Comparing Alignment of a State Test and District Formative Assessments with State Content Standards using Three Methods

Elizabeth L. Greive

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the School of Education (Educational Psychology, Measurement, and Evaluation).

Chapel Hill
2012

Approved by:

Dr. Gregory J. Cizek

Dr. William Ware

Dr. Gary Henry

Abstract

ELIZABETH L. GREIVE: Comparing Alignment of a State Test and District Formative Assessments with State Content Standards using Three Methods
(Under the direction of Gregory J. Cizek)

Alignment between tests and content standards is an essential piece of validity evidence. This study examines the alignment of a district, fourth grade, mathematics, formative assessment to state content standards using three commonly-used methods, including the Webb method, Achieve method, and the Surveys of an Enacted Curriculum (SEC) method. Alignment sessions were conducted separately for each method by the researcher with educators and graduate students. Findings across methods suggest that each alignment method highlights different components of alignment. Suggestions are made for integrating the essential pieces of alignment evidence across methods. The alignment of the district formative assessment to the state standards is compared to the alignment of the state test and the state content standards using the Webb method. The findings comparing the alignment of the state test and the formative assessment to the state content standards using the Webb method indicate that the state test is more aligned to the standards than the district formative assessment.

Table of Contents

List of Tables

List of Figures

Comparing Alignment of a State Test and District Formative Assessments with State Content Standards using Three Methods

The 2001 passage of No Child Left Behind (NCLB; Public Law 107-110) heightened the focus on accountability by mandating that states develop rigorous content standards and standardized tests to measure students' academic progress. This emphasis on student achievement and teacher quality continues today with the initiatives of the Race to the Top Fund, the adoption of the Common Core State Standards, and the development of the Common Core Assessments in many states (Common Core, 2011; Common State Assessments, 2011; U.S. Department of Education, 2011). Objective criteria for assessing the alignment of these standards and assessments are not widely agreed upon. Because content standards and performance standards vary across states, cognitive demands and stringency of passing requirements have remained idiosyncratic across the nation. With the current movement toward national standards and assessments, the identification of agreed upon alignment criteria is critical for the successful implementation of the Common Core.

Instructional coherence and a common framework are necessary components for wide-scale educational reform (Stuart & Rinaldi, 2009). Teachers and schools need to know where to focus their attention in order to drive instructional improvement. If the content of the instructional program, the state standards, and the state assessment contradict one another, more pressure and stress are created for the teachers and students. NCLB started in the 2005-2006 school year by requiring schools to administer state summative tests in reading and mathematics in grades three through eight, and once in high school. All students

were meant to meet state-defined criteria for proficiency by the 2013-2014 school year. Under NCLB, schools, districts, and states are required to demonstrate that the number of students achieving the defined levels of proficiency increases each year (known as Adequate Yearly Progress or AYP) until all students have reached proficiency. If schools, districts, and states are not able to meet the set AYP goals, a system of consequences exists, such as loss of funding or restriction of local decision making and control. If these high-stakes decisions are made based on tests that are not aligned to the instruction and standards, there could be serious consequences for schools, including mislabeling of student performance and teacher job loss (Roach, Niebling, & Kurz, 2008).

Because NCLB requires alignment of tests to state standards, research on alignment has emerged in the last decade (Bhola, Impara, & Buckendahl, 2003; Flowers, Browder, & Ahlgrim-Delzell, 2006; Kurz, Elliot, Wehby, & Smithson, 2010; Martone & Sireci, 2009; Resnick, Rothman, Slattery, & Vranek, 2004; Roach, Elliot, & Webb, 2005; Roach et al., 2008; Webb, 2007). In the past, results across the three methods produced varying results pertaining to the alignment of state standards and state tests (Bhola et al., 2003; Kurz et al., 2010; Lui, Zhang, Liang, Fulmer, & Yuan, 2008; Polikoff, Porter, & Smithson, 2011; Roach et al., 2005). The findings suggested that alignment varies across states, with no one state demonstrating exemplary alignment. A press release by the American Federation of Teachers found only 11 states had strong content standards and tests aligned to those standards (AFT Teachers, 2006). Different criteria and methodologies make alignment unclear, subjective, and at times contradictory between different sources (Lui et al., 2008). However, a published comparative state alignment study within the last five years could not be found at the time of this thesis.

Although NCLB mentions alignment dozens of times and states are mandated by NCLB to conduct alignment studies, the literature on the extent to which standards-based reform has resulted in coherence of standards, instruction, and assessments is thin (Polikoff et al., 2011). Webb (1997) defined alignment as "the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students' learning what they are expected to know and do" (p. 4). Several alignment models exist at varying levels of complexity (Bhola et al., 2003). These models will be examined and explained in depth in the next section of this thesis, but it is important to note that high complexity models, like the Webb and Achieve models, include several interrelated dimensions, such as content match, depth match, emphasis, and performance match. Moderately complex models, like the SEC model, look at the relationship between cognitive demand and topic. Low complexity models, which were widely accepted before NCLB passed but are no longer commonly used, focus only on objective and item matching without accounting for cognitive demand and other criteria (Bhola et al., 2003). The Webb, Achieve, and SEC models are the three models examined in this study.

Because the three most commonly used and widely accepted alignment methods have not yet been applied simultaneously in a single study, differential aspects and the utility of alignment results across methods cannot accurately be described (Martone & Sireci, 2009). For the purpose of this thesis, a state's test will be examined through the lenses of the three most widely used alignment methods identified by the Council of Chief School Officers (CCSSO) as preferred models to examine alignment. These methods include the Webb method, the Achieve method, and the Surveys of an Enacted Curriculum (SEC) method (Roach et al., 2008). The CCSSO is a nonpartisan, nationwide, nonprofit organization of

public officials who head departments of elementary and secondary education in states and other equivalent agencies. The organization provides leadership, advocacy, and technical assistance on major education issues (Lui et al., 2008). The goal of this research is to identify the strengths and weaknesses of the three models suggested for use by the CCSSO. Because the cohesiveness of standards, instruction, and assessments is essential for student learning to take place, this study will address three questions:

- Do district formative assessments align with the state content standards as measured by the Webb method, the Achieve method, and the SEC method?
- In what ways do the three methods of aligning the formative assessments to the state content standards produce different results?
- What is the alignment of a state test with the state's content standards as measured by the Webb method?

Components of each alignment method including the Webb, Achieve, and SEC shed light on the complex picture of alignment. Without strong alignment, accurate inferences about student academic performance cannot be made, and achievement of goals is unlikely. According to Herman, Webb, and Zuniga (2007), alignment is a validity issue. In order to provide content-based validity evidence, the assessments must work coherently with curriculum and instruction. Validity refers to the degree to which evidence and theory support the interpretations that are suggested by the test scores (Messick, 1989). In order to make a valid inference about a student’s ability, the proposed uses of the test must be clearly stated along with sufficient evidence of validity. These evidences include test content, relationships to other variables, internal consistency, response process, or test consequences. By examining content and cognitive demand, alignment studies provide validity evidence by

linking test items to academic standards. Evidence of strong alignment heavily influences the rationale and justifies the use of a particular test for a specific purpose.

This research will investigate the alignment of the formative and summative assessments to a state's content standards. In order for student learning to occur, educational standards, instruction, and assessments must work in accordance with one another. Using formative assessments to track student achievement throughout the year, teachers are better able to understand students' needs and design effective and differentiated instruction in preparation for the comprehensive summative state test, which is mandated for accountability purposes through Title I and the NCLB (Webb, Herman, & Webb, 2007). The formative assessments used throughout the year must be focused and information-rich in order to accurately formulate a picture of an individual student's academic achievement and readiness for the annual summative test. Therefore, questions regarding the quality of formative and summative assessments are frequently asked by educators, policy makers, administrators, and parents.

The level of alignment between the formative assessments and the content standards directly leads to opportunity to learn and the possibility of high achievement on the state summative test (Lui et al., 2008). Opportunity to learn is defined as adequate coverage, exposure, emphasis, and quality instruction related to the content covered in the test (Lui et al., 2008). The alignment of a state's summative test and district-level formative assessments to the content standards is essential for accurate inferences about student performance to be made by teachers, parents, district leaders, and state representatives. The system of standards, instruction, and assessments must work together to focus vested individuals on what students

should be able to know and do in order to be successful at any particular point in their PreK-12 experience.

In today's standards-based and achievement-driven context, alignment must clearly communicate the degree to which assessments yield results that provide accurate and detailed information about students' achievement in regards to academic content standards (Martone & Sireci, 2009). The assessment must adequately cover the content standards with the appropriate depth, reflect the emphasis of the content standards, provide scores that cover the range of performance standards, allow all students an opportunity to demonstrate their proficiency, and be reported in a manner that clearly conveys student proficiency as it relates to the content standards (Martone & Sireci, 2009). Another important consideration to standards-based reform is high quality standards. Current state standards have been suggested to cover a wide variety of topics and content, but not place much emphasis or instructional intensity on content (Lui et al., 2008). In other words, the U.S. state standards are each extensive in breadth, but limited in depth (Roach et al., 2008).

In a review of the three commonly-used alignment methods, the Webb method provided the strongest quantitative information for evaluating alignment on multiple criteria, which is why the Webb method was chosen to examine the alignment of the summative test and formative assessments (Polikoff et al., 2011). The Achieve method provided the most useful narrative summary of alignment. The SEC method provided applicability to instructional issues and took instruction into account along with standards and assessments; however, it was the least detailed evaluation of alignment. The next section of this proposal will review the components and applications of the Webb, Achieve, and SEC methods of alignment in depth and will end with issues related to their use and projections for the future

of alignment research and applications. Whereas the summary included in this section is designed to be sufficient, the reader's understanding of the methods is essential for the benefit of interpreting the future results. Summaries of all of the methods can be found at http://programs.ccsso.org/projects/Alignment_Analysis/.

Webb Alignment Method

The Webb alignment procedure is conducted in two phases, standards review and items review. During training, five to eight content-area experts are trained on the method including the operational definitions of a general standard, which is composed of a specific number of goals, which are comprised in turn of specific objectives (Webb, 2007). The reviewers are trained on depth of knowledge (DOK) levels and are encouraged to write notes about the quality of the standards or the items if there is an extraneous source of challenge in the item (Webb, 2007). Extraneous source of challenge includes student knowledge that is necessary to answer the item but is not relevant to the tested standards. For example, if the language in a mathematics word problem is not written at an appropriate grade level or a graph necessary to answer an item is not clearly labeled, the rater would note this as an extraneous challenge. Mean and standard deviations are reported for all reviewers' ratings and discussed (Webb, 2007).

After training and during the first phase of the alignment method known as standards review, the reviewers examine the content standards and assign an appropriate DOK level for each objective. According to Webb (2007), DOK levels measure the level of cognitive demand and are labeled with Level 1 (recall), Level 2 (skill/concept), Level 3 (strategic thinking), or Level 4 (extended thinking). During the second phase known as items review, the reviewers examine the test items, code the items with an appropriate DOK level for the

item, and link the items to corresponding curriculum objectives. The assessment is then judged along four dimensions: depth of knowledge consistency, categorical concurrence, range of knowledge consistency, and balance of representation. The training materials for this study were retrieved from the free web-based version of the Webb method, called the Web Alignment Tool (WAT), which is available at http://wat.wceruw.org/index.aspx. The current study uses a paper and pencil version of the WAT which, perhaps more time consuming on the part of the researcher, avoids potential loss of data and technological malfunction.

Depth of knowledge consistency. In Webb's model, depth of knowledge consistency requires that at least 50% of the test items corresponding to a given standard should be at or above the DOK level of the items' corresponding objective (Webb, 2007). If the standard has between 40% and 50% of the items at or above the DOK levels of the objectives, then it is reported that the criterion is weakly met (Webb, 2007). The rationale for this cutoff is that if three of the six, or 50% of the items, are at or above the DOK level of the standard, then in order for a student to achieve a proficient score on the overall standard, he or she would be required to answer correctly at least one of the items at or above the DOK level of the standard (Webb, 2007). According to Webb (2007), DOK level 1 (recall) includes recalling information such as a fact, definition, term, or a simple procedure. DOK level 2 (skill/concept) includes the engagement of some mental processing beyond a habitual response and requires students to make a decision about how to approach the problem. DOK level 3 (strategic thinking) requires some reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. DOK level 3 typically requires students to explain their thinking, which should be complex and abstract. DOK level 4 (extended

thinking) requires complex reasoning, planning, developing, and thinking, which most likely occurs over an extended period of time, and typically requires developing and proving conjectures, designing, and conducting experiments, or critiquing experimental designs (Webb, 2007).

Categorical concurrence. Categorical concurrence examines the extent to which at least some element of each standard appears on the assessment. Webb (1997) specified that at least six items on the assessment should address each standard in order to indicate acceptable categorical concurrence. A hit is used to designate that a reviewer has mapped an assessment item to an objective (Webb, 1997). Each item can have up to three hits, each to a different objective. The average number of hits assigned to each standard is meant to describe the weight of information from the assessment in making judgments about a student's performance. The rational for the six-item cutoff per standard was developed using a procedure by Subkoviak in 1988 (Webb, 2007). Assuming the reliability for each item is 1.0, the estimated six items would provide an agreement coefficient of 0.63, which is somewhat acceptable according to Webb (2007). Webb does not encourage reporting scores on subscales of the test or by objective, because this agreement coefficient would be mediocre (Webb, 2007). The reliability of 1.0 for each item assumes that the items are well designed, written clearly, and function similarly across the population (Webb, 2007)

Range of knowledge. Range of knowledge (ROK) suggests that at least 50% of the objectives under any curriculum standard should have at least one matching item (Webb, 2007). This ensures that, on average, at least half of the objectives under each standard are included on the test, and that student knowledge is measured on at least half of the content from a given curriculum standard. ROK correspondence is used to judge whether a

comparable span of knowledge expected of students by a standard is the same as the span of knowledge that students need to correctly answer the assessment items. Having at least one item for each objective for at least half of the objectives under a standard provides a decision rule that ensures that the assessment is measuring some breadth in content knowledge and is at least sampling half of the most important partitions of content identified by the objectives (Webb, 2007). This assumes that a student's knowledge should be tested on at least half of the domain of knowledge for a standard. This increases the likelihood that students will need to demonstrate knowledge on more than one objective per standard to achieve a minimal passing score (Webb, 2007). If 50% of the objectives have a matching item for a given standard, the ROK is met, but if only 40% to 49% have a match, ROK is weakly met (Webb, 2007).

Balance of representation. Balance of representation takes into account how the hits are distributed among the objectives under a standard. A hit is defined as a match between an objective and an item. Balance of representation is calculated by summing the differences between the total number of objectives hit under a standard and the proportion of the hits assigned to each objective to the total number of hits for a standard. This calculation is subtracted from one. This formula results in the balance of representation index, which was formulated by Webb (1997). The index calculates the degree to which the distribution of hits for objectives within each standard is balanced across objectives under each standard, taking into account only objectives that have hits. The formula for the balance of representation index is

$$Balance = 1 - (\sum_{i=1}^{K} \left| \frac{1}{O} - \frac{I_k}{H} \right|)/2$$

where O is the total number of objectives hit for the standard, I_k is the number of items hit corresponding to objective k, and H is the total number of items hit for the content standard.

