Review of Educational Research

http://rer.aera.net

Evaluating Alignment Between Curriculum, Assessment, and Instruction

Andrea Martone and Stephen G. Sireci

REVIEW OF EDUCATIONAL RESEARCH 2009; 79; 1332 originally published online Sep 18, 2009;
DOI: 10.3102/0034654309341375

The online version of this article can be found at:
http://rer.sagepub.com/cgi/content/abstract/79/4/1332

Published on behalf of

http://www.aera.net

By

http://www.sagepublications.com

Review of Educational Research
December 2009, Vol. 79, No. 4, pp. 1332–1361
DOI: 10.3102/0034654309341375

Evaluating Alignment Between Curriculum, Assessment, and Instruction

Andrea Martone
The College of Saint Rose
Stephen G. Sireci
University of Massachusetts Amherst

The authors (a) discuss the importance of alignment for facilitating proper assessment and instruction, (b) describe the three most common methods for evaluating the alignment between state content standards and assessments, (c) discuss the relative strengths and limitations of these methods, and (d) discuss examples of applications of each method. They conclude that choice of alignment method depends on the specific goals of a state or district and that alignment research is critical for ensuring the standards-assessment-instruction cycle facilitates student learning. Additional potential benefits of alignment research include valuable professional development for teachers and better understanding of the results from standardized assessments.

KEYWORDS: assessment, test theory and development, test validity and reliability, teacher education and development, psychometrics.

A great deal of discourse and debate exists, both professional and political, regarding state-mandated testing including testing under the No Child Left Behind (NCLB) legislation. The main criticisms of mandated testing in our nation's schools are reduced teaching time, a narrowed curriculum, limited opportunity to assess higher order thinking skills, and decreased morale of teachers and students (Roach, Niebling, & Kurz, 2008; M. L. Smith & Rottenberg, 1991). There is evidence, however, to support the view that mandated testing provides a necessary lens to view the educational opportunities presented to students. Without a means to understand what goes on in the classroom and a way to compare how students are performing, it is difficult to truly understand if all students are provided with adequate educational opportunities. Well-designed tests provide important data to learn about student performance and aid in decisions regarding funding (Cizek, 2001).

Although politicians, educators, and parents debate the merits of standardized testing, the psychometric characteristics of the tests are rarely the basis of concern. Rather, criticisms have focused on "opportunity to learn" issues such as failure to test students on what they are taught and a narrowing of the curriculum because of mandated testing (Resnick, Rothman, Slattery, & Vranek, 2004; Roach et al., 2008). Ideally, to address such claims, researchers must demonstrate that what is covered on mandated tests aligns with what occurs in the classroom, both

in terms of the curriculum and the instruction. Alignment research is one means to demonstrate or evaluate the connection between testing, content standards (i.e., curriculum), and instruction. If these components work together to deliver a consistent message about what should be taught and assessed, students will have the opportunity to learn and to truly demonstrate what they have achieved.

In this article, we discuss different methods used to evaluate alignment and the types of information alignment studies can provide. Although there is very little research on the use of alignment research in the classroom (Roach et al., 2008), the results of an alignment study could potentially help policymakers, assessment developers, and educators make refinements so curriculum, assessment, and instruction support each other in what is expected of students. Alignment research may also allow the public to understand how testing does or does not support what is purported to occur in classrooms and what changes may be needed in components of educational systems.

As part of the NCLB legislation, alignment between state standards and assessments is a prerequisite to achieving adequate yearly progress (U.S. Department of Education, 2002). States must demonstrate how their assessment tools align with their state standards (e.g., Johnson, 2005; Leffler, Carr, Griffin, & Gates, 2005). Norman L. Webb[1] (1997) stated "Better aligned goals and measures of attainment of these goals will increase the likelihood that multiple components of any district or state education system are working toward the same ends" (p. 2). Beyond just the alignment of standards and assessments, the instructional content delivered to students also needs to be in agreement. If this were not the case, if teachers are teaching what they want irrespective of what the curriculum calls for, students could potentially do well in the classroom and then fail on the assessments without understanding where they need additional help (McGehee & Griffith, 2001). Through alignment research, policymakers and educators can see where they are headed and will know where they stand relative to agreed on goals.

In this article, we review three popular methods for evaluating alignment. Our review focuses on the use of alignment methodology to facilitate strong links between curriculum standards, instruction, and assessment. The purpose of our review is to describe why an understanding of alignment should be an important characteristic of a statewide testing process. Our review is structured around three areas and builds on earlier descriptions of these alignment processes (Bhola, Impara, & Buckendahl, 2003; Council of Chief State School Officers [CCSSO], 2002; Porter, 2006). First, we present an overview of how alignment is defined in the educational measurement literature. This overview includes formal definitions of alignment and describes how alignment builds on earlier notions of content validity. In the second section, we describe the three most widely used alignment evaluation methods. Although these methods share some common components, a closer look at each approach highlights the relative strengths and limitations of each method. We also provide an example of a specific application of each methodology. In the final section, we discuss the importance of considering the goals of an alignment study before determining the best method to use. We argue there are many potential reasons for conducting an alignment study and we believe some studies can serve as a form of professional development for teachers and others involved in curriculum development. We argue that the process of alignment research itself, more than just the results, can help educators see how assessments can connect to what happens in the classroom.

