessment to Guide the Learning of All Students*. Washington, D. C. Council of Chief State School Officers

Tindal, G. & Nese, J. F. (2011). Applications of curriculum-based measures in making decisions with multiple reference points. *Advances in Learning and Behavioral Disabilities*, 24, 31-58. doi: 10.1108/S0735-004(2011)0000024004

Tindal, G., Nese, J., & Alonzo, J. (2009). *Criterion-related evidence using easyCBM ® reading measures and student demographics to predict state test performance in grades 3-8* (Technical Report No. 0910). Eugene, OR: Behavioral Research and Teaching, University of Oregon. Retrieved from http://www.brtprojects.org/publications/technical-reports

U.S. Department of Education (2010), ESEA Blueprint for Reform, *Office of Planning, Evaluation and Policy Development*, Washington, D.C., 2010. Retrieved November 2011, from http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

Webb, N. L. (1997a). *Research monograph No. 6. Criteria for alignment of expectations and assessments in mathematics and science education.* Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1997b). *Determining alignment of expectations and assessments in mathematics and science education. NISE Brief.* National Center for Improving Science Education, University of Wisconsin, Madison.

Webb, N. L. (1999). *Research monograph No. 18. Alignment of science and mathematics standards and assessments in four states.* Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2002). *An analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Webb, N. L. (2007a). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education, 20*, 7–25.

Webb, N. L. (2007b). Aligning assessments and standards. *Wisconsin Center for Education Research.* Retrieved from http://www.wcer.wisc.edu/news/coverstories/aligning_assessments_and_standards.php.