The index ranges from 0 to 1. A balance representation index of one, or near one, indicates that the assessment is well balanced across the objectives within a particular standard. A balance of representation index of zero, or near zero, indicates that the assessment is unbalanced in the distribution of hits (Webb, 2007). Assessments that are unbalanced lead to biased inferences about students' ability. The index only considers objectives that have at least one hit. Therefore, objectives that do not have a matching item are not taken into the equation of balance. If all of the items assigned to a standard are evenly distributed among the objectives, then the index will be one. For example if a particular standard has 10 objectives, but only 7 objectives have hits and there are 12 hits distributed across the 7 objectives such that one objective has four hits, five objectives have one hit, and one objective has three hits; the formula for the standard would be calculated as such: 1 – (|1/7-4/12| + |1/7-1/12| +|1/7-1/12| +|1/7-1/12| +|1/7-1/12| +|1/7-1/12| +|1/7-3/12|)/2 such that 1 - 0.595/2 = 1 - 0.2975 = 0.7025. Index values greater than 0.7 are deemed acceptable and 0.6 to 0.7 indicate that balance is weakly met (Webb, 2007). According to Webb (2007), seven tenths was chosen as a cutoff because it indicates that the items are distributed among all of the hit objectives to at least some degree (e.g., every objective with a hit has at least two items).

Webb (2007) pointed out some possible issues with the alignment tool, questioning all of his developed cut scores, suggesting that the accuracy of the scores is dependent on the coherent structure, clarity, and quality of the standards, as well as questioning the progression of cognitive functioning throughout grade levels, for which the Webb method currently does

not account. The Webb method was built on five different dimensions to understand the degree of alignment including content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability (Webb, 1997). While the area of focus for the Webb tool is content, the Webb method is comprehensive in its item and objective level analysis, its view of alignment through four quantitative dimensions, and the proposed guidelines for acceptable minimum levels. Webb's method does not take into account objectives that do not have hits; therefore, its alignment measures of range and balance may overestimate alignment (Martone & Sireci, 2009).

Applications of the Webb Alignment Method

Recent applications of the Webb alignment model have examined the importance of inter-rater agreement, differences across raters based on job title, the alignment of alternate assessments for students with special needs, and the alignment of assessments and standards across transitional years, such as high school to college or preschool to kindergarten. This section will provide a brief review of this literature relating to the current applications of the Webb method.

Because DOK levels and objective matches can relate to the unique perspectives of the raters, the role of reviewer agreement can potentially influence results of alignment studies (Webb et al., 2007). Three approaches were considered by Webb et al. (2007): 1) reviewer agreement was not specifically addressed, 2) a bare majority or more than half of reviewers needed to agree on the content or depth of knowledge match in order for the match to be included in the study, and 3) a clear majority or two-thirds of the raters needed to agree upon ratings before matches were entered into analysis. High school state tests and standards from Tennessee, California, and Michigan were analyzed. In the first approach, the average

of all reviewers' ratings was used. The results were considerably different when a minimum level of reviewer agreement was required. The authors suggested that requiring agreement on the objective and item level was too strict for categorical concurrence. Requiring a minimum level of reviewer agreement for an objective matching an item resulted in mixed results across the different tests and standards, varying from none of the standards meeting the ROK criterion to all of the standards meeting the ROK criterion. Whereas results for different tests and standards were mixed, the authors suggested that taking into account the reviewer agreement reduced the number of items and objectives taken into account and generally provided weaker alignment evidence than a case where all reviewers' ratings were considered. Categorical concurrence and ROK were most influenced when taking into account reviewer agreement. When compared to each other, the selection of bare or clear majority made little difference in the results. Standards need to be clear, detailed, and complete in order to match test items and the average of all reviewers' ratings should be used in determining item and objective matches and DOK levels.

Differences in the raters' job titles may influence alignment results. Teachers, test publishers, and college faculty tend to make item and objective matches and rate DOK labels differently. A study of the Nebraska English Language Arts and Reading (ELAR) standards and assessments found that teachers consistently found less alignment compared to test publishers, who viewed more standards at being met by their assessments (Bhola et al., 2003). Teachers sometimes rated items differently than higher education faculty, who viewed items as less multi-faceted and of lesser cognitive demand compared to teachers (Martone & Sireci, 2009). Test publishers tend to find the most robust and multi-faceted alignment results with the Webb method, compared to teachers, who find more robust alignment than college

faculty. The 2001 Golden State Examination in High School Mathematics and the 1997 University of California Statement on Competencies in Mathematics Expected of Entering College Students (which is intended to give a clear picture of what students need to know and be able to do in order to be successful in college) were compared. To examine consistency among raters, kappa coefficients were calculated to determine inter-rater reliability, and generalizability analyses with items crossed with raters were conducted. Looking at categorical concurrence, kappa coefficients were .55 and .58 for faculty and teachers respectively. Teachers rated more items as multidimensional, which is defined as an item matching to more than one objective, than faculty. On average, teachers rated 45% of the test as multidimensional, and faculty rated only 26% of the test as multidimensional. Teachers tended to rate DOK higher than faculty. DOK is a difficult feature to rate because students' developmental levels and teachers' instructional experiences may play a role. Tremendous variation can result across 6-rater subsets of 20 raters. With modest training, 20 raters can achieve acceptable levels of agreement (Bhola et al., 2003).

Perhaps because it is one of two most complex models available, and it is available online for no cost, the Webb method has been widely used for various alignment studies across a variety of age groups and populations (Bhola et al., 2003; Brown & Niemi, 2009; Flowers et al., 2006; Polikoff et al., 2011; Roach et al., 2005; Roach et al., 2010). The Webb method has been applied to alternate assessments for students with disabilities. Roach et al. (2005) examined the alignment of the Wisconsin alternate assessment for students with special needs and found that the alternate assessment overall met the specified criteria for mathematics, ELAR, and social studies. Science demonstrated the weakest alignment, including only 13% of the academic standards. On the other hand, Flowers et al. (2006)

applied the Webb method to three state alternate assessments for students with special needs in mathematics and ELAR and found that none of the alternate assessments met the recommended levels of alignment criteria. Two of the three assessments were portfolio-based and one was performance-based. Because the performance-based assessment demonstrated the best alignment to the state standards, the Webb method may not work well for alignment information regarding portfolio-based assessments and works slightly more than not at all for constructed-response, performance-based assessments (Flowers et al., 2006).

Studies have been conducted comparing the alignment of assessments and standards vertically across transitional years for high school to college and preschool to kindergarten. Brown and Niemi (2009) found that the California Standards Tests in ELAR demonstrated sufficient alignment with the California Community College placement objectives as measured by two placement exams used in community colleges; however, the mathematics test showed adequate alignment values only with respect to DOK consistency and balance of representation, falling short in categorical concurrence and ROK. These findings for mathematics indicated that high school standards and assessments are not consistent with college expectations for success (Brown & Niemi, 2009).

Using a modified Webb alignment method to examine the Indiana Kindergarten content standards and the items on the Indiana Standards Tool for Alternate Reporting (ISTAR) which is available in five versions to monitor development throughout ages birth to five, Roach et al. (2010) found that the ROK was adequate across assessments, but the mathematics ISTAR did not meet the criteria for DOK consistency. The ROK expected in the battery of assessments was inconsistent and did not progressively build toward the

kindergarten standards across the assessments. DOK consistency was weakly met across assessments.

These studies show the breadth and depth of the application of the Webb method and some of its shortcomings. The Webb model provides in-depth quantitative descriptive information that can be used to provide evidence of test validity. Despite limitations, the Webb method continues to provide important information about alignment to assessment developers, educators, policy makers, and researchers. Because the quality and quantity of standards, objectives, and test items influences alignment results, well written items and objectives are essential to make accurate judgments about test alignment from Webb's method.

Achieve Alignment Method

The Achieve model was developed in 1998 at the Learning Research and Development Center at the University of Pittsburgh. Achieve, Inc. is an independent and bipartisan organization created by governors and chief executive officers. The Achieve model has been used in 14 states to assess the overall quality of the tests and alignment to state standards (Roach et al., 2008). The Achieve method uses both quantitative and qualitative alignment comparisons of the assessment and the standards based in a specific subject area including ELAR, mathematics and science. The Achieve method was developed to provide a story of alignment designed around three questions (Resnick, Rothman, Slattery, & Vranek, 2004):

- Does the assessment measure only content and skills reflected in the standards?
- Does the assessment fairly and effectively sample the important knowledge and skills in the standards?

- Is the assessment sufficiently challenging?

The Achieve method starts with accessing or creating a test blueprint to map the items to the objectives (Resnick et al., 2004). The test blueprint is developed by a senior reviewer. The senior review is someone who was involved the test development or has extensive experience with the test and standards. When the senior reviewer maps the items back to the objectives, potential for human error by raters called subject-matter experts (SMEs) is minimized and the purpose of the test is validated (Roach et al., 2008). The test blueprint allows for a comparison of the intentions of the assessment and what the assessment actually accomplishes (Resnick et al., 2004).

The alignment of each individual item is assessed in relation to the standards, and then the extent to which the test as a whole adequately measures the set of standards is examined (Resnick et al., 2004). First, individual items are judged by SMEs for their content centrality, performance centrality, and source of challenge. Next sets of items for each standard are examined for content centrality, performance centrality, challenge, balance, and range. The Achieve method does not have clear cut offs for each dimension, but rather focuses on the holistic picture of alignment (Resnick et al., 2004).

Content centrality. The degree of the match between an item and an objective is measured with content centrality. SMEs evaluate the quality of the item and objective matches, which are done prior to the alignment study in the test blue print. SMEs rate a 2 for *clearly consistent*, 1A for *not specific enough* meaning the standard or objective is too broad to be assured of the item's strong alignment, 1B for *somewhat consistent* meaning that the item only assesses part of the objective and the less central part of a compound objective, or 0 for *inconsistent* (Roach et al., 2008). The process is confirmatory and serves to provide

information about the intended purpose of the items. SMEs scores are averaged. Like the Webb method, an item can be mapped to two objectives on the blue print, or SMEs can indicate that they believe the item measures other objectives. With the use of the Likert scale previously described, the Achieve method allows SMEs to rate an item as only measuring part of the objective. The way in which items are coded in the Achieve method provides analysts with more information regarding the quality of the items and objective matches compared with the Webb method, which only asks raters to make a match, not judge the quality of the match.

Performance centrality. Performance centrality indicates the extent to which the item's cognitive demand level matches the level specified in the objective (Roach et al., 2008). According to Roach et al. (2008), cognitive demand refers to the type of thinking required to successfully complete the item. The SMEs have to decide whether the test item demands the same type of performance task as the related objective. Levels of cognitive demand include Level 1 (recall), Level 2 (application/skill), Level 3 (strategic thinking), or Level 4 (extended analysis). Performance centrality focuses on the match between the performance called for in the objective and the performance that the item is intended to measure. Performance centrality is also measured with a Likert scale with a rating of a 2 for *clearly consistent*, 1A for *not specific enough* meaning that the objective is too broad to be sure of the item's alignment, 1B for *somewhat consistent* meaning that the objective uses more than one verb, but the item matches only one verb, or 0 for *inconsistent* (Roach et al., 2008).

Challenge. Challenge, which is the extent to which the item has a range of difficulty that is both matched to the level of difficulty in the objective and appropriate for the target

students. Looking at item sets (all of the items linked to a particular standard), the source of challenge is measured to confirm that the items are constructed fairly and measure the intended construct. For both items and items sets, reviewers assess source of challenge, with the assignment of 1 for *appropriately difficult* or 0 for *inappropriate for grade level*. If an item scores a 0 for both content and performance centrality, then it is automatically rated a 0 for source of challenge (Roach et al., 2008).

Range and Balance. After assessing the item level, SMEs continue to evaluate the test as a whole, looking at the extent to which item sets cover the range of content from the standards and the extent to which emphasis is balanced across topics. Item sets are created with all items relating to a particular standard. Range is expressed as the proportion of objectives assessed by at least one test item and thus represents a basic indicator of overall coverage (Roach et al., 2008). The range of the items should present simple to complex items. Range is a quantitative measure of the proportion of the objectives within a standard that are measured by at least one item. Range is expressed as the fraction of the total objectives under a standard that are assessed by at least one item. According to Resnick et al. (2004), ranges from 0.50 to 0.66 are acceptable and above 0.67 are considered good.

Looking at the item sets, balance is a measure of how well particular content and skills in the items reflect the emphasis that the standard and its related objectives require. Looking at item sets, SMEs are asked to make qualitative judgments as to whether a set reflects the corresponding standard's emphasis on content and skills along two questions (Roach et al., 2008):

- What objectives in a standard seem to be over-assessed?
- What objectives in a standard seem to be under-assessed or not assessed at all?

Reviewers evaluate each question from two perspectives: (a) their reading of the standards; and (b) their personal judgments of what is most relevant for the particular grade level. Balance judgments fall into four categories: *good*, *appropriate*, *fair*, or *poor* (Roach et al., 2008). Sets of items are further evaluated on their level of challenge, a global judgment on the test's overall difficulty according to assessed concepts and cognitive demands placed on students. Reviewers make qualitative judgments regarding the cognitive demands of an entire set in relation to the demands specified in the matching standards, as well as if items skew toward more or less challenging concepts, types, or parts of objectives. The level of challenge for sets of items is rated as *easy*, *medium* or *hard*. A short written evaluation by each SME on each item set's level of challenge concludes the alignment process (Roach et al., 2008).

Application of the Achieve Method

Perhaps because of the extensive qualitative nature of the Achieve method and because versions to collect data regarding the criteria are not available online, the Achieve method is not widely used in research. Achieve method literature has found that assessments and standards are not well balanced (Resnick et al., 2004). Whereas individual items tend to align well to the standards, the tests, when looked at holistically, are not well aligned (Resnick et al., 2004). One key difference between the Webb and Achieve method is that in the Achieve method, SMEs are asked to rate the degree of alignment between a stated objective and item on a multipoint scale rather than match an item to objectives. In the Achieve method, SMEs rate the content and cognitive demand congruence between item-objective links, which is based on the test specifications developed by a senior reviewer. In a study, using the Arizona Instrument to Measure Standards (AIMS) 2004 high school

mathematics exam and the state's academic standards, reviewers were assigned to either matching as specified in Webb method or rating as specified in the Achieve method (D'Agostino, Walsh, Cimetta, Falco, Smith, VanWinkle, & Powers, 2008). SMEs practiced rating the alignment between items and objectives as *consistent*, *somewhat consistent*, or *not consistent* in three separate areas: content, intellectual skill, and overall match. Raters in the Webb method focused on both the content and intellectual challenge while matching items and objectives. According to D'Agostino et al. (2008), when comparing the two methods, a moderate correlation was found between Webb's overall match and Achieve's overall rating scores (r = .59). The item alignment decision agreement between the two methods converged moderately (kappa = .39). Eighty percent of the items, 32 out of 40 items, received similar alignment scores across the two methods. Matching error occurred for 2 of 40 items or 5% of the time in Webb's method. Matching is more flexible in Webb's model because the rater can link objectives to an item based on their judgment. Rating in the Achieve model is less susceptible to error because the items are already linked, and the raters evaluate the quality of that link, which saves them time in searching through the standards.