Overview of Alignment

Alignment means many things in the world of education. La Marca, Redfield, Winter, and Despriet (2000) point out that the dictionary defines "to align" as "to bring into a straight-line; to bring parts or components into proper coordination; to bring into agreement, close cooperation" (p. 1). In a classroom setting, instructional alignment refers to agreement between a teacher's objectives, activities, and assessments so they are mutually supportive (Tyler, 1949). On a schoolwide level, curricular alignment refers to the degree to which the curriculum across the grades builds and supports what is learned in earlier grades (Tyler, 1949). Alignment, as described in this review, takes curricular alignment a step further to look at "the degree to which expectations [i.e., standards] 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" (Webb, 1997, p. 4). In describing an aligned educational system, La Marca and colleagues (2000) emphasized that the assessments must allow students to demonstrate their knowledge and skills with respect to the expectations set up in the curriculum frameworks so that proper interpretations of their performance can be made. As they put it,

> Alignment is . . . the degree to which assessments yield results that provide accurate information about student performance regarding academic content standards at the desired level of detail, to meet the purposes of the assessment system . . . . 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. (p. 24)

In a perfect world, what a student is tested on should be derived from what is expected of the student as detailed in the state or district standards, as well as from what is taught to the student by his or her teachers. Although not everything that is listed in the standards or taught to the student can or should be assessed, alignment research can illuminate how much and to what degree the standard coverage or instructional content has been assessed. The theory underlying alignment research is that a consistent message from all aspects of the educational structure will result in systemic, standards-based reform (M. S. Smith & O'Day, 1991). Porter (2002) describes this type of consistent message as follows:

> An instructional system is to be driven by content standards, which are translated into assessments, curriculum materials, and professional development, which are all, in turn, tightly aligned to the content standards. The hypothesis is that a coherent message of desired content will influence teachers' decisions about what to teach, and teachers' decisions, in turn, will translate into their instructional practice and ultimately into student learning of the desired content. (p. 5)

Assessments, standards, and instruction are all integral to student achievement, but they have each been determined and enacted at multiple levels of the educational structure. State content standards (embodied in state curriculum frameworks) represent state level policy documents, but the policymakers do not create the assessments, and the curriculum standards and assessments are implemented

at the local level. Alignment studies allow researchers systematically to study the different components of an educational system to compare their content and make judgments about how well they are in agreement.

Webb (1997) noted that the Education Goals 2000 Act supported the development of a consistent message about student learning between the policy, assessment, and instruction perspectives. As he stated, that act "indicated alignment of curriculum, instruction, professional development, and assessments as a key performance indicator for states, districts, and schools striving to meet challenging standards" (p. 1). Additionally, NCLB requires that a state's academic achievement standards be aligned with the state's academic content standards. If the alignment between academic achievement and content standards is low, a state is likely to have trouble meeting the requirements of NCLB. Alignment research culminates in a report about the relationships of the components that can be used for future decision making rather than just a simple yes or no response (Rothman, Slattery, Vranek, & Resnick, 2002). The results of an alignment study should provide a measure of how well assessments cover the underlying standards. Some alignment approaches also provide information regarding the degree to which assessments and standards match classroom instruction. Once the degree of alignment is understood, subsequent changes in any of the educational components can be made to improve the standards-assessment-instruction cycle.

In summary, alignment studies provide data that can be combined with the priorities of educational stakeholders to guide changes in assessments, standards, and/or instruction. By focusing on the match between test content and what is intended to be taught, alignment research shares some common goals and methodology with traditional methods for studying content validity. In the next section, we discuss some similarities between contemporary evaluations of alignment and traditional studies of content validity.

The Relationship of Alignment to Content Validity

Generally defined, content validity refers to the degree to which a test appropriately represents the content domain it is intended to measure. When a test is judged to have high content validity, its content is considered to be congruent with the testing purpose and with prevailing notions of the subject matter tested. Thus, content validity does not specify particular aspects of the educational process such as curriculum frameworks or instruction. Rather, it is more general and refers to tests both within and outside educational systems (e.g., licensure and certification tests).

As we describe in a subsequent section, there are several different aspects of an alignment study, and the specific aspects within a given study depend on the methodology used. With respect to a content validity study, there are at least four potential aspects—domain definition, domain representation, domain relevance, and appropriateness of the test construction procedures (Sireci, 1998a, 1998b). Domain definition refers to the process used to define operationally the content domain tested. In the case of K–12 achievement testing, the domain is typically derived from state-established curriculum frameworks. Domain representation refers to the degree to which a test represents and adequately measures all facets of the intended content domain. To evaluate domain representation, inspection of all the items and tasks on a test must be undertaken. Studies of domain representation typically use subject matter experts (e.g., teachers) to scrutinize test items and

judge the degree to which they are congruent with the test specifications (Crocker, Miller, & Franks, 1989; Sireci, 1998a). Domain relevance addresses the extent to which each item on a test is relevant to the domain tested. An item may be considered to measure an important aspect of a content domain and so it would receive high ratings with respect to domain representation. However, if it were only tangentially related to the domain, it would receive low ratings with respect to relevance. Appropriateness of test development procedures refers to all processes used when constructing a test to ensure that test content faithfully and fully represents the construct intended to be measured and does not measure irrelevant material. The content validity of a test can be supported if there are strong quality control procedures in place during test development and if there is a strong rationale for the specific item formats used on the test.

Traditional studies of content validity use subject matter experts (SMEs) to rate test items with respect to their congruence to the test specifications or their relevance to the intended domain. Hence, traditional content validity studies and contemporary alignment studies are similar in that they both gather data from SMEs, and they structure the data collection procedures in a way that independently evaluates specific aspects of content domain representation. The specific tasks given to the SMEs differentiate content validity and alignment studies.

Sireci, Robin, Meara, Rogers, and Swaminathan (2000) provided an example of a traditional content validity approach to alignment using the Grade 8 1996 National Assessment of Educational Progress (NAEP) Science Assessment. A primary goal of their study was to evaluate the congruence between the NAEP Science Framework and the NAEP Science Assessment. Ten carefully selected SMEs reviewed a sample of NAEP Science items and were asked to assign each item to (a) one of the three content areas ("fields of science"), (b) one of the three cognitive levels ("ways of knowing and doing science"), and (c) one of the four "themes of science" listed in the NAEP test specifications (framework). Each item was given an item congruence index rating based on the number of raters who agreed with the original classification. For example, if an item was intended to measure Earth Science and 8 out of 10 SMEs rated it as Earth Science, it had an item-content area congruence rating of .8. Following the suggestion of Popham (1992), an index of .7 and greater was used to judge an item as adequately congruent with its content area, cognitive level, or theme. (See Sireci, 1998a, for other examples of traditional and innovative content validity studies in several contexts.)

Although the traditional content validity approach involves rating or matching items to more global levels within test specifications (such as "domains," "strands," or "content areas"), contemporary alignment research uses the same expert rating approach but delves deeper to examine the match between items and the objectives or benchmarks *within* a strand. For example, a state's curriculum framework may have the strand Grade 4 Number Sense (4N), which is the level at which test specification tables are typically written. However, within strand 4N there are multiple objectives. For example 4N-1.1 might be "Read, write, order, and compare numbers up to 1,000,000." In this example, the objective provides the detail regarding the specific skill being measured by an item. Alignment research often matches items to these detailed objectives and then reports findings summarized by objective and/or by strand. Additionally, in some cases alignment research considers what was actually taught to the students. In this way, alignment research can

offer a deeper view of the educational process, and can be thought of as an extension of a more traditional content validity evaluation. However, as we discuss later, traditional content validity studies may have some advantages for evaluating the congruence of a particular test form to its test specifications.

Valid educational assessment requires significant overlap between the assessment and the curriculum measured to ensure the decisions made based on test results are defensible. As the *Standards for Educational and Psychological Testing* state, "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). In educational assessments related to NCLB, the proposed uses of tests include evaluation of students' current proficiencies and their progress with respect to state-defined performance standards. Thus, an evaluation of the appropriateness of the state test for such purposes involves consideration of both the test content and the state standards. To understand students' performance on the test, the instruction received by the students must also be considered. Because alignment research considers all three of these aspects, it provides validity evidence for evaluating not only the tests, but also the curriculum and the instruction.

Although the definition of validity from the *Standards* cited above is succinct, there have been many different "types" or "aspects" of validity that have been proposed for educational tests (Sireci, in press). In addition to content validity, alignment research has also been associated with evaluation of testing consequences.[2] Research questions relating to the consequences of achievement tests that may be addressed in an alignment study include "Have state-mandated tests led to changes in teachers' instruction?" (Porter, Smithson, Blank, & Zeidner, 2007) and "Do mandated assessments narrow the curriculum?" (Achieve, Inc., 2006).

Alignment research may address potential assessment or instructional deficiencies by systematically comparing the different pieces of the educational process. If educational components are not well aligned, the system will not send a consistent message about what is valued in the educational process (Webb, 1999). Thus, alignment research can be used to evaluate concerns that the curriculum has been dumbed down (Linn, 2000), that students have not received a fair chance to learn the material on which they were tested (Winfield, 1993), and that states have not addressed the need to improve instructional quality (Rothman et al., 2002). These evaluations are important extensions of the information provided by a typical content validity study that focuses on how well test items represent the domains specified in a test blueprint.

Approaches to Alignment Research

The development and application of alignment methods came about from a desire to ensure that students' test scores reflect their performance with respect to specific curricular expectations (La Marca, 2001). Some alignment studies have focused on the content of the standards compared to the assessments, and others have included the content of instruction. In the following section, we elaborate on the three most common alignment methods—the Webb, Achieve, and Surveys of Enacted Curriculum methods. An application of each of these methods is also presented to illustrate their processes and findings.

Webb Methodology

Webb developed a comprehensive and complex methodology to investigate the degree of alignment between assessments and standards. His method explores five different dimensions to understand the degree of alignment: content focus, articulation across grades and ages, equity and fairness, pedagogical implications, and system applicability (Webb, 1997). However, only the area of content focus has been applied in alignment studies. Therefore, this review focuses on the applied piece of the Webb methodology. In Webb's method, "standards" are the broad content domains within a subject and the skills within this domain are referred to as "objectives." Understanding these definitional terms is critical to seeing how the alignment process has been applied because these terms and levels of analyses differ across the different alignment methods.

Webb Alignment Dimensions

Webb's content focus dimension comprises six subcategories for analysis: categorical concurrence, depth of knowledge, range of knowledge, balance of representation, structure of knowledge, and dispositional consonance. Each of these subcategories explores the relationship between the assessment and the standards in a different way. However, only the first four (categorical concurrence, depth of knowledge, range of knowledge, balance of representation) have been applied in alignment studies so those will be discussed in depth here. Together these subcategories contribute to a thorough understanding of the degree of alignment between assessments and standards. An important aspect of the Webb methodology is the term "hit," which is any item–objective match. Given that participants could match an item to up to three objectives, each item could potentially have three hits. As originally formulated, the hits did not need to be within the same standard, although this flexibility has been removed in recent years because that approach could lead to concluding more standards are supported than actually are (Webb, 2007).