D'Agostino et al. (2008) concluded that rating seems most suitable for confirming the quality of the test specifications; whereas, matching can be used to confirm specifications or explore other possible item-objective connections that were not included in test specifications. Rating in the Achieve method is more time efficient, provides information about the quality of the fit on a Likert scale, and is less likely to result in error. Webb's matching provides more explorative information and should be used to gather a more global picture of fit. D'Agostino et al. (2008) suggested that both methods used in conjunction provide the most comprehensive alignment method.

Surveys of an Enacted Curriculum (SEC)

The Surveys of an Enacted Curriculum (SEC) method, also known as the Porter method (Fulmer, 2011), provides information for teachers and other stakeholders about the intended, the enacted, and the assessed curriculum (Kurz et al., 2010). The SEC method is the only method designed to take into account factors of alignment beyond assessments and standards. The intended curriculum is specified in the content standards for a particular subject or grade level. The content of instruction delivered by classroom teachers designates the enacted curriculum. Because the SEC reviewers map the alignment elements to a common framework, the SEC method can be used to analyze a variety of elements depending on purpose of alignment. Elements of alignment analysis can include comparisons across assessment, standards, curriculum, instruction, and student input (Kurz et al., 2010; Polikoff et al., 2011). The SEC results in a single statistic and a graphical output of alignment called a content map. The framework on which the content is graphed is represented with more general, big picture topics related to the elements of alignment. The content maps provide information on the depth of cognitive ability and coverage of topics. For the purposes of this study, the following review of the literature will focus on using the SEC method to compare topic coverage and cognitive demand for only assessments and standards. The uses and applicability of the SEC outreach the scope of this review.

Content maps are used to display the content coverage and emphasis data in order to visually assess alignment (Roach et al., 2008). The SEC method maps the standards and assessments onto a common framework—a content taxonomy. The taxonomy defines content with topics on one axis and cognitive demand on the other axis. SMEs place assessment items and objectives from standards into the taxonomy, and the documents are then

represented as a matrix of proportions, where the proportion in each cell (topic and cognitive demand) indicates the proportion of the total content in the document that emphasizes that particular combination of topic and cognitive demand (Lui et al., 2008). The matrices for standards and assessments are compared, cell by cell, and an alignment index is calculated, indicating the proportion of content in common (Lui et al., 2008).

The SEC method assesses alignment by calculating the Porter index (Lui et al, 2008). For the purposes of this study, the SEC method will result in two content maps, one representing the alignment between the topics and cognitive demands required for the formative assessments and one content map for the content standards. To make the content maps comparable, all cell values are standardized, that is converted into ratios totaling to 1. The rows and columns in the content maps visually represent relative emphasis of different topics and cognitive demands.

Survey. The SEC is typically comprised of three main alignment dimensions: (a) content match, which can be difficult to manage so the analysts should keep topics broad; (b) expectations for student performance or cognitive demand; and (c) instructional content, which asks teachers to self-report how much time is spent on each topic (Martone & Sireci, 2009). According to Martone and Sireci (2009), three or more SMEs are needed to complete the alignment ratings. Cognitive demand is a common dimension by which elements are scored; teachers are asked to identify items and standards as (a) memorize; (b) perform procedures; (c) communicate understanding; (d) solve non-routine problems; or (e) conjecture/generalize/prove (Lui et al., 2008). Studies have found higher response rates when teachers complete the survey in groups. Individual reports of results can be provided for teachers' professional development (Martone & Sireci, 2009). The researcher should use at

least five teachers and conduct a generalizability study to see if the raters are reliable (Martone & Sireci, 2009).

Output. On a graphical matrix with two axes, each cell-by-cell unit analyzed is a proportion of the whole. The alignment index is the sum of all the cell-by-cell intersection points and expresses alignment as a matter of degree, rather than an absolute. Content data from each survey is reduced to cell by cell proportions with the sums across all rows and columns equaling 1.00. The sum of all ratings for a particular content map for K-12 mathematics consists of cells in columns by topics with the sum of all ratings across cells equaling 1.00. The SEC method provides categorical concurrence (which looks at matching topics), balance of representation (which is a measure of relative emphasis of topic coverage), cognitive complexity (which is a measure of relative emphasis of cognitive demand), and an overall alignment index (which examines everything in a single index). Content maps are used to display the content coverage and emphasis data in order to visually assess alignment. Examples of content maps comparing the test and curriculum in New York State (NYS) are depicted in Figure 1.

Figure 1: Content Maps for the Curriculum and Test in NYS

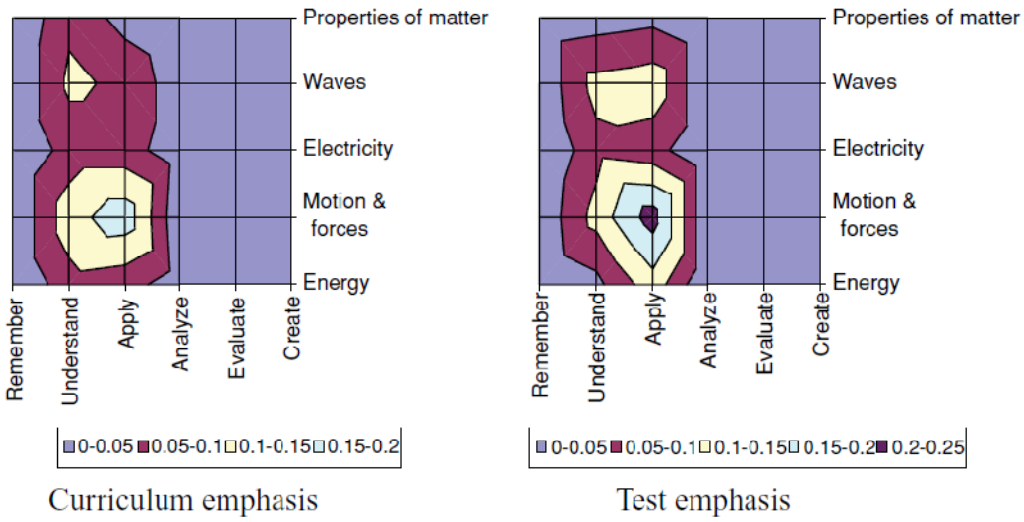


Figure 1. Adapted from "Alignment between the physics content standard and the standardized test: A comparison among the United States-New York State, Singapore, and China-Jiangsu," by X. Liu, B. Zhang, L. Liang, G. Fulmer, B. Kim and Y. Haiquan, 2008, *Science Education, 93,* p. 787.

The content maps in Figure 1 depict the relationship between the topics on the *y* axis and the cognitive demand on the *x* axis for the curriculum and test in physics. The pictorial analysis demonstrates that both the test and standards do not address cognitive levels above the level of analysis and focus primarily on understanding and application. Given the information in Figure 1, comparisons can be made between congruent intersection points on the content maps. For example, the dark purple area on the test emphasis content map indicates that between .20 to .25 percent of the items assess motion and forces at the application level. This can be compared with the curriculum content map, which shows that .15 to .20 percent of the curriculum targets motion and forces at the application level. After comparing each point of intersection across the content maps, a concern Lui et. al (2008) observed is that the test emphasizes the content at a higher level than the standards. The test focuses on application of concepts related to motion and forces and waves, while the

standards do not place emphasis on application of concepts related to waves and place a less concentrated focus on application of concepts related to motion and forces. This information could help state curriculum and test developers create more rigorous curriculum and more aligned tests. The test and the curriculum in this analysis could focus on higher order thinking skills such as analyze, evaluate and create, of which none are currently addressed in the curriculum and the test (Lui et al., 2008).

Index. Because Porter's index is determined independent of standards and assessments and each document is coded with the same rubric, policy makers are able to make decisions about the degree of alignment across multiple jurisdictions, not limited to tests and standards. Virtually any two categorical variables can be included in Porter's method, such as language complexity or gender neutrality (Kurz et al., 2010). The Porter method has relative simplicity in calculation and broad application compared with other methods. Looking at the two frequencies across content maps and producing a single alignment index, ranging from 0 to 1, which indicates how closely the distribution of points in the first table, for example relating to standards, aligns with the second table, perhaps relating to the assessment. The index (*P*) is determined by creating a table of frequencies for the two documents being compared, which are labeled *A* and *B*. For each cell in tables *A* and *B*, a ratio of points in the cell with the total number of points in the respective tables is computed, which are labeled as *a* and *b*. For every row *j* and column *k* in tables *a* and *b*, an absolute value of discrepancy between the ratios in cells a_{jk} and b_{jk} is calculated (Fulmer, 2011). In the equation, *J* is the number of rows and *K* is the number of columns in each table, and a_{jk} and b_{jk} are the ratios of points in the cells at row *j* and column *k* for the respective ratio

tables, *a* and *b*. The total number of cells in the table is called $N = (J * K)$. The alignment index is then computed used the following equation (Fulmer, 2011):

$$P = 1 - \frac{\sum_{k=1}^{K} \sum_{j=1}^{J} |a_{jk} - b_{jk}|}{2}$$

A greater number of cells in the table will yield a range of likely values that is lower than for tables with fewer cells. As the number of cells increases, there is much more room for discrepancy between the ratios, and the values for the index are likely to be lower (Fulmer, 2011). It is difficult to know whether a higher or lower alignment is meaningful or is a consequence of the table size, which highlights the need for established criteria for assessing the strength of alignment indices (Fulmer, 2011). Analyzing the results of 5,000 random alignment calculations, the mean alignment index was higher for tables of greater size, and the mean index was lower in cases with fewer points in the standards. Fulmer (2011) identified the mean and critical values for alignment indices and reexamined observed alignment values from previous research using these criteria. The results provided researchers and policymakers the first opportunity to draw conclusions as to whether or not observed alignment indices differ significantly from what could occur by chance. The average alignment index that might occur by chance is dependent on the size of the frequency tables being compared and the number of test items or standards involved in the comparison. Any effort to gauge the strength of alignment is affected by the scoring rubric that is used to code the test items or other document (Fulmer, 2011).

Application of the SEC Model

In a study of the SEC approach, two content-area experts examined the physics exam for NYS, Singapore and China (Lui et al., 2008). Based on the critical value of .78, there was

statistically significant alignment between the test and standards in New York, but not for China and Singapore. Both physics tests from China and Singapore shifted toward higher cognitive skills by de-emphasizing lower level cognitive skills and emphasizing higher level skills. It is not simply coverage of topics that is predictive of student achievement on standardized tests, but coverage and a focus on cognitive emphasis together that predict students' performance (Lui et al., 2008). The differences found may be a result of the various test formats across the nations. The researchers suggested that the curriculum in the United States as a whole is unfocused and does not promote depth of coverage and requires too many understandings (Lui et al., 2008).

In another study, Kurz et al. (2010) applied the SEC alignment methodology to examine differences in alignment between instructional content and state standards for eighth-grade general and special education mathematics teachers. Teachers reported on their instructional content coverage via an online or paper and pencil survey, done retrospectively at the end of the year. The SEC survey was completed at three points—at the beginning of the year to assess the planned curriculum, mid-year to measure the enacted for first half of year, and at the end of the year to measure the entire school year enacted. Using formative assessments throughout the year, which were aligned the curriculum according to an outside publisher, gain scores were calculated for each group of students. The findings did not suggest significant differences in the general and special education teachers planned and enacted curriculum. Low alignment indices were found across the board for general and special education teachers looking at a sample of 18 teachers.

In an analysis of 19 states' standards and assessment alignment—including 11 for ELAR, 14 for mathematics, and 9 for science—a total of 138 documents across the three

subjects were compared (Polikoff et al., 2011). Content maps were generated using Microsoft Excel, which resembled topographical maps where specific topics were displayed as lines of latitude and cognitive demands as lines of longitude. This provided a visual record of the content contained in the particular standards document or assessment that can be used to compare the content of standards or assessments within or between the states. Across the 19 states the average test-standards alignment index was .19, indicating that 19% of the content was shared between tests and standards (Polikoff et al., 2011). The average alignment index was slightly higher for mathematics at 0.27 and science at 0.26. The alignment of state standards with assessments of student achievement was typically in the range of 0.20 to 0.30. These results may be under-estimating alignment, because the number of items on the test and the number of cells influences alignment results in the SEC method (Polikoff et al., 2011).

Polikoff et al. (2011) suggested that there was no apparent pattern in misalignment across grades. About 24% of test content in grade 3 through 8 was at the wrong level of cognitive demand, and across grades 3 through 12, the right topic and wrong cognitive demand levels were closer to 51% in mathematics (Polikoff et al., 2011). When cognitive demand was ignored, agreement increased on average to 0.80. In mathematics 34% of standards were typically not tested at all, 52% for ELAR and 23% for science. About half of the content in mathematics and science standards and two-thirds in ELAR were misaligned with test content. In mathematics the standards tended to place a greater emphasis on the two highest levels of cognitive demand, and the average alignment indices for state standards and assessments were below .30 in mathematics and science and below .20 in ELAR. No alignment index was above .50 for any state, grade, or subject included in the study. There

was some consistency across states in what was over and under-tested across the included subjects. This review demonstrates the broad uses and applications of the SEC method, which results in less complex alignment results but can be applied to a broader range of contexts compared to the Webb and Achieve methods.

Issues and Future Directions

According to Bhola et al. (2003), some generalizable issues arise when using alignment models. Many objectives are multidimensional, and the items that are identified as corresponding only focus on one dimension within an objective. For example an objective may specify the use of whole numbers, fractions, and ratios, and an item corresponding to that objective may only represent the use of whole numbers. This makes rating and matching items to objectives difficult. A second issue relates to students of various levels needing an opportunity to demonstrate a range of levels of proficiency. The items for a particular standard must span a wide range of difficulty to permit students throughout the proficiency continuum to demonstrate their ability. Having enough items to accurately classify students into performance criterion and adequately cover all standards is very difficult. A third issue is that alignment may be influenced by content area. A review of Nebraska's content standards and assessments found that in science, no objective had more than three aligned items, but almost every objective was covered. On the other hand, most social studies objectives had no items corresponding but had some objectives which were heavily hit. A fourth issue relates to training, which is difficult because many teachers tend to be expansive in their decisions of what constitutes a content match. All of these limitations apply to this study.

Alignment studies will need to be conducted as states adopt new assessments and standards. This alignment study is designed to shed light on the weaknesses and strengths of

alignment studies in order to make suggestions for the new Common Core State Standards (CCSS) and assessments, which are currently being developed. Using the SEC method, Porter, McMaken, Hwang, and Yang (2011) found a lack of alignment between the Common Core State Standards (CCSS) and state standards and assessments. Beach (2011) wrote a response to Porter et al.'s findings stating that the CCSS focused on argumentative writing and expository text to a greater extent than current state standards. Substantive curriculum and instructional changes will need to take place over the next few years in order for successful adoption of the new CCSS. Current standards-based reform is intended to result in more rigorous curriculum to better prepare students for college, hence a need exists for more research on alignment.

Method

In a suburban North Carolina district, teachers in grades three through five are provided with formative assessments to assess student achievement according to the North Carolina's content standards, called the North Carolina Standard Course of Study (NCSCoS). The NCSCoS provides specific goals, standards, and objectives for each area of study. The formative assessment results are used to guide instruction in preparation for the North Carolina End-of-Grade Tests (EOGs) and are common across the district, meaning each school uses the same assessment. The EOGs measure student achievement in the areas of math and reading for students in grades three through five. Therefore, in order for the system to work coherently, the NCSCoS, the formative assessments, and the EOGs must align to one another. In this study, alignment methods including the Webb method, Achieve method, and the SEC method were used to measure the level of connection between the NCSCoS with the formative assessments. In order to understand how the alignment of the formative assessments compared to the alignment of the EOG, an additional alignment study was conducted between the NCSCoS and the EOG using the Webb method.

The 2008-2009 Fourth Grade Mathematics North Carolina EOG Test (Form T) consists of 50 items, including 14 calculator inactive items and 36 calculator active items. Students were permitted to use a calculator on the calculator active items, but use of a calculator was not allowed on the calculator inactive items. The test was given at the end of the 2008-2009 school year to assess student learning in accordance with the NCSCoS, which

were implemented in 2003. The third quarter 2010-2011, district, grade four, mathematics formative assessments were compared to the NCSCoS using the three alignment methods. Use of the formative assessment system was required by the district for all third through fifth grade teachers in the district during the 2010-2011 school year. This study examines the alignment between the formative assessment and the full NCSCoS, as well as the district-specified third quarter mathematics standards, which were obtained from the district pacing guide. The formative assessment is meant to only measure student knowledge associated with a subset of the content standards for the entire year. Thus, for this study, only the third quarter, district-specified standards were included in the formative assessment. The district-specified standards are a subset of the full NCSCoS, excluding a total of four objectives from the full standards. The excluded four objectives are meant to be taught during other quarters throughout the school year.

Participants were recruited from a North Carolina school and from the researcher's university. Upon receiving permission from the district and school principal and upon IRB approval, the researcher sent recruitment information via email to the staff at a local school and requested participation from individuals with experience teaching fourth grade mathematics. Participants indicated their schedule availability on a Google form. The researcher assigned educators to one of three methodologies depending on the number of those whom agreed to participate and their availability. As suggested by Martone and Sireci (2009), three participants were assigned to the SEC method; three participants were assigned to Achieve; and six participants were assigned to the Webb method. Graduate students from the University of North Carolina School of Education were recruited because less than 12 educators expressed interest in participating in the study. Six teachers and six graduate

students in education participated in the study. Current teaching assignments included three fifth grade teachers, two fourth grade teachers, one third grade teacher, and one high school science teacher. Four of the six teachers had experience teaching fourth grade mathematics. The teachers included one first year teacher, one fourth year teacher, one sixth year teacher, one ninth year teacher, and two teachers with over ten years of experience. All teachers taught in the district where this study's formative assessments were used and were familiar with the assessments. Graduate students included two master's students in educational psychology, one master's student in the early childhood, special education, and literacy, one doctoral student in social foundations of education, one doctoral student in educational psychology, and one doctoral student in the early childhood, special education, and literacy. Among the graduate students, three had no experience teaching at the K-12 level; one taught third grade for two years; one taught middle school mathematics for six years; and one taught high school English as a Second Language for one year.

The researcher met with participants in their alignment groups and conducted a 30-minute training, which included calibration on the appropriate alignment method. Following the training, teachers rated 49 items if assigned to the Achieve or SEC method and 99 items if assigned to the Webb method. The researcher remained at the alignment session to clarify directions if questions arose. The data were collected in pencil-and-paper form from the participants and entered into an Excel document by the researcher. The forms and definitions that were used for data collection can be found in the Appendices A through F.

In the Webb alignment method group, six participants were trained to recognize and apply four depth of knowledge (DOK) levels including recall, skill/concept, strategic thinking and extended thinking to items and objectives (see Appendix A). The panel

reviewed the objectives and reached consensus on the DOK levels. The panel then independently rated the DOK levels and matched objectives to each assessment item on the summative EOG and formative assessments, using a common data collection instrument developed by the researcher (see Appendix B). Categorical concurrence, ROK, balance of representation, and DOK consistency were calculated by the researcher based on the participants' responses. The participants in the Webb session included one teacher and five graduate students. The session occurred over a three hour period at a university restaurant. Participants were given snacks, drinks, and a ten dollar restaurant gift card. Participants were permitted to discuss items with each other if they had concerns or questions.

The Achieve method session also included a half hour training with calibration. Following the training, participants rated the quality of the content and performance match between individual items and their respective objectives, which were suggested by the formative assessment's test specifications designated by the test publisher. Each item was examined for source of challenge. Following the items review, the participants judged whether the item sets relating to a standard represented comparable balance and challenge. Based on the test specifications, the range statistic was calculated for each standard by dividing the number of objectives with matches by the total number of objectives under a standard. Content centrality, performance centrality, and source of challenge were examined across the three subject-matter experts (SMEs) and the majority response was recorded for each item, along with standard deviations and means. Looking at item sets, qualitative notes and labels for balance and challenge were examined across raters. The three SMEs for the Achieve method included a fifth grade teacher, a third grade teacher, and a graduate student. The teachers both had experience teaching fourth grade mathematics. The SMEs were

encouraged to discuss rating throughout the session, but were not required to reach consensus, which would typically be required in an Achieve session, due to time constraints. The Achieve session lasted for two and a half hours at a local elementary school after the instructional day ended. Participants were given pizza, drinks, snacks, and the researcher volunteered to help in the teachers' classrooms for a day following the session. The forms used to collect the Achieve data can be viewed in Appendices C and D. The full form including the test blueprint is not included due to formatting considerations, but is available upon request from the researcher.

After a half hour training including calibration and before rating the individual items, the participants in the SEC alignment group rated each objective on its level of cognitive demand (e.g., memorize facts/definitions/formulas, perform procedures, demonstrate understanding of mathematical ideas, conjecture/generalize/prove or solve non-routine problems/make connections). Participants rated the items for their cognitive demand and matched each item to a goal (e.g., number and operations, measurement, geometry, data analysis and probability, or algebra). The full forms used for data collection can be viewed in Appendices E and F. The SEC method included three teachers (one fourth grade teacher and two fifth grade teachers). The session lasted for two hours at an elementary school on a teacher work day. Participants were given pizza, drinks, snacks, and the researcher volunteered to help in the teachers' classrooms for a day following the session.

Results

In this section, the results for the Webb, Achieve, and SEC method will be reported and compared. The first research question focuses on the alignment of the formative assessment with the state content standards as measured by the Webb method, Achieve method, and SEC method. These findings will be reported after describing the results for the three methods. The second research question investigates how the alignment results compare across methods, which will also be assessed after reporting the results for the three methods. The third research question focuses on the alignment of the summative test to the state content standards using one method, the Webb method. To answer the third research question, the results from the Webb method for summative test will be reported as part of the results for the Webb method. After reporting the Webb, Achieve, and SEC results, the first and second research questions will be answered by examining and comparing the results across methods.

Webb Results for the Formative Assessment and Summative Test

The results for each test item in the Webb method consist of the majority responses across the six raters trained in the Webb method. The primary match was decided based on the clear majority response of the raters. In all cases, a clear majority for the primary match was attainable. The secondary match was decided based on the majority secondary response of the raters. If no clear majority response was indicated, a secondary match was not identified. Next, the average of the raters' DOK assignments for individual items was calculated and rounded to the nearest whole number. On the formative assessment, only one

item's assignment of DOK indicated a standard deviation above 1 (item 26); therefore, the small standard deviations indicate that assignment of the DOK levels was somewhat consistent among raters. On the summative assessment, four items' assignments of DOK indicated standard deviations above 1 (items 15, 16, 29, and 48), suggesting that there was not as much agreement on DOK level assignments on the summative test as there was on the formative assessment.

DOK Consistency. DOK Consistency examines the extent to which the DOK levels of the items match or are above the DOK levels of the corresponding objectives. DOK consistency is met acceptably if at least 50% of the items corresponding to an objective are written at or above the DOK level of the objective. DOK is weakly met if between 40-49% of the items corresponding to an objective are written at or above the level of the objective. Finally, if less than 39% of the items are written at or above the DOK level of the corresponding objective, the DOK consistency is considered unacceptable. After comparing individual item and objective matches and indicating whether the item was written below, at, or above the DOK level of the objective, the DOK consistency for the formative assessment was examined by standard, which is shown in Table 1. The results in Table 1 demonstrate that when compared to the full NCSCoS and the district-specified standards for third quarter, the formative assessment exhibits acceptable levels of DOK consistency on all standards with the exception of the geometry standards. The summative test demonstrated acceptable alignment to the NCSCoS on all standards.

Table 1: DOK Consistency for Formative and Summative Assessments

DOK Consistency		Below DOK		At DOK		Above DOK		DOK Level	% at or above
		#	%	#	%	#	%		
Formative Assessment	Number and Operations	4	12.9	22	70.97	5	16.13	Acceptable	87
	Measurement	0	0	7	63.64%	4	36.36%	Acceptable	100
	Geometry	5	71.42	2	28.57	0	0	Unacceptable	28.57
	Data Analysis and Probability	2	40	3	60	0	0	Acceptable	60
	Algebra	6	28.57	4	19.05	11	52.38	Acceptable	71.43
Summative Assessment	Number and Operations	8	27.59	13	44.83	8	27.59	Acceptable	72.41
	Measurement	0	0	13	44.83	8	27.59	Acceptable	100
	Geometry	3	50	3	50	0	0	Acceptable	50
	Data Analysis and Probability	1	12.5	4	50	3	37.5	Acceptable	87.5
	Algebra	8	47.06	4	23.53	5	29.41	Acceptable	52.94

In short, these results suggest that the summative assessment is slightly more aligned to the standards with regard to DOK consistency than the formative assessment. With only 28.57% of the geometry items measuring at the DOK levels of the objectives, the DOK for the geometry items needs to be improved before teachers can make conclusions about what students know about geometry in preparation for the summative test, which has 50% of the items written at the DOK level of the objectives.

Categorical Concurrence. Categorical concurrence measures the extent to which at least some of each standard is represented on the test. According to Webb (2007), in order for categorical concurrence to be acceptable, at least six items must correspond to a standard. If four to five items correspond to a standard, categorical concurrence is deemed weak. Finally, if three or less items are linked to a standard, categorical concurrence is labeled unacceptable. For the purpose of calculating categorical concurrence, if an item was assigned to a primary and secondary objective within the same standard, the item counted as one item toward

achieving acceptable categorical concurrence for the standard. On the other hand, if the item was linked to two objectives across two different standards, the item counted as one item for each corresponding standard. In other words, items corresponding to two objectives under the same standard counted as one item in this analysis. Items corresponding to two objectives under different standards counted as two items in this analysis or one item for each standard. This decision was made so as to not overestimate categorical concurrence. Overestimation of categorical concurrence may have resulted if one item with two objectives within one standard qualified as two items for one standard. Looking at the formative assessment results presented in Table 2, the categorical concurrence results for both tests were overall acceptable. The results listed in Table 2 for the formative assessment are compared to the district-specified NCSCoS, but the results were very similar when comparing to the full NCSCoS and the district-specified third quarter standards.

Table 2: Categorical Concurrence of Formative Assessment and Summative Test

Formative Assessment	# of Objectives	# of Hits	Cat. Con. Level	% Acceptable
Number and Operations	5	26	Acceptable	80
Measurement	2	6	Acceptable	
Geometry	1	7	Acceptable	
Data Analysis and Probability	2	5	Weak	
Algebra	3	16	Acceptable	
Summative Test	Objectives	# of Hits	Cat. Con. Level	% Acceptable
Number and Operations	5	22	Acceptable	100
Measurement	2	6	Acceptable	
Geometry	3	6	Acceptable	
Data Analysis and Probability	4	8	Acceptable	
Algebra	3	14	Acceptable	

To summarize, the results for categorical concurrence shown in Table 2 indicate that all of the standards were met at the acceptable level on the summative test, but 80% of the standards were met at the acceptable level on the formative assessment. If the criteria proposed by Webb were considered, one more item should be added to the data analysis and

probability standard in order to achieve acceptable categorical concurrence with the expectation of a minimum of six items. However, the expectation of six items for each standard is somewhat arbitrary because some of the standards have more objectives, which may necessitate more items. While the formative assessment is close to make the six item cutoff for all standards, the need for a minimum of six items may contribute to the overestimation of the alignment with the Webb method. Whereas, it makes sense that at least six items are necessary to make a conclusion about a student's knowledge relating to a standard, some standards may require more than six items because of the structure or the number of the objectives.

Range of Knowledge. Range of knowledge examines the relationship between the total number of objectives in the standards and the total number of objectives hit by at least item on the test. Range of knowledge is designed to examine the distribution of hits across standards. At least 50% of the objectives under each standard should have at least one corresponding item in order for range of knowledge to be considered acceptable. On the contrary, if less than 50% of the objectives under a standard are matched to items on the test, the range of knowledge is considered unacceptable. Table 3 shows the range of knowledge results for the formative assessment and the summative test. As can be seen in the table, the results for the range of knowledge indicate that when comparing the summative test to the state standards, all of the standards reached acceptable levels for range of knowledge. This is shown by a comparison of the values for the formative assessment compared to the district-specified standards; four out of five standards were met at an acceptable level of range of knowledge. Data analysis and probability was not met at an acceptable level of range of knowledge on the formative assessment.

Table 3: Range of Knowledge for Formative Assessment and Summative Test

Formative Assessment	# of Objectives	# of Objectives with Hits	ROK Level	% of Objectives Hit
Number and Operations	5	4	Acceptable	80
Measurement	2	2	Acceptable	100
Geometry	3	2	Acceptable	66.67
Data Analysis and Probability	4	1	Unacceptable	25
Algebra	3	3	Acceptable	100
Summative Test	# of Objectives	# of Objectives with Hits	ROK Level	% of Objectives Hit
Number and Operations	5	5	Acceptable	100
Measurement	2	2	Acceptable	100
Geometry	3	3	Acceptable	100
Data Analysis and Probability	4	4	Acceptable	100
Algebra	3	3	Acceptable	100

To sum up, the results for range of knowledge shown in Table 3 indicate that the summative test covers the objectives under each standard better than the formative assessment. Requiring 50% as general cutoff for acceptable seems like a generous expectation for range, which could also contribute to overestimating alignment with the Webb method. In the researcher's opinion, all objectives included on the standards should be measured by at least one item on the formative and summative test. The summative test accomplishes this with 100 % of the objectives acquiring hits; however, the formative assessment does not accomplish full coverage of all of the objectives for the third quarter standards or the full standards. The formative assessment clearly does not measure enough breadth across the objectives with only two standards meeting 100% coverage.

The formative assessment is identified as having unacceptable range of knowledge for only the data analysis and probability standard, but by visually comparing the tests, the researcher noted that for numbers and operations, the formative assessment does not require students answer any questions using fractions, while the summative test requires students to

use fractions to solve many items. The NCSCoS specifies using fractions as a component of objective 1.03. In order for the formative assessment and summative test to work coherently, the tests must focus on similar cognitive demand and skills, or the formative assessment must include at least as challenging or preferably more challenging items compared to the summative test, which would secure student success on the summative test by engaging students to more challenging material and instruction than they would see on the summative test. At the most basic level, the formative assessment needs to emphasize fractions to same extent that the summative test emphasizes fractions, and the Webb method does not measure this emphasis, which could drastically affect how students perform on the summative test. If students are prepared for the summative test using the items on the formative assessment and instruction is tailored for students around their performance on the formative test, then the students will be missing the important basic instruction relating to fractional concepts and the system will not be working coherently to ensure student success.