Categorical concurrence. This subcategory compares the similarity of the expectations for student learning, as expressed through the content categories in the standards, to the assessments. Categorical concurrence is most similar to traditional content validity and is a minimum requirement in alignment research. Like the test blueprint comparison in a traditional content validity study, categorical concurrence looks at broad content areas, such as number sense and geometry. The total number of item–objective matches, hits, within a standard is averaged across all participants to determine the average number of items per standard. Webb (2002) suggested using a criterion of at least six hits measuring a standard for successful alignment of a test on this dimension. His logic was that at least six items would be needed if students were to receive scores on a standard because fewer than six items would not likely result in scores of sufficient reliability. Using this approach, if there are four standards, an assessment needs at least 24 hits to establish categorical concurrence. However, unlike a traditional content validity study where a test item is matched to its standard by SME consensus (e.g., 70% of SMEs match an item to its intended standard[3]), Webb's criterion is simply that, across the SMEs, an average of at least 6 hits is matched to the standard. That is, a standard

could theoretically be considered adequately represented even if the items matched to it were specified to measure a *different* standard in the test blueprint.

Depth of knowledge. This subcategory of consistency compares the complexity of knowledge expressed in the specific objectives within each standard to the complexity of knowledge in each item that is matched to that objective. Webb initially defined the cognitive areas as recall, skill and/or concept, strategic thinking, and extended thinking, but these areas may be modified for a particular study (Webb, 1999). The main criterion is that what is tested should be at or above the same cognitive level as what is expected to be taught based on what is in the standards. To have alignment relative to this criterion, at least 50% of the items matched to an objective must be at or above the cognitive level of that objective (Webb, 2002). Fifty percent is based on the assumption that most cutoff points require students to answer more than half the items to pass, but some flexibility is allowed with this criterion. The main concern in this aspect of alignment is that assessment items should not be targeting skills that are below those required by the objectives to which the item is matched.

Range of knowledge. This subcategory of consistency analyzes the breadth of the standards as compared to the breadth of an assessment. This dimension looks at the number of objectives within a standard measured by at least one assessment item. To have sufficient alignment relative to range of knowledge, at least 50% of the objectives within a standard need to be measured by at least one assessment item (Webb, 2002). This logic assumes that students should be tested on at least half of the domain of knowledge. This part of the alignment process also assumes all of the objectives have equal weighting and all of the objectives accurately cover the skills needed to complete that standard. The level of complexity within a state's standards influences this aspect of alignment, as more complexly written objectives might be only partially assessed but would still be considered a match from the perspective of this dimension.

Balance of representation. This subcategory focuses on the degree to which items are evenly distributed across objectives within a standard to represent the breadth and depth of the standards. Given the limited time for assessment, this dimension highlights what aspects of the standards are prioritized. Balance of representation focuses on the objectives assessed by the items and then looks at the proportion of objectives measured compared to the number of items. The calculation for the balance index is:

$$1 - \frac{\left(\sum_{k=1}^{0} \left|\frac{1}{O} - \frac{I_k}{H}\right|\right)}{2} \quad (1)$$

where O = total number of objectives hit for the subject domain; $I_{(k)}$ = number of items corresponding to objective (k); and H = total number of items hit for the subject domain (Roach, Elliott, & Webb, 2005). If the proportion approaches zero, it signifies one or more objectives are measured by relatively fewer items. If it approaches one, it signifies the items are evenly distributed across all objectives.

Ideally, in time, assessments should shift in the balance of representation to cover the entire standards.[4] Evaluating balance of representation across grades can also demonstrate shifts in priorities as the content develops.

These first four areas of Webb's content focus dimension—categorical concurrence, depth of knowledge, range of knowledge, and balance of representation—are used by alignment researchers as the basis for their alignment studies. These four aspects serve as the most direct way to view the degree of match between an assessment and the standards.

Application of Webb's Method

Webb (1999) applied his methodology in a study of mathematics and science assessments and standards in four states. Here, we focus on the mathematics alignment process and results. The purpose of Webb's study was to better understand how his alignment methodology functioned, to examine in greater detail the different alignment dimensions, and to understand ways to improve the alignment process. Six reviewers compared the match between assessment items and standards and/or objectives in mathematics. The results of this matching were used to judge the degree of alignment based on four of Webb's criteria: categorical concurrence, depth-of-knowledge consistency, range-of-knowledge consistency, and balance of representation.

The review process involved multiple decision points by the reviewers. Applying this process across four states, the reviewers noted differences between the standards in terms of content covered, level of detail, and overall organization, which impacted the comparability of the states. The first step was a review of each state's standards to match each objective to a depth of knowledge level representative of the highest level of knowledge needed to achieve that objective. This process allowed for systematically linking items to objectives and cognitive levels. The reviewers reached an agreement about the depth-of-knowledge of the objectives based on a group discussion. These decisions were used as a baseline comparison to the assessment items to determine if the items were at or above the cognitive level in the objective.

The items within an assessment were then matched to the objectives within the standards and coded based on the depth of knowledge required by that item. Any match was called a "hit." However, one item could be matched to more than one objective. This increased the content and range alignment criteria areas but proved to be an area of confusion for the reviewers. The reviewers also noted when items appeared not to match any objective. The results were aggregated to report by standard. The mean and standard deviation for each criterion were computed for each reviewer.