Balance of Representation. The results for balance of representation take into account how the hits are distributed under each standard. Balance of representation is expressed in an index and is calculated with a formula that examines the proportion of hits assigned to each objective relative to other objectives with hits under the standard. The index does not account for objectives that do not have hits; objectives that do not have hits are meant to be examined with range of knowledge. Balance of representation is designed to measure how evenly distributed the hits are across the assessed objectives under a standard. Although the results for balance of representation are the same levels when comparing the formative assessment to the full NCSCoS and the district-specified third quarter standards, the results listed in Table 4 for the formative assessment are for the full NCSCoS. If a

standard has a balance of representation index over .70, the standard is considered to be met at an acceptable level.

Table 4: Balance of Representation for Formative Assessment and Summative Test

Formative Assessment	# of Total Objectives	# of Objs. Hit	# of Total Hits	Bal. Index	BOR Level	% Acceptable
Number and Operations	5	5	31	0.5645	Unacceptable	80
Measurement	2	2	11	0.9545	Acceptable	
Geometry	3	2	7	0.9286	Acceptable	
Data Analysis and Probability	4	1	5	1	Acceptable	
Algebra	3	3	21	0.9048	Acceptable	
Summative Test	# of Total Objectives	# of Objs. Hit	# of Total Hits	Bal. Index	BOR Level	% Acceptable
Number and Operations	5	5	29	0.8138	Acceptable	100
Measurement	2	2	12	1	Acceptable	
Geometry	3	3	6	1	Acceptable	
Data Analysis and Probability	4	4	8	0.875	Acceptable	
Algebra	3	3	17	0.8039	Acceptable	

As shown in Table 4, for the balance of representation the standards are met at the acceptable level in the summative test, but 80% of the standards meet the acceptable cutoff for the formative assessment. This suggests that the items corresponding to number and operations are weighted more heavily toward some objectives and are not fairly distributed across hit objectives. The serious limitation of this index is that it only includes hit objectives in its calculation. This works acceptably for the summative test because 100% of the objectives are hit by at least one item, but the formative assessment did not demonstrate adequate range on all standards and did not have 100% of the objectives hit by at least one item on all standards, so the balance of representation indices need to be interpreted with caution. Even with limited range, the balance of representation indices for the formative

assessment indicate the items for the hit objectives are not distributed evenly across the objectives under a standard.

In order to answer my third research question about the comparison of the alignment of the formative assessment and summative test to the state standards, the Webb method suggests that the summative test is well aligned to the NCSCoS and that the alignment of the formative assessment and NCSCoS is acceptable. The Webb method misses a few very important characteristics of alignment. For example, while the NCSCoS specifies that students must be able to use fractions, the formative assessment does not include any items related to the use of fractions. Yet, the majority of the items on the summative test require the use of fractions. Students who do not understand fractions will likely not be equipped for success on the summative assessment. If instruction is tailored using the formative assessment, concepts relating to fractions might not be taught sufficiently. Lastly, for the standard of algebra on the formative assessment, more than 50% of the items are written above the DOK level of the objective. This suggests that some of the algebra items are written at a level that is not appropriate for measuring the algebra objectives specified in the standards. Webb does not account for this distinction in his method.

Achieve Results

The Achieve method is a holistic assessment of alignment including quantitative and qualitative information. For the purposes of this study, the reported results are focused on the quantitative results of the Achieve method. The individual items (n = 49) on the formative assessment were analyzed by three SMEs for their content centrality, performance centrality, and source of challenge. The results for the item level analysis were based on majority responses across the three raters. If there was not a clear majority, the responses were

averaged and rounded to the nearest whole number to find content centrality, performance centrality and source of challenge scores. The standard deviations across the SMEs for each criterion were calculated. Standard deviations were all below 0.577, indicating that at least 2/3 of SMEs agreed on most items across the criteria. The raters assigned item sets, which included all items pertaining to a standard, an overall judgment score for balance (poor, fair, appropriate, good) and level of challenge (easy, medium, hard). The range was calculated for item sets by dividing the number of objectives under a standard hit with at least one item by the total number of objectives listed under a standard in the NCSCoS. Range in the Achieve method is an expression of the portion of the standards represented by at least one item on the test.

Content Centrality. The item ratings for content centrality were assigned by SMEs based on the level of match between the item and the objectives assigned to the item on the test blueprint, which was made available through the formative assessment developer. SMEs examined the item and the assigned objective(s) and indicated a score of 2 if the item and objective or objectives clearly and consistently matched. If the objective was written in a way that was not specific enough or too vague to match the item, the SMEs rated the item with a score of 1A. On the other hand, if the objective was written with too much specificity and listed relevant content, but the item only measured one part and the less essential part of a compound objective, the SMEs assigned the item's content centrality a score of 1B. Finally, if the assigned objectives and the item did not match, the SMEs assigned a score of 0. SMEs were in total agreement on these ratings 73.47% of the time (for 36 out of 49 items), and 2/3 agreement on 26.53% of the time (13 out of 49 items).

Table 5: Content Centrality for Formative Assessment

Rating	# of Items	% of Assessment
2	6	12.24%
1A	7	14.29%
1B	35	71.43%
0	1	2.04%

The results shown in Table 5 show that approximately 71% of the test measure the less essential part of compound objectives. This is evidence that the quality and the clarity of the objectives influence the results of the Achieve method. Many of the objectives were written with long lists of content; for example, objective 5.02 requires that students translate among symbolic, numeric, verbal, and pictorial representations of number relationships, and objective 1.03 requires students to solve problems using models, diagrams, and reasoning about fractions and relationships among fractions involving halves, fourths, eighths, thirds, sixths, twelfths, fifths, tenths, hundredths, and mixed numbers. The lists of content associated with one objective make aligning items difficult. At times, SMEs questioned the most essential pieces of the objectives because of the lists of content within the objectives.

Performance Centrality. The ratings for performance centrality were assigned based on the level of congruence between the performance specified in the item and the objectives assigned to the item on the test blueprint. SMEs examined the item and the assigned objective(s) and labeled the item with a score of 2 if the item and objective(s) clearly and consistently matched with regard to performance. If the verb of the objective was written in a way that was not specific enough or too vague to match the item, the SMEs rated the item with a score of 1A. By contrast, if the verbs in the objective were compound and consisted of lists of verbs, but the item only measured one part and the less essential part of a compound

objective, the SMEs assigned the item's performance centrality with a score of 1B. Hence, if the assigned objectives and the item did not match, the SMEs assigned the item with a score of 0. SMEs were in total agreement on these ratings 65.31% of the time (32 out of 49 items), and 2/3 agreement on 30.61% of the time (15 out of 49 items). The SMEs did not agree on a majority rating for performance centrality 4.08% of the time (2 out of 49 items).

Table 6: Performance Centrality for Formative Assessment

Rating	# of Items	% of Assessment
2	23	46.94%
1A	10	20.41%
1B	14	28.57%
0	2	4.08%

The results for Performance Centrality shown in Table 6 suggest that almost half of the assessment was written at a performance level consistent with the objectives, but the lists of verbs in some of the objectives made approximately 28.57% of the items difficult to rate as consistent with the objectives because the items only measured part of the objective. Finally, about twenty percent of the items were matched to objectives with a vague or unclear verb in the objective.

Source of Challenge. Source of challenge is rated as 1 if the challenge is appropriate for the grade level, in this case fourth grade, and is written clearly, without misleading language or information that does not relate to the objective. If the item contains an extraneous challenge, the source of challenge for the item is rated a score of 0. The SMEs rated the source of challenge for the items consistently 81.63% of the time (40 out of 49 items) and with 2/3 agreement 18.37% of the time (9 out of 49 items).

Table 7: Source of Challenge for Formative Assessment

Rating	# of Items	% of Assessment
1	45	91.84%
0	4	8.16%

These results for source of challenge shown in Table 7 indicate that the majority of the assessment (91.84%) is written at a level is appropriate for fourth graders and does not include many extraneous sources of challenge.

Range, Challenge, and Balance. Range is designed to express the portion of the objectives under a standard that are represented on the assessment by at least one item. These calculations are based on the test blueprint. The ranges are reported for the full NCSCoS and the district-specified third quarter standards in Table 8. In order to create cutoffs for the range results, percentages of objectives hit higher than 67% are considered good, between 50-66% are considered acceptable, and below 49% is considered unacceptable.

Table 8: Range Levels for Item Sets on the Formative Assessment

Third Quarter Standards				
Standard	# of Objectives	# of Objectives Hit	% of Objectives Hit	Range Level
Number and Operations	5	2	40%	Unacceptable
Measurement	2	2	100%	Good
Geometry	1	2	100%	Good
Data Analysis and Probability	2	1	50%	Acceptable
Algebra	3	3	100%	Good
Full NCSCoS				
Standard	# of Objectives	# of Objectives Hit	% of Objectives Hit	
Number and Operations	5	2	40%	Unacceptable
Measurement	2	2	100%	Good
Geometry	3	2	66.66%	Acceptable
Data Analysis and Probability	4	1	50%	Acceptable
Algebra	3	3	100%	Good

Similar to the results for the Webb study, the results for Range of Item Sets shown in Table 8 indicate that overall the formative assessment does not sufficiently cover the range specified in the district-specified third quarter standards or the full NCSCoS at an overall *good* level. Range of knowledge for the Webb method indicated that the standard data

analysis and probability was not acceptably met, and the Achieve results indicated that the standard number and operations was not acceptably met, which highlights inconsistent findings across methods based on the varying matching procedures. The Achieve method blueprint suggests different item/objective matches than the raters generated in the Webb method, such that the range levels were calculated at different levels across the standards.

The challenge component of the item set analysis asks SMEs to rate the overall level of challenge for the item sets as easy, medium, or hard. The balance component asks SMEs to rate the overall balance of the item set across the objectives as poor, fair, appropriate, or good. The SMEs ratings for challenge and balance are located in Table 9, which includes the majority responses of the SMEs, along with their percent agreement.

Table 9: Challenge and Balance for Item Sets for the Formative Assessment

Standard	Challenge	% SME Agreement	Balance	% SME Agreement
Number and Operations	Medium	100%	Poor	66.67%
Measurement	Medium	100%	Appropriate	100%
Geometry	Medium	100%	Poor	100%
Data Analysis and Probability	Medium	100%	Poor	100%
Algebra	Medium	100%	Poor	100%

Overall, the results shown in Table 9 reveal a medium level of challenge, which is desirable. The Webb results for DOK give a more detailed analysis of how the DOK levels of the objectives relate to the DOK levels of the items compared to the qualitative judgment in the Achieve method for challenge. The qualitative notes written by SMEs indicated overall satisfaction with the level of challenge associated with the items on the formative assessment and did not indicate that almost half of items associated with the algebra standard were written above the cognitive demand associated with the algebra objectives, which was highlighted in the Webb method. Table 9 also suggests that the balance ratings for the item

sets are of concern because overall the balance across objectives is poor, with the exception of the item set relating to measurement. However, the Webb method found that the balance of representation was acceptable for all standards on the formative assessment with the exception of the standard number and operations. Keep in mind that Webb's balance of representation may overestimate balance because the formula only includes objective with hits. Balance in the Achieve method requires SMEs to qualitatively compare the breadth of the items assigned to a standard by the blue print to the objectives listed under a standard, allowing for a more holistic and subjective, but possibly more accurate understanding of balance than Webb's balance of representation index, which is limited to only accounting for objectives with hits.

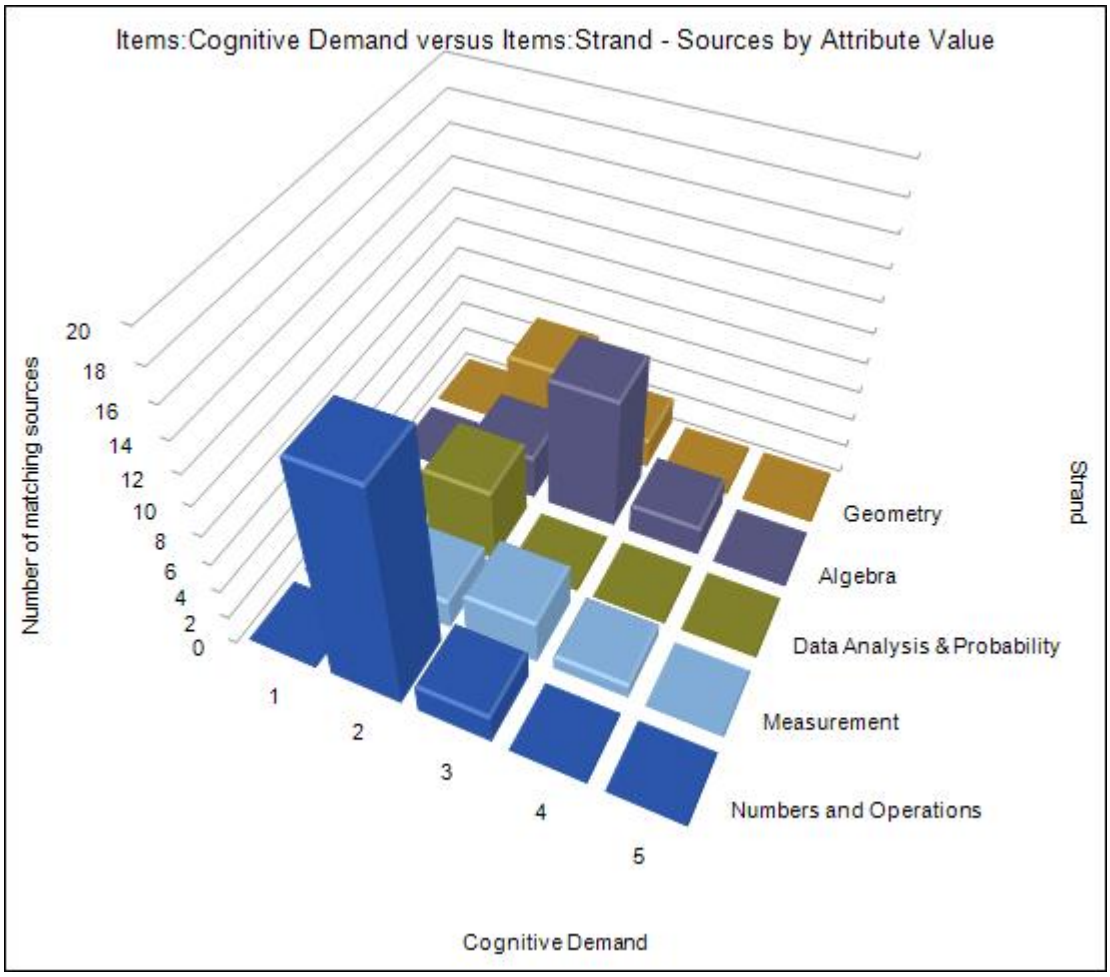
As a result, the Achieve method indicates poor alignment between the formative assessment and the NCSCoS. Most items were written to assess part and usually the less essential part of the objectives. A limitation of the Achieve method is that if the objectives are written to vague or too wordy from the perspective of the SMEs, the test items are consequently harshly rated. This raises concerns about the quality of the standards and objectives, as well as concerns about the alignment methodologies sensitivity to differences across standards, which will be discussed in the conclusions section.