The results showed varied levels of alignment across grade levels and states. The strongest area of alignment was for the categorical concurrence criterion. Three out of the four states fulfilled this criterion with at least six items measuring a standard, but in each state one fourth or more of the standards were measured by fewer than six items. The balance-of-representation criterion was satisfied because the standards that were assessed had items evenly distributed among the objectives.

The weakest aspects of the alignment were the depth-of-knowledge consistency and range-of-knowledge criteria. The results demonstrated that test items generally targeted a lower level of knowledge and did not sufficiently cover the range of knowledge laid out in the standards. This finding lends some support to the

common criticism that standardized testing does not test complex thinking and narrows the curriculum by testing a small part of the content domain. Armed with these results, the states could accurately address these issues in their assessment design. This study also demonstrated that each of the four criteria measured different aspects of alignment.

Webb (1999) noted that the reviewers could have benefited from more training at the beginning of the process. Some reviewers wanted to code near matches instead of exact matches, which confused the analysis. The reviewers needed more guidance about making distinctions relative to the depth-of-knowledge criteria and more explicit guidance about how to match an item to more than one standard based on the central content of an item. Webb also found that it could be helpful to put the standards in context so the reviewers know each state's purpose for the standards and how they were created. During the review process, the reviewers focused purely on the item–objective match and did not have an opportunity to critique the quality of each component. Webb concluded the reviewers were frustrated by this constraint. Although it is important to stay focused on the task at hand, it could be helpful to gather this feedback throughout the process as a means to inform future standard or assessment development work.

Webb (1999) concluded that trade-offs between these four alignment variables are realistic, but it is important to look at broader approaches to assessment to understand how other pieces (e.g., those discussed in the general Webb methodology, but not specifically studied in his alignment process) complement the process. Unfortunately, these aspects are harder to measure and to include in a formal study and may involve validity issues that go beyond alignment per se. One limitation of this study was that the range of knowledge criterion did not look at the breadth of the measured objective in terms of how many different ideas were combined under one objective. If an objective were very broadly stated, it was still considered assessed if it had an item matched to it, regardless of what else within that objective was not assessed. If objectives combined many different skills, it would be easier to meet the range-of-knowledge criterion as there would be fewer objectives to measure. However, combining skills within a single objective might result in an increased cognitive complexity as students are asked to do more with a range of skills. This might result in a lower depth-of-knowledge conclusion. Another limitation with this study was that it did not capture the fact that assessments may purposefully contain items to measure standards from more than one grade. This misalignment by design should be carefully detailed in the alignment process.

In looking at the alignment study process, Webb (1999) developed a number of recommendations. If the goal were to analyze standards from more than one state, Webb recommended starting with the most detailed state standards. It would be helpful to repeat the alignment study over time to capture the changing content of the assessment and how this may or may not impact the alignment results. More recently, Webb and colleagues (N. M. Webb, Herman, & Webb, 2007) noted that averaging reviewers' ratings across standards and objectives might mask the different views of what the item is truly measuring and inflate the degree of alignment across the four dimensions. Other more recent studies examined setting a minimum reviewer agreement requirement at the standard and/or objective level as to what the item is measuring (Herman, Webb, & Zuniga, 2007; N. M. Webb et al., 2007). Recent applications have also emphasized gathering more qualitative data from the participants. For

example, participants can be encouraged to record observations about the quality of the items and the standards when making their ratings and through a debriefing process (Webb, 2007; Webb, Alt, Ely, Cormier, & Vesperman, 2005). Such comments would help determine whether objectives are too broad or too multifaceted.

In summary, the Webb model is comprehensive and provides a point of reference for the next two models reviewed. The strength of this model is its comprehensive analysis of the objective level detail, its view of alignment through four different dimensions, and the proposed guidelines for acceptable levels of alignment. Sample reports for the Webb methodology can be found in the Web Alignment Tool Training Manual (Webb, Alt, Ely, & Vesperman, 2005). The results of a study using the Webb approach illustrate the relationship between what is being asked of the students, how that is being assessed, and what trade-offs are made in the process.

Achieve Methodology

The Achieve methodology is an alignment protocol that is adapted to reflect the concerns of specific subject areas (English language arts, mathematics, and science). It yields both a quantitative and qualitative alignment comparison of a state's assessment to its related standards. Rothman and colleagues (2002) laid out the components of the initial Achieve methodology, which was designed to judge the quality of the overall assessment and its individual items. Since that time, Achieve's protocol has been further refined. A first step in the method is a verification of the blueprint that maps the test items to the objectives. Then the method is based on a team of carefully trained SMEs reaching consensus on the degree of match between the standards and the assessment based on specific criteria (dimensions). Verification of the test blueprint and using a consensus requirement are significant differences from the Webb methodology. The Webb methodology does not utilize the test blueprint, which could obscure the intentions of the assessment, and the SMEs' results are averaged, which could mask significant disagreements about what the item was measuring. Understanding these two differences provides an important foundation to build on when examining the Achieve methodology.

Achieve Alignment Dimensions

The Achieve methodology is applied in two stages. The first stage is an item-by-item analysis to confirm the test blueprint, determine the content and performance "centrality" of each item compared to the objective to which it is matched, evaluate the source of challenge, and determine the level of cognitive demand. The second stage is a holistic evaluation of a set of items matched to an overarching standard in terms of the overall level of challenge, the balance, and the range.