SEC Results

The SEC method can be applied to a variety of elements of analysis in education, including tests, standards, curriculum, textbooks, instruction, and student feedback (Kurz et al., 2010). The results of the SEC alignment method include content maps, which were created with the chart features in Nvivo 9, a qualitative software program. The content maps demonstrate the relative emphasis of the standards or the formative assessment on content

topics and cognitive demand. The strands or topics included in this study included algebra, data analysis and probability, geometry, measurement, and numbers and operations. These strands or topics were based on the standards represented in the NCSCoS. The cognitive demands included in this study were memorize, perform, demonstrate, generalize, and problem solve. After training and calibration, three raters labeled the cognitive demand and the topic assignment for the items on the formative assessment and the objectives specified on the NCSCoS. The inter-rater agreement was calculated for the items across raters using Stata 10. The combined unweighted kappa coefficient across the three raters was 0.222 ($Z = 3.86$, $p = 0.0001$). The inter-rater reliability correlations between the three raters ranged from 0.356 to 0.414. This is less than ideal agreement across raters on the items cognitive demand level. The inter-rater agreement across raters for the objectives was low, with a combined unweighted kappa coefficient of 0.157 ($Z = 1.77$, $p = 0.0381$). The inter-rater reliability for the objectives across raters ranged from 0.299 to 0.642. Again, this is less ideal agreement for the cognitive demand associated with objectives. However, the raters agreed 100% of the time on the strand identified with each objective. Along the same lines, the inter-rater reliability correlations for the strand identified for each item ranged from 0.55 to 0.59. The inter-rater agreement combined unweighted kappa coefficient was 0.761 ($Z = 17.10$, $p < 0.0000$), which indicates acceptable agreement. Individual rater content maps were examined for consistency, and the majority response for each item and objective was included in the overall content map for the formative assessment and the standards. Figure 2 shows a content map representing the coverage of the items on the formative assessment according to majority topic assignment and cognitive demand.

Figure 2: Content Map for the Formative Assessment



This content map shown in Figure 2 indicates that the overall coverage of the topics and cognitive demands is low (between 0-5 items), and the highest number of items assess numbers and operations at the cognitive level performance. The most concentrated grouping of items with a high cognitive demand is written for the topic algebra, requiring students to think at the demonstration level. Few items require students to generalize and problem solve. In fact, there are three items written at the generalization level and no items at the problem solving level. The majority of the items are written at the perform level. There is a large concentration of number and operations items at the cognitive level two (performance), and the less concentrated but still distinct grouping of algebra items at the cognitive level

demonstrate. Two of the generalization items are assigned to the algebra topic and one is assigned to the measurement topic. Comparatively, the content map in Figure 3 demonstrates the distribution of objectives across the topics and cognitive demands.

Figure 3: Content Map for Full NCSCoS (Full Standards)

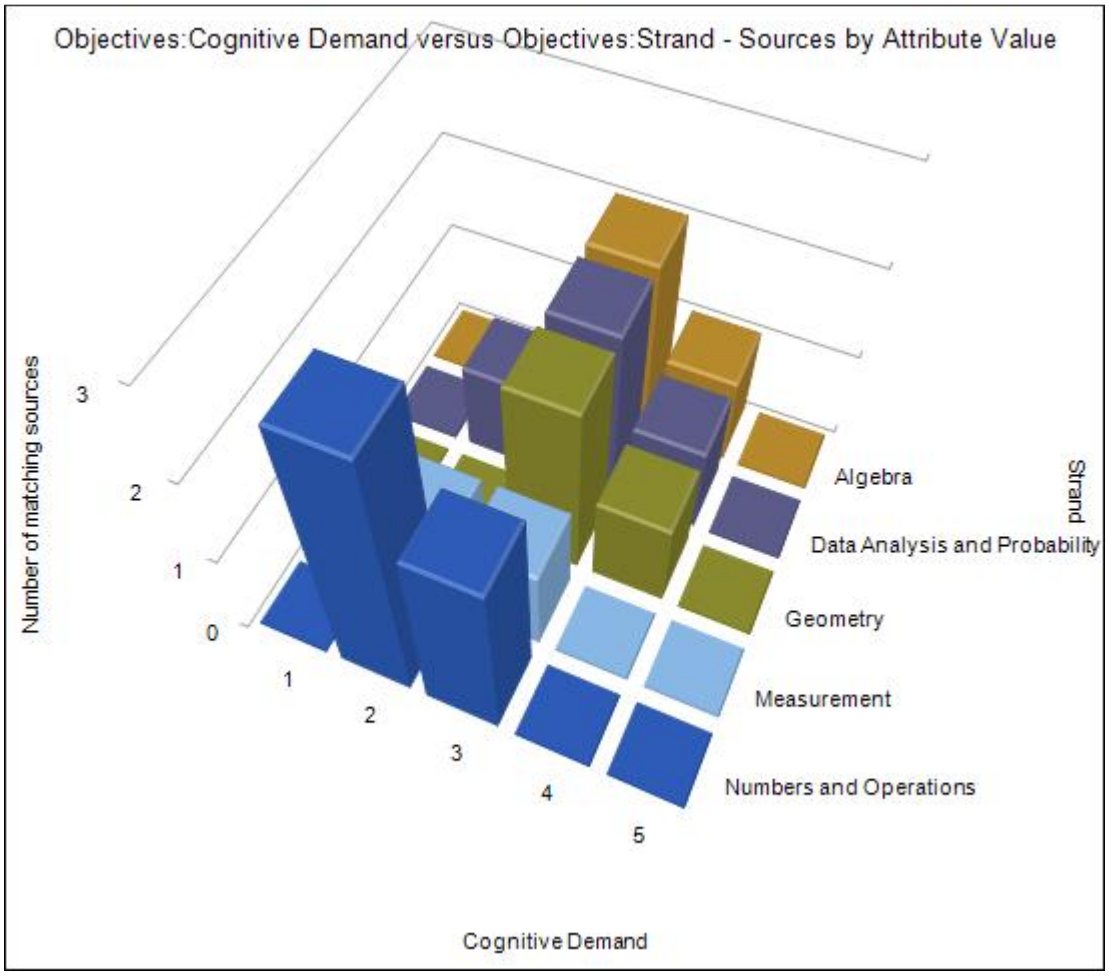
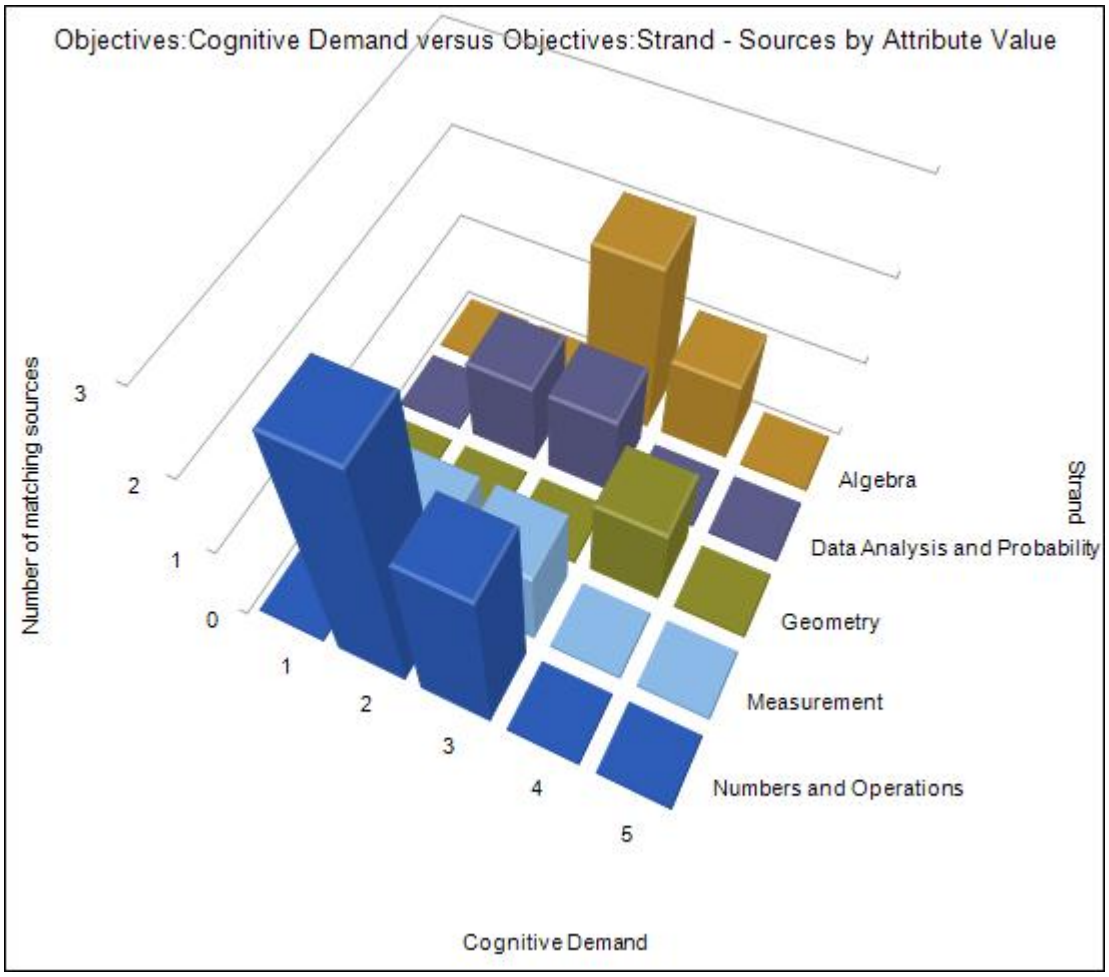


Figure 3 reveals that, similar to the content map in Figure 2, the full standards emphasize numbers and operations at the performance level; however, unlike the content map in Figure 2, the content map for full standards has a more evenly distributed emphasis on all of the other topics at the demonstrate level. The emphasis appears evenly distributed across the strands and cognitive demands, with a concentration at the demonstrate level

across all strands. This suggests that the items on the formative assessment in Figure 2 should be more evenly distributed across the objectives and not quite as heavily concentrated on algebra and numbers and operations.

The content map in Figure 4 examines the district-specified content standards for third quarter, which are the standards indicated by the district as the subset of the NCSCoS that teachers should focus on in their instruction for their students to do well on the formative assessment. These abbreviated standards are designed to provide focus for teachers in the third quarter. The third quarter standards do not include a total of four objectives that are included in the full standards. Because this subset of standards is specified by the district to accompany the formative assessment, the assessment should align more clearly to these district-specified standards than the full standards, which are meant to be taught throughout the entire school year.

Figure 4: Content Map for Third Quarter District-specified Standards



Looking across the three content maps, the content map shown in Figure 4 looks more similar to the formative assessment content map (Figure 2) than the full standards' content map (Figure 3), suggesting that the formative assessment is well aligned to the third quarter district-specified standards. Comparing the content maps for the items (Figure 2) and for the third quarter standards (Figure 4), the items in Figure 2 are mostly assessing the material at the cognitive demand level of perform and the objectives in Figure 4 are more evenly distributed across all standards at the perform and demonstrate level, with two objectives at the generate level. Basically, in order for the formative assessment to be well aligned to the full or third quarter standards, the items in the formative assessment need to be

more evenly distributed across the five content strands and have a more concentrated focus on measuring skills at the demonstrate level.

The SEC results also include a statistical index, designed to express the alignment between two content maps. The formula to calculate the index compares the proportion of the number of objectives or items in each congruent cell on the two content maps. A cell is defined as a topic-by-cognitive demand intersection on the table used to create the content maps. In Table 10, the indices for the formative assessment compared to the full standards and third quarter standards are both weak. This is not due to the small number of objectives (17 objectives) compared to the formative assessment (49 items) because the formula standardizes each cell, but could be a result of the small number of cells. Indices above .83 are considered strong; indices between .70-.82 are considered acceptable; and indices below .69 are considered weak (Fulmer, 2011).

Table 10: Formative Assessment Indices for SEC Alignment Method

Formative Assessment compared with:	Index	Level
Third Quarter Standards	.640	Weak
Full Standards	.575	Weak

The content maps suggest that improvements should be made to the formative assessment to more evenly distribute items across the cognitive demands and strands, placing less emphasis on numbers and operations and algebra. The index suggests weak alignment. In the next section, the results will be further compared to examine whether or not the results present a consistent picture of alignment.

Results Comparison

In order to understand if the three alignment methods produce consistent or inconsistent results, the statistical results across methods were compared. Table 11 displays the percentage of the statistical data that fit into each cutoff across methods. The level of alignment indicated by majority for each criterion is highlighted in gray. The cutoffs are depicted in Table 12. Dashed lines indicate that the cutoffs for the level of alignment are not specified by a method.

Table 11: Comparisons across Alignment Methods in Percentages

	Webb				Achieve						SEC
Alignment Level	Cat. Con.	DOK	ROK	Bal. of Rep.	Con. Cen.	Per. Cen.	Source of Chall.	Range	Bal.	Chall.	Porter Index (5x5)
Strong	---	---	---	---	---	---	---	40	0	---	0
Acceptable	80	80	80	80	12.2	46.9	91.8	40	20	100	0
Weak	---	0	0	0	85.7	48.9	---	20	80	0	100
Unacceptable	20	20	20	20	2.04	4.08	8.2	---	0	---	---

Table 12: Comparisons across Alignment Methods Cutoffs

	Webb				Achieve						SEC
Alignment Level	Cat. Con.	DOK	ROK	Bal. of Rep.	Con. Cen.	Per. Cen.	S. Of Chall.	Ran.	Bal.	Chall.	Porter Index (5x5)
	At the standard level				At item level			Item Sets			
Strong	---	---	---	---	---	---	---	.67+	Score of good	---	.83+
Acceptable	6 items +	50% +	50%+	70% +	Score of 2	Sc. of 2	Score of 1	.50-.66	Score of app.	Score of med.	.70-.82
Weak	---	40-49%	40-49%	60-69%	Score of 1A or 1B	Sc. of 1A or 1B	---	Less than .50	Score of fair	Score of easy or hard	Less than .69
Unacceptable	Less than 6 items	Less than 40%	Less than 40%	Less than 60%	Score of 0	4.08	Score of 0	---	Score of poor	---	---

To answer my first research question about the alignment of the formative assessment to the standards across the three methods, the percentage of results supporting each alignment level across methods is presented in Table 11. The overall results did not differ across comparisons of the formative assessment to the full standards versus the district-specified standards, which was unexpected but might be explained by the spiraling curriculum in the district, which returns frequently to previously taught concepts. The Webb method demonstrated overall acceptable alignment; the Achieve method indicated mixed levels of alignment; the SEC indicated weak alignment. In summary, the alignment methods did not portray consistent findings about the alignment of the formative assessment to the standards; therefore, the alignment between the formative assessments and the standards is unclear and different depending on the method selected.