Like the Webb methodology, "objectives" are defined as the most specific level of outcome (i.e., the smallest level of grain size used by a state in delineating its content standards). Another similarity to the Webb methodology is that the Achieve protocol compares individual items on an assessment to the related objectives based on the skill and type of thinking required. Beyond item level matching, however, this methodology also qualitatively considers how a set of items matched to an overarching standard (e.g., literary response or algebra) functions as a group. Although potentially more time-consuming than other approaches, these qualitative data provide a more thorough understanding of the degree of alignment.

Unlike the Webb approach, Achieve does not have clear cutoff criteria to determine acceptable alignment. Each aspect of the alignment process is analyzed by the expert panel independently and as a part of the overall alignment. Then the results of the study are presented in a comprehensive report to the state. The steps within each stage of the Achieve methodology are detailed below.

Stage 1: Item level analysis. The first stage in the Achieve method focuses on item level detail only. This stage of the analysis will confirm the test blueprint, determine the content and performance "centrality" of each item compared to the objective to which it is matched, evaluate the source of challenge, and determine the level of cognitive demand.

The Achieve methodology begins with a confirmation of the test blueprint. An expert reviewer compares items to the objectives within the state standards that they are mapped to in the blueprint. This comparison is then verified by the SMEs to ensure that every item is matched to at least one objective in the state standards. A match between the test blueprint and the item–objective determination by the SMEs requires only that the item address the *same* content; the level of cognitive demand or the associated objective is not considered. Items that are viewed as inappropriately mapped in the test blueprint are reassigned to a more closely related objective, whereas items that do not match a standard or objective are eliminated from further analysis. Where a state lacks a test blueprint or the blueprint does not allow for fruitful application of the protocol, Achieve constructs a blueprint. In these instances, Achieve provides a brief rationale and communicates the findings to the state. Achieve scrutinizes the test blueprint because of its importance in developing score reports. This step allows for a comparison of the intentions of the assessment with what it actually accomplishes. For example, the blueprint might have items coded to number sense, but the reviewers think a match to algebra is more appropriate. The result is the test might be skewed to measure more algebra than intended and not provide enough items for the number sense area (Achieve, Inc., 2006). Understanding this mapping at the beginning of the analysis provides an important foundation to build from.

Each item can have a primary and a secondary match to the objectives (Rothman et al., 2002). The primary match is used in judging content and performance centrality, source of challenge, and level of cognitive demand (described below). The secondary match is taken into account in evaluating level of challenge, balance, and range. The use of a secondary match is similar to the Webb method where items could be mapped to more than one objective, but this model is more explicit about the degree of match and how it can be used in the alignment process. After the test blueprint has been confirmed, the reviewers delve deeper into the actual content of the item and how it specifically relates to the identified objective.

To judge content centrality, SMEs rate each item based on the degree of content match between the item and the objective it is measuring (Resnick et al., 2004; Rothman et al., 2002). Currently, the rating system uses a 5-point scale where a "2" is a clearly consistent content match; "1A" is a match where the degree of alignment is unclear (generally because the objective is too broad to conclude that the item is clearly consistent with the objective);[5] "1B" is a somewhat consistent match in that the item assesses only part of a compound objective;[6] "1C" is a match but

the objective is too specific to fully meet the item task;[7] and "0" signifies an inconsistent match. This rating dimension addresses a limitation of the Webb (1999) study where a broadly stated objective may be considered adequately measured even if the item addressed only a part of the objective. Unlike the Webb approach, however, there is not a clear guideline as to what determines an acceptable level of alignment relative to content centrality. The ratings are reviewed holistically by the reviewers, who then summarize the findings in a report.

In considering performance centrality, the Achieve protocol focuses on the quality of the match between the performance called for in the item and the performance described by the objective the item is intended to measure. This is similar to Webb's (1997) method, but in the Webb approach the cognitive level of the objectives is coded in the beginning and the performance rating is made simultaneously with the content rating. The Webb method might be more efficient, but the Achieve method allows the reviewers to focus on each aspect of the process in isolation. The performance centrality rating process calls reviewers' attention to the verbs in the objectives as compared to what the items actually demands of the student (e.g., the objective asks the student to complete a pattern and the item is pattern completion). The same 2, 1A, 1B, 1C, and 0 scoring system is used for this dimension. Again, there is not a definitive guideline as to acceptable alignment relative to this dimension. Rather the coding results are reviewed across all items to determine an overall view of the performance centrality. That view is then expressed qualitatively in a summary report.

Source of challenge is measured to ensure that items are fairly constructed and not designed to trick students. The items are reviewed to ensure they are not technically flawed (from a content perspective and by reviewing results from item analyses). For example, mathematical items are reviewed to ensure the reading level is appropriate for the grade level of the assessment and unnecessary reading is not required, whereas reading items are examined to ensure they measure comprehension and not prior knowledge. Reading passages are reviewed by the SMEs to ensure, based on a consensus agreement, the vocabulary, sentence structure, literary techniques, plot line, and organizational structure are all appropriate, based on the grade level of the assessment. Writing prompts are similarly reviewed for accessibility, appropriate vocabulary, clarity of purpose and audience, and inclusion of basic criteria by which the sample will be scored. Each assessment item is scored as 1 for an appropriate source of challenge and 0 for an inappropriate source of challenge. If the item received a 0 for content and performance centrality, then it would receive a 0 for source of challenge, as it is not a good measure of that objective.[8]

Level of cognitive demand is concerned with the kind and level of thinking required by students to respond to an item. The level of demand can stem from the nature of the concept assessed (some concepts are more readily understood than others) or from the kind of thinking required to arrive at a response (an item may demand routine or concrete thinking as opposed to complex reasoning or abstract thinking.) Achieve has refined the way in which it tracks the level of cognitive demand of individual items to better inform the evaluation of overall level of challenge. (J. Slattery, personal communication, December 15, 2006). SMEs formally rate each item on a scale: Level 1 (recall or basic comprehension), Level 2 (application of skill/concept), Level 3 (strategic thinking), to Level 4 (extended analysis,

typically during an extended period of time). Level 4 items are not usually found on large-scale, on-demand tests.

Stage 2: Set-of-items analysis. The next stage in the Achieve methodology is a holistic evaluation of a set of items matched to an overarching standard in terms of the overall level of challenge, the balance, and the range. Level of challenge is a global judgment (not item specific) that qualitatively captures whether the collection of items mapped to a given overarching standard appropriately challenges students in a given grade level. Ideally, items within each standard should range from simple to more complex. SMEs provide a brief written evaluation of the level of challenge for each set of items tied to a specific standard, describing how the "overall demand" compares to that expressed in the standard. They base their judgment, in part, on the level of cognitive demand scores previously assigned to individual items in the set. SMEs look to see if a set of items is skewed toward one level of demand, if the items are focused only on the more demanding or least demanding objectives within a standard and, where there are compound objectives, if the items are skewed toward the most or least demanding part of the overall objective. The next step of the Achieve methodology examines the balance and range of sets of items relative to the expectations expressed in the standards.

Balance, like level of challenge, is a holistic evaluation at the levels of the standards rather than the objectives. It looks at a set of items mapped to a given standard to determine how closely the set of items measures the breadth and depth of the content and performances expressed in the relevant standard. The assumption is that the relative importance the test items give to content and skills should be proportionately similar to what is stated in the standards. The SMEs comment on objectives within a standard that are over- or underassessed, redundant items, and how the overall set of items measures content they think is important for that level. The analysis allows the experts to focus on how they view the balance of the assessment as compared to the standards (Rothman, 2003). Again, this is captured qualitatively and builds on the expert knowledge of the SMEs, which is similar to Webb's (1997) balance criterion, although that measure is quantitative. Webb's balance calculation only determines if the objectives are equally represented, but that might not be meaningful if one area of the standards should be emphasized more through the assessment (Rothman, 2003). The quantitative measure facilitates comparison across states or districts, whereas the qualitative measure provides information more informative to the standards and/or assessment revision process.

The range criterion also considers a set of items matched to a standard, but it measures the standard coverage. Range is a quantitative measure of the proportion of the objectives within a standard that are measured by at least one item. Ranges between .50 and .66 are acceptable and above .67 is considered good coverage. This is similar to Webb's (1997) range calculation, although his methodology uses 50% coverage criterion. It is possible for a test to be well balanced, but have low coverage (and vice versa) and so it is important to consider both of these criteria.

At the close of the alignment review, SMEs look across all of the overarching standards (i.e., at the assessment as a whole) to determine the overall rigor of the assessment and how closely it succeeds in measuring the content and performances described by the standards. When Achieve analyzes state assessments at multiple grade levels, SMEs comment on the comparative strengths and weaknesses of the

assessment system taken as a whole, which provides SMEs with insights regarding the quality of a state's standards. For example, if a great many items are scored a "1A" for content centrality, it signals that many objectives are written at too high a level of generality. Achieve transmits all its findings in a comprehensive, technical report to the state that is kept secure because it contains detailed commentary on actual test items. Achieve also produces a policy level report meant for the state to release publicly. Sample policy alignment reports can be found at www.Achieve.org.

An Application of the Achieve Model

Rothman and colleagues (2002) applied the Achieve methodology to the evaluation of assessments in five states. The process began with a training of expert reviewers. The reviewers represented a diversity of viewpoints and included classroom teachers, curriculum specialists, and content and assessment experts. They were trained through the use of carefully selected items to illustrate each of the rating criteria in the Achieve protocol.

Rothman and others (2002) found that states with objectives written in global terms received low ratings because it was more difficult to determine accurate item–objective matches. Overall, they found that items were well matched to content and performance standards. Most states also fared well with respect to the source of challenge criterion. However, they found that the states were not doing a sufficient job of assessing the full range of standards and objectives and that the most challenging standards and objectives were undersampled or omitted (similar to Webb, 1999). With respect to balance, they found that the sets of items were too focused on the less important objectives, a finding that was also supported by the level of challenge results.

Rothman and colleagues (2002) emphasized the need to focus on the issues of balance and challenge in the design and selection of state assessments. Their study illustrated both the drawbacks and strengths of the Achieve alignment method—the process can be time-consuming and expensive to undertake, but it can result in a thorough understanding of the strengths and weaknesses of a state's assessment system.

Surveys of Enacted Curriculum (SEC) Methodology

Although many teachers may think they are assessing what is taught and vice versa, assessments present different stimulus conditions than those used in the classroom, and teaching and assessing are often "institutionally dichotomized" (Cohen, 1987). Porter and Smithson (2001) developed the SEC alignment methodology to help people involved in the education process see the connection between what is taught in the classroom and what is assessed, and they applied it in 11 states and four urban districts. This methodology was developed to quantitatively compare degrees of alignment for standards, assessments, and instruction across schools and states. The SEC methodology builds on a content validity approach but also measures the instructional content purportedly taught and captures this information at both a detailed and more general level of analysis.