In order to answer my second research question and make inferences and general conclusions about the differences in alignment results across the three methods, the results were compared on two common criteria: breadth and depth. In order to understand differences and similarities across the methods, results for Webb's categorical concurrence, ROK, and balance of representation were compared with the content centrality, range, and balance for item sets for the Achieve method, which was compared to the topic categories for the SEC method. The ratings that are good or acceptable have been highlighted to make the table more readable. If no items were rated inconsistent for content centrality, the criterion was considered to have been acceptably met.

Table 13: Comparison of Breadth Criteria across Methods on Formative Assessment

	Webb			Achieve			SEC
Standard	Categorical Concurrence (# of Hits)	Range of Knowledge (% of Obj. Hit)	Bal. of Rep. (Index)	Content Centrality (Ratings for items)	Range (% of Obj. Covered)	Balance (Overall Rating)	Topic Categories (#of Items)
Numbers and Operations	26 Accept.	80 Accept.	.56 Unaccept.	Consis.: 1 Partial: 17 Incon.: 1	40 Unaccept.	Poor	17
Measurement	6 Accept.	100 Accept.	.95 Accept.	Consis: 2 Partial: 2 Incon: 0	100 Good	Appropr.	6
Geometry	7 Accept.	66.67 Accept.	.93 Accept.	Consis: 0 Partial: 7 Incon: 0	66.66 Accept.	Poor	7
Data Analysis and Probability	5 Weak	25 Unaccept.	1 Accept.	Consis.: 0 Partial: 5 Incon.: 0	50 Accept.	Poor	5
Algebra	16 Accept.	100 Accept.	.90 Accept.	Consis.: 3 Partial: 11 Incon.: 0	100 Good	Poor	12

Table 13 shows the findings by standard and across breadth criteria and, within each cell, the numbers of items or hits for each method. The items associated with measurement were the strongest across the criteria and were consistently rated as high quality across methods. The items under the geometry and algebra standards were acceptable on Webb's criteria and were rated as partially or clearly consistent on content centrality and acceptable or good on range. However, the balance judgment given by the raters on geometry and algebra was poor. The items relating to numbers and operations and data analysis and probability did not rate consistently across the breadth criteria. However, none of the ratings on the criteria relating the standards were clearly consistent with one another, suggesting that each criterion shows something different related to breadth.

After examining the breadth criteria, the depth criteria were compared by examining the evaluating the congruence of DOK consistency in the Webb method, performance centrality for the items, source of challenge for the items, level of challenge for the item sets in the Achieve method, and cognitive demand in the SEC method. The ratings that are good

or acceptable have been highlighted to make the table more readable. A challenge in comparing the alignment data is gauging what cutoffs are acceptable or unacceptable for Achieve method, and SEC does not have cutoffs; the items are listed in descriptive form as they relate to the standards.

Table 14: Comparison of Depth Criteria across Methods on Formative Assessment

	Webb	Achieve			SEC
Standard	DOK consistency (% at or above DOK)	Perf. Cent.	Source of Challenge (judgments for items)	Level of Challenge (judgment for item sets)	Cognitive Demand (by level)
Numbers and Operations	87 Accept.	Consis.: 3 Partial: 14 Incon.: 1	App: 16 Inapp:2	Medium	Perform: 15 Demon: 2
Measurement	100 Accept.	Consis.: 4 Partial: 0 Incon.: 0	App: 4 Inapp: 0	Medium	Perform: 2 Demons.: 3 Gener.: 1
Geometry	28.57 Unaccept.	Consis.: 3 Partial: 4 Incon.: 0	App: 7 Inapp: 0	Medium	Perform: 5 Demon.: 2
Data Analysis and Probability	60 Accept.	Consis.: 4 Partial: 0 Incon.: 1	App: 4 Inapp: 1	Medium	Perform: 5
Algebra	71.43 Accept.	Consis.: 8 Partial: 6 Incon.: 0	App: 11 Inapp: 2	Medium	Perform: 3 Demons: 9 Gener.: 2

In sum, the comparisons across depth shown in Table 14 suggest that the results for the standard measurement are consistently rated acceptable across criteria. Numbers and operations, data analysis and probability, and algebra were somewhat consistently rated as acceptable across methods. Another observation is that the results for depth are more consistent with one another than the results across criteria for breadth. Finally, the DOK consistency for geometry was rated unacceptable for Webb, but none of the items were rated as inconsistent with the objectives in the Achieve method.

Thus the depth and breadth criteria across methods indicate mixed levels of alignment. The Achieve method resulted in the results indicated the lowest levels of alignment, and the Webb method suggested overall acceptable alignment. The SEC

suggested weak levels of alignment. In the next section, a recommendation is made to use aspects of the alignment methods congruently in order to gain a more descriptive and accurate understanding of alignment for tests and standards. The goal of combining aspects of methods is to not over or under estimate alignment and result in valid and reliable evidences of alignment.

Conclusions, Recommendations, and Limitations

This study is the first comparison of the most commonly-used alignment methodologies across a common assessment and standards. This study is also the first to compare the alignment of a formative assessment to state content standards. In this section, I discuss the implications of on the results, make recommendations for the uses of alignment methods, and suggest future studies on alignment methodologies. I also discuss the limitations of the present study and its contribution to the literature on alignment.

Conclusions and Recommendations

The findings of this study suggest that in order to ensure the alignment of the formative assessments because the present methods do not result in consistent results, test developers need to combine results across methods to obtain a full and accurate picture of alignment. The results of each method contribute differently to understanding alignment. The accuracy of the formative assessment scores is essential for system coherence and for teachers to know the areas in which students are struggling academically. If the system is aligned coherently, the formative assessments will be indicative of performance on the summative test. The results of this study indicate that the formative assessments' alignment to the standards needs to be improved in order for the system to work towards a common, clear goal. Quality instruction and academic success require strong content standards, formative assessments, and summative tests. Frustration on behalf of teachers and families will result if the content on the formative assessment is aligned to the content of the summative assessment such that it acceptably predicts success on that assessment. Because

the formative assessments used by districts are used by teachers for instructional guidance and intervention purposes, test developers have a responsibility to provide high quality assessments that align well with the standards and are predictive of the success of summative test. Looking at the alignment results for the formative assessment included in this study, the quality of the formative assessment is not nearly as well aligned to the standards as an educator would expect or hope, which suggests that instruction is being tailored using assessment items that do not fully assess the standards and do not fully allow for opportunity to learn or opportunity to teach before the summative test.

According to the Webb method, the alignment of the summative test to the standards is somewhat better than the formative assessment's alignment to the standards. The alignment of the summative test according to the Webb method is acceptable across all standards and criteria; however, one standard on each criterion is rated unacceptable or weak for the formative assessment. This is concerning, considering that the Webb method is perhaps overly generous with cutoffs for the criteria. On the formative assessment, measurement and algebra are the two standards rated as acceptable across all criteria. The standard data analysis and probability is rated weak for categorical concurrence because less than six items are available. Data analysis and probability is also rated unacceptable for range of knowledge because only one of four objectives is hit with items. The 31 hits associated with the standard number and operations are distributed poorly across the objectives under the standard, and depth of knowledge was unacceptable for geometry because the geometry items are written at a level 1. These flaws in the test need to be addressed in order for the assessment to validly assess student ability before the summative test, the scores of which determine the allocation of federal and state incentives and consequences.

The formative assessment includes flaws across all criteria according to the Webb method, but compared to the other methods, the Webb results suggest that the formative assessment is somewhat aligned to the standards; when in fact, the other two methods suggest that the alignment between the formative assessment and the standards is mixed or unacceptable. According to Martone and Sireci (2009), the Webb method provides the most detailed statistics. Although rich in a quantitative sense, the Webb method has a tendency to over-estimate alignment compared to the results of other methods. This claim is made on the basis that although all three of the methods suggest that the formative assessment is not strongly aligned to the content standards, the Webb method suggests somewhat acceptable evidence of alignment, indicating only one standard as unacceptable on range, balance, and DOK consistency. Overall in the Webb method, at least 80% of the standards were met at an acceptable level of each criterion. Using the results of the Webb method, a test developer could claim that the alignment of the formative assessment to the standards is acceptable. The Achieve method emphasizes another perspective of alignment. Many objectives are multi-faceted and sometimes items only assess parts of the objectives. For example, the objective may state that students need to describe, generalize, and predict geometric transformations, but the items only require students to describe. The findings of this study show that the majority of the items do not assess the full objectives for content or performance centrality and four out of five item sets (standards) were rated poor for balance. The SEC results across Figure 2 through 4 and the Porter indices indicate weak alignment. The accuracy of the Porter index is subject to table size (larger tables results in higher indices); thus in some cases, the Porter index may under-estimate alignment. In summary, in

order to fully understand alignment, components across methods must be utilized, which will be discussed further in this section with five recommendations.

The Achieve results highlight the importance of the content standards. During the alignment session, the standards included in this study were described by the teachers in the Achieve method as frequently vague, wordy, or too broad. The quality of the standards drives the quality of the test. Based on the findings of the Achieve method, the first step that educators and policy makers must take is to adopt strong, clear, and specific standards. This is the researcher's first recommendation to improve alignment methodologies: content standards must be focused on specific and measurable skills with specific indication of expected cognitive demand. Overall, the Achieve results suggest that the items on the formative assessment do not match the content and performance specified in the objectives. For example, less than 13% of the items matched their assigned objectives clearly for content centrality. The items matched slightly better for performance centrality than content centrality. Almost 47% of the items clearly matched the performance centrality of the objectives. Clearer, more specific standards would result in more accurate alignment results. Because the standards included in this study are multi-faceted, including lists of verbs for expected performances and content for expected skills within each objective, an item may measure part of the objective but not fully measure the objective itself. The Webb method does not account for objectives with multiple expectations; whereas, the Achieve method rates the quality of the match as completely consistent or partially consistent between the item and the objective. If the match is not completely consistent, it is the responsibility of the test developer to ensure that other items measure the various expectations set forth in objectives with multiple performances and skills. Adoption of clear standards would make

alignment clear and measureable. The current NCSCoS are not amenable to clear results using the alignment methods.

The second recommendation based on this study is that rating the match between the item and the assigned objective should be a component of all alignment studies because the ratings allow test developers to understand which compound objectives are not fully assessed by the items. To rate the quality of the match between objectives and items, participants should use a scale similar to the Achieve scale. Based on the alignment sessions in this study, looking at the content centrality, raters should assign 2 for clear consistent, meaning that the item and objective match clearly; 1 for somewhat consistent, meaning that the item assesses part of a compound objective or an objective that is too broad to capture in one item, or 0 for inconsistent, meaning that the objective and item do not match. Content centrality looks only at the match between the content in the objective and the content being assessed by the item. Considering that raters should also be making item-objective matches autonomously, combining 1A and 1B makes rating efficient and provides information for the test developer on what items need to further examination to ensure that all parts of compound and vague objectives are being assessed. If partially consistent as indicated by a score of 1, raters could circle the portion of the objective measured by the item, and an analysis could be conducted to ensure that the entire objective was measured by at least one item on the assessment. Looking at Tables 5 and 6 for the Achieve method, only 12.24% of the content and 46.94% of the performances in the objectives were completely consistent with the expectations of the objectives, suggesting that the Webb method does not factor in the multiple expectations of multi-faceted objectives to an acceptable extent.

The third recommendation is that raters should make item-objective matches autonomously as in the Webb method. This action allows raters more autonomy in the matching process and for various interpretations of the items, as well as verification of the test blue print. During the Achieve alignment session, the participants were at times frustrated with the test blueprint and would have appreciated the opportunity to assign objectives themselves to the items. Even though the participants were encouraged in the Achieve alignment session to disagree with the test specifications, they did not suggest changes to the document, perhaps because they felt the blue print was already created and suggesting changes was not the goal of the session. The participants expressed confusion with the clarity of the objectives and in identifying the most essential piece of the objectives, even when reminded by the researcher that they were the experts and could talk about their thoughts in an effort to reach consensus.

Webb's balance of representation index needs to be improved to include all objectives and not limited to the objectives with hits. The Achieve method does this to an extent by asking SMEs to qualitatively compare the objectives under a standard to the items linked to a standard, but without offering a statistic. Balance across the standards is an important component of alignment and a more demanding quantitative method for expressing balance of representation needs to be developed to fully understand alignment. A more demanding and accurate quantitative measure of balance could be calculated using the current Webb calculation if objectives were clear and not multi-faceted, and the statistic required that all objectives had at least one hit for the calculation to be valid.

The fourth recommendation is that regardless of the levels of cognitive demand used in an alignment study, the raters should label the cognitive demand of the objective and item

to allow for Webb's depth of knowledge calculation, and raters should rate the performance centrality of the item-objective match using a scale similar to the suggestion for content centrality. Raters should assign 2 for clearly consistent, meaning that the item and objective match clearly; 1 for somewhat consistent, meaning that the item assesses part of a compound objective or an objective that is too broad to capture in one item, or 0 for inconsistent, meaning that the objective and item do not match.

The fifth recommendation is to include the content maps from the SEC method, which capture a more complete picture of alignment between tests and standards than what is currently offered by any one of the current methods. These maps are useful because they contribute to a more large-scale, big picture understanding of alignment between an assessment and standards.

Overall, the alignment results are dependent on the method chosen by the test developer. This is concerning because this study's findings suggest that results of any one of the current alignment methodologies will not be consistent with results of other methodologies. In order for the education system to work coherently towards student academic success and improved academic outcomes, the alignment between the instruction, assessments, and standards is vital. Further steps should be taken to improve the current standards, such that the objectives are not multi-faceted. District and classroom formative assessments must expect the same or higher levels of achievement than summative tests, since the summative tests are typically not used for instructional purposes and intervention design. Finally, current alignment methodologies have room for improvement, specifically in their justification for the criteria cutoffs and in their application to the standards.

Limitations

This research has several limitations. The first limitation concerns the setting and small number of participants. The alignment analyses were conducted within one district across two public schools with six educators. The alignment analyses included the results of six graduate students, some of whom had little to no classroom experience. The trainings associated with the alignment analyses were less than an hour, and the participants had little to no experience with alignment methodologies prior to participating in the study.

Second, concluding that the state test is better aligned to the state content standards than the formative assessments may not be completely accurate because only the Webb alignment method was used to compare the state test and formative assessment to the state content standards, and the Webb method tends to over-estimate alignment. Other alignment methodologies may highlight significant flaws in the state test compared to the state content standards.

Third, the study involved only one subject area and one grade level. Therefore, the results may not generalize across other subject areas. Further research should be conducted to understand how content area influences the findings of alignment studies. The results might be different depending on the sample, such as including different teachers or graduate students.

Finally, the research questions are limited because each alignment method may be developed to offer a purposefully different perspective on alignment compared to other methods, not necessarily comparable to other methods. Some comparisons across methods are based on the researcher's perception. Test developers should gather the suggested

alignment evidence before looking at one method's results and concluding that the assessment or test is well-aligned to the standards.