SEC Alignment Dimensions

The SEC alignment methodology comprises alignment analyses of standards, assessments, and instruction by use of a common content matrix or template that allows comparison across schools, districts, or states. The methodology begins

with a coding process where the content and cognitive levels are determined for the standards, the assessment items, and the instructional focus. The frameworks are coded at the smallest unit possible. Coding at the objective level is similar to the Webb and Achieve methods, as the results can be aggregated and reported at the strand level. The assessments are coded at the individual item level. Content experts, teachers, and people familiar with the frameworks code both the standards and the assessments.

There are three main alignment dimensions in the SEC methodology: content match, expectations for student performance, and instructional content. The SEC employs a content matrix of two dimensions: content topic and expectations for student performance (CCSSO, 2002). The task for SMEs is to review items and match them to the topical content and type of thinking required in the matrix. These dimensions are discussed below, as is an application of the SEC methodology.

Content match. In the SEC content matrix for mathematics there is a list of topics across the K–12 levels. One potential disadvantage of this method is that the number of topics can be difficult to manage. However, the benefit is an exhaustive common view of all the content. The results can also be reported at a fine- or coarse-grain level. The fine-grain level displays all of the topics and the coarse-grain level rolls up the results to the 16 broad topic areas, which are similar to strands of content (e.g., number sense and patterns; CCSSO, 2004). Thus, the method provides information similar to that gained from traditional content validity studies, but also provides information at a more micro level, which is more likely to better inform instructional and curricular changes (Porter & Smithson, 2002).

Expectations for student performance. The items, standards, and instruction are also coded based on expectations for student performance. This measure is similar to Webb's depth criterion and Achieve's performance centrality measure. The SEC method utilizes five levels of cognitive demand or expectations for student performance. These are: memorize, perform procedures, communicate understanding, solve nonroutine problems, and conjecture/generalize/prove (Porter, 2002). These nonhierarchical terms were chosen to be more behaviorally oriented and indicate knowledge and skills required of students as a way to help teachers describe the cognitive expectations they hold for students (Porter & Smithson, 2001).

Porter and Smithson recommend using the same cognitive levels for each area of analysis as a means to accurately make comparisons across the instructional content, standards, and assessments. The cognitive areas are an important part of the alignment process to address the criticism that standardized tests "dumb down" the curriculum. Through an evaluation of the match between the cognitive demands of each of the educational components (assessment items, standards, instruction), the alignment measure can accurately reflect where differences occur to address the issue of less challenging curricula. The common mapping language allows alignment results to illustrate comparisons of classroom practice to standards and assessments, as well as comparisons between states, districts, and individual teachers.

Instructional content. Unlike the other two alignment methods, the SEC method includes a measure of instructional content. Porter and Smithson (2002) emphasized

the importance of including an instructional content component because it serves as an intervening variable when looking at student achievement gains because of standards-based reform. Through surveys, teachers code the instructional content as they think about a preselected target class during a specified period of time. Then, the teachers estimate the emphasis allotted to that topic for each of the cognitive areas. This is then summed to determine the proportion of each topic relative to the total instructional time (Porter, 2002).

The SEC methodology provides a snapshot of practice during a period of time, which is useful in determining the extent to which teaching reflects standards and assessments (Blank, Porter, & Smithson, 2001). This is a critical question that is not directly addressed by the Webb or Achieve alignment approaches. The benefit of the survey approach is that it allows data collection from a large number of respondents and is relatively inexpensive. Other data collection approaches such as daily logs or classroom observations will be more expensive, time-consuming, and intrusive on the classroom. Porter (2002) acknowledged the weaknesses of the SEC approach: The findings are limited to what is asked, the approach can be subject to self-report bias because teachers complete the survey at the end of the year, and it may be difficult to capture the complexity of instructional practice. Nevertheless, the survey tool has been piloted in multiple settings (Blank et al., 2001) and is being used to address the many questions educators and policymakers have about patterns and differences in curriculum and instructional practices across classrooms, schools, districts, and states (Roach et al., 2008).

The result of the SEC coding across standards, assessments, and instructional content is that each cell in the two-dimensional matrix (content by performance expectations) represents the proportion of content, assessment, or standards in that cell and these three pieces can then be compared to determine the degree of alignment. Each area matrix is compared to another to determine the degree of alignment. This resulting alignment index is:

$$1 - \left[\left(\sum |X - Y|\right)/2\right] \tag{2}$$

where X represents the cell proportions in one matrix (e.g., assessment topics by cognitive demand) and Y represents the cell proportions in the other (e.g., standard topics by cognitive demand; Porter, 2002). The values range from .0 to 1.0. The results can be presented on topographical map layouts to show the relative areas of concentration and facilitate easier comparisons. The results of an SEC alignment analysis illustrate gaps in the assessment, the curriculum, or the instruction, which can then be used to guide additional discussions about what, if any, steps need to be taken to address these gaps. The SEC methodology does not provide specific guidance for the alignment index to represent acceptable alignment as the Webb methodology does. Instead the alignment index should be viewed in relation to the different educational components being studied. For example, the alignment index between a state test and that state's content standards should be higher than the alignment of that state's test to other standards, assuming the state standards differ (Porter, 2006).

An Application of the SEC Methodology