Summary and Future Studies

Overall, the findings of this study suggest that in order to fully understand alignment between tests and standards, components of various common alignment methods should be used. The statistics associated with the Webb method are useful in understanding the extent of alignment, but the statistics do not capture the complete picture of alignment. Along with making matches between items and objectives, raters should rate the quality of the match using the component from the Achieve method. Content maps can be generated to accompany the Webb method, which can help visually assess alignment. This is important because the Webb method alone has a tendency to suggest that the alignment is stronger than the results of the Achieve method and SEC method tend to indicate.

Further research should be conducted examining alignment results across different grade levels and content areas. Replication of the current study would be interesting to see if results generalize to other assessments. A study combining components across methods would be interesting to see if a combination method portrays a more complete and useful understanding of alignment. A combination of matching and rating could lead to a new methodology. Comparisons of how results differ when using matching and rating or only matching or rating would be interesting to know how the procedure affects the outcome of alignment. When assessing the alignment of a test and the standards, researchers and test developers should make judgments about alignment based on as much evidence as possible.

Appendix A: Webb Alignment Method DOK Level Definitions

Level 1 = Recall and Reproduction

An item or objective that requires the recall of information such as a fact, definition, term or a simple procedure, as well as performing a simple algorithm or applying a formula.

Level 2 = Skills and Concepts

An item or objective that requires the engagement of some mental processing beyond recalling a response; requires students to make some decision as to how to approach the problem or activity, whereas Level 1 requires students to give a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps.

Level 3 = Problem Solving and Strategic Thinking

An item or objective requiring complex or abstract reasoning, planning, using evidence, drawing conclusions, and/or justifying an approach to a problem that has multiple solutions. This is a higher level thinking than the previous two levels.

Level 4 = Extended Thinking

An item or objective that requires complex reasoning, planning, developing, and thinking, generally requiring an extended period of time. Students are required to make several connections—relate ideas within the content area or among content areas—and have to select one approach among many alternatives on how the situation can be solved.

(Herman, Webb, & Zuniga, 2007)

Appendix B: The Webb Alignment Method

1. Please rate each objective for DOK level
   1 = Recall
   2 = Skills and Concepts
   3 = Problem Solving and Strategic Thinking
   4 = Extended Thinking

Objective	DOK	Comments
1.01		
1.02		
1.03		
1.04		
1.05		
2.01		
2.02		
3.01		
3.02		
3.03		
4.01		
4.02		
4.03		
4.04		
5.01		
5.02		
5.03		

2. Please identify objectives to which the items correspond, identifying both a primary and secondary topic if appropriate.
3. Please judge the depth of knowledge (DOK) associated with each item
   1 = Recall
   2 = Skills and Concepts
   3 = Problem Solving and Strategic Thinking
   4 = Extended Thinking

Form 1: Calculator Active

Item	Primary Match	Secondary Match	DOK	Comments:
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				

DOK Levels:

1 = Recall
2 = Skills and Concepts
3 = Problem Solving and Strategic Thinking
4 = Extended Thinking

Form 1: Calculator Active (cont.)

Item	Primary Match	Secondary Match	DOK	Comments:
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				

DOK Levels:

1 = Recall
2 = Skills and Concepts
3 = Problem Solving and Strategic Thinking
4 = Extended Thinking

Form 1: Calculator Inactive

Item	Primary Match	Secondary Match	DOK	Comments:
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				

DOK Levels:

1 = Recall
2 = Skills and Concepts
3 = Problem Solving and Strategic Thinking
4 = Extended Thinking

Form 2: Formative Assessment

Item	Primary Match	Secondary Match	DOK	Comments:
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				

DOK Levels:

1 = Recall
2 = Skills and Concepts
3 = Problem Solving and Strategic Thinking
4 = Extended Thinking

Form 2: Formative Assessment (cont.)

Item	Primary Match	Secondary Match	DOK	Comments:
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				
39				
40				
41				
42				
43				
44				
45				
46				
47				
48				
49				

Appendix C: Achieve Alignment Method Definitions

Content Centrality

2 = Clearly consistent: The item assesses the exact content articulated in the objective.

1A = Not specific enough: Objective is too broad to confidently judge item alignment.

1B = Somewhat consistent: Item samples only part of the objective.

0 = Inconsistent: The item only marginally assesses what is prescribed by the standard.

Performance Centrality

2 = Clearly consistent: The item and the objective require the same type and number of cognitive tasks.

1A = Not specific enough: The objective is too broad.

1B = Somewhat consistent: The item samples only part of the cognitive demands expressed in the objective.

0 = Inconsistent: The cognitive demand of the test item and objective do not match.

Source of Challenge

1 = Appropriate: The item difficulty is appropriately located in the subject matter and performance demanded by the objective.

0 = Inappropriate: The item's difficulty stems from extraneous sources such an inappropriate grade-level language, misleading graphs, or unfair assumptions about a student's background knowledge.

Roach, Niebling, & Kurz, 2008

Appendix D: Achieve Alignment Method

1. **Please rate the content centrality of the match between the item and the objective**
   2 = Clearly consistent
   1A = Not specific enough
   1B = Somewhat consistent
   0 = Inconsistent

2. **Please rate the performance centrality of the match between the item and the objective**
   2 = Clearly consistent
   1A = Not specific enough
   1B = Somewhat consistent
   0 = Inconsistent

3. **Rate the source of challenge on whether the item's difficulty is due to appropriate or inappropriate sources of challenge.**
   1 = Item difficulty is appropriately located in the subject matter and performance demanded by the objective

   0 = Item's difficulty stems from extraneous sources such an inappropriate grade-level language, misleading graphs, or unfair assumptions about a student's background knowledge

Standard 1: Number Sense

All items under a particular objective have been considered a set for each test.

1. What objectives in a standard seem to be overassessed?

2. What objectives in a standard seem to be underassesed or not assessed at all?

3. Based your reading of the standards and your personal judgment of what is most relevant for the particular grade level in question, is the balance of this test good, appropriate, fair or poor?

4. Looking at the set of items relating to objectives and thinking about the cognitive demands of the entire set in relation to the demands specified in the matching objectives as well as items skewing toward more or less challenging concepts, types, or parts of objectives, rate each set as easy, medium or hard. Write a short evaluation of the set's level of challenge.

Standard 2: Measurement

1. What objectives in a standard seem to be overassessed?

2. What objectives in a standard seem to be underassesed or not assessed at all?

3. Based your reading of the standards and your personal judgment of what is most relevant for the particular grade level in question, is the balance of this test good, appropriate, fair or poor?

4. Looking at the set of items relating to objectives and thinking about the cognitive demands of the entire set in relation to the demands specified in the matching objectives as well as items skewing toward more or less challenging concepts, types, or parts of objectives, rate each set as easy, medium or hard. Write a short evaluation of the set's level of challenge.

Standard 3: Geometry

1. What objectives in a standard seem to be overassessed?

2. What objectives in a standard seem to be underassesed or not assessed at all?

3. Based your reading of the standards and your personal judgment of what is most relevant for the particular grade level in question, is the balance of this test good, appropriate, fair or poor?

4. Looking at the set of items relating to objectives and thinking about the cognitive demands of the entire set in relation to the demands specified in the matching objectives as well as items skewing toward more or less challenging concepts, types, or parts of objectives, rate each set as easy, medium or hard. Write a short evaluation of the set's level of challenge.

Standard 4: Data Analysis

1. What objectives in a standard seem to be overassessed?

2. What objectives in a standard seem to be underassesed or not assessed at all?

3. Based your reading of the standards and your personal judgment of what is most relevant for the particular grade level in question, is the balance of this test good, appropriate, fair or poor?

4. Looking at the set of items relating to objectives and thinking about the cognitive demands of the entire set in relation to the demands specified in the matching objectives as well as items skewing toward more or less challenging concepts, types, or parts of objectives, rate each set as easy, medium or hard. Write a short evaluation of the set's level of challenge.

Standard 5: Algebra

1. What objectives in a standard seem to be overassessed?

2. What objectives in a standard seem to be underassesed or not assessed at all?

3. Based your reading of the standards and your personal judgment of what is most relevant for the particular grade level in question, is the balance of this test good, appropriate, fair or poor?

4. Looking at the set of items relating to objectives and thinking about the cognitive demands of the entire set in relation to the demands specified in the matching objectives as well as items skewing toward more or less challenging concepts, types, or parts of objectives, rate each set as easy, medium or hard. Write a short evaluation of the set's level of challenge.

Appendix E: SEC Alignment Method Cognitive Demand and Strand Definitions

Level 1 = Memorize facts/definitions/formulas

- Recite basic mathematics facts
- Recall mathematics terms and definitions
- Recall formulas and computational procedures

Level 2 = Perform Procedures

- Use numbers to count order, or denote
- Do computational procedures or algorithms
- Follow procedures or instructions
- Solve equations, formula, and routine word problems
- Organize or display data
- Read or produce graphs and tables
- Execute geometric constructions

Level 3 = Demonstrate Understanding of Mathematical Ideas

- Communicate mathematical ideas
- Use representations to model mathematical ideas
- Explain findings and results from data analysis strategies
- Develop and explain relationships between concepts
- Show or explain relationships between models, diagrams, and/or other representations
- Develop flexibility in thinking and reasoning

Level 4 = Conjecture/Generalize/Prove

- Determine the truth of a mathematical pattern or proposition
- Write formal or informal proofs
- Recognize, generate, or create patterns
- Find a mathematical rule to generate a pattern or number sequence
- Make and investigate mathematical conjectures or predictions
- Identify faulty arguments or misrepresentations of data
- Apply mathematical properties and rules to reason inductively or deductively

Level 5 = Solve Non-routine Problems/Make Connections

- Apply and adapt a variety of appropriate strategies to solve non-routine problems (i.e., solve a novel problem by collecting and analyzing data)
- Apply mathematics in contexts outside of mathematics
- Analyze data and recognize patterns
- Synthesize content and ideas from several sources

Kurz et al, 2010

SEC Alignment Method Strand Competency Goals for Fourth Grade
(http://www.ncpublicschools.org/curriculum/mathematics/scos/2003/k-8/24grade4)

NO = Number and operations

The learner will read, write, model, and compute with non-negative rational numbers.

M = Measurement

The learner will understand and use perimeter and area.

G = Geometry

The learner will recognize and use geometric properties and relationships.

DP = Data analysis and probability

The learner will understand and use graphs, probability, and data analysis.

A = Algebra

The learner will demonstrate an understanding of mathematical relationships.

Appendix F: Surveys of an Enacted Curriculum

3. Please rate each objective for cognitive demand.

1 = Memorize facts/definitions/formulas
2 = Perform procedures
3 = Demonstrate understanding of mathematical ideas
4 = Conjecture/generalize/prove
5 = Solve non-routine problems/make connections

4. Please match each objective to a strand.

NO = Number and operations
M = Measurement
G = Geometry
DP = Data analysis and probability
A = Algebra

Objective	Cognitive demand	Strand	Comments
1.01			
1.01a			
1.01b			
1.01c			
1.01d			
1.02			
1.02a			
1.02b			
1.02c			
1.02d			
1.02e			
1.03			
1.04			
1.04b			
1.04c			
1.05			
2.01			
2.02			
3.01			
3.02			
3.03a			
3.03b			
3.03c			

Objective	Cognitive demand	Strand	Comments
4.01			
4.02			
4.03			
4.04			
5.01a			
5.01b			
5.02			
5.03			
5.03a			
5.03b			

Please rate each item for cognitive demand.
1 = Memorize facts/definitions/formulas
2 = Perform procedures
3 = Demonstrate understanding of mathematical ideas
4 = Conjecture/generalize/prove
5 = Solve non-routine problems/make connections

Please match each item to a strand.
NO = Number and operations
M = Measurement
G = Geometry
DP = Data analysis and probability
A = Algebra

Formative Assessment			
Item	Cognitive demand	Strand	Comments:
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
21			
22			

Item	Cognitive demand	Strand	Comments:
23			
24			
25			
26			
27			
28			
29			
30			
31			
32			
33			
34			
35			
36			
37			
38			
39			
40			
41			
42			
43			
44			
45			
46			
47			
48			
49			

References

AFT Teachers. (2006, July). *Smart testing: Let's get it right.* Retrieved from http://www.aft.org/pdfs/teachers/pb_testing0706.pdf.

Beach, R. W. (2011). Issues in analyzing alignment of language arts common core standards with state standards. *Educational Researcher, 40*(4), 179-182.

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.

Brown, R. S., & Niemi, D. N. (2009). Content alignment of high school and community college assessments in California. *Journal of Applied Research in the Community College, 16*(2), 109-118.

Common Core. (2011). Retrieved from http://www.corestandards.org/.

Common State Assessments (2011). Retrieved from http://www.ets.org/k12/commonassessments.

D'Agostino, J. V., Walsh, M. E., Cimetta, A. D., Falco, L. D., Smith, S., VanWinkle, W. H., & Powers, S. J. (2008). The rating and matching item-objective alignment methods. *Applied Measurement in Education, 21*(1), 1-21.

Flowers, C., Browder, D., & Ahlgrim-Delzell, L. (2006). An analysis of three states' alignment between language arts and mathematics standards and alternate assessments. *Exceptional Children, 72*(2), 201-215.

Flowers, C., Wakeman, S., & Browder, D. M. (2009). Links for Academic Learning (LAL): A conceptual model for investigating alignment of alternate assessments based on alternate achievement standards. *Educational Measurement: Issues and Practice, 28*(1), 25-36.

Fulmer, G. (2011). Estimating critical values for strength of alignment among curriculum, assessments, and instruction. *Journal of Educational and Behavioral Statistics, 36*(3), 381-402.

Herman, J. L., Webb, N. L., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measurement in Education, 20*(1), 101-126.

Kurz, A., Elliott, S. N., Wehby, J. H., & Smithson, J. L. (2010). Alignment of the intended, planned, and enacted curriculum in general and special education and its relation to student achievement. *The Journal of Special Education, 44*(3), 131-145.

Lui, X., Zhang, B., Liang, L. L., Fulmer, G., Kim, B., & Yuan, H. (2008). Alignment between the physics content standard and standardized test: A comparison among the United States-New York State, Singapore, and China-Jiangsu. *Science Education, 93*(5), 777-797.

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1359.

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

Polikoff, M. S., Porter, A. C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? *American Educational Research Journal, 48*(4), 965-995.

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

Roach, A. T., McGrath, D., Wixson, C., & Talapatra, D. (2010). Aligning an early childhood assessment to state kindergarten content standards: Application of a nationally recognized alignment framework. *Educational Measurement: Issues and Practice, 29*(1), 25-37.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools, 45*(2), 158-176.

Stuart, S. & Rinaldi, C. (2009). A collaborative planning framework for teachers implementing tiered instruction. *TEACHING Exceptional Children, (42)*2, 52-57.

U.S. Department of Education. (2011). Retrieved from http://www2.ed.gov/ programs/racetothetop/ index.html.

Webb, N.L. (1997). Criteria for alignment of expectations and assessments in mathematics and science education (NISE Research Monograph NO. 6). Madison: University of Wisconsin-Madison, National Institute for Science Education.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and curriculum. *Applied Measurement in Education, 20*(1), 7-25.

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The roles of reviewer agreement. *Educational Measurement: Issues and Practice, 26*(2), 17-29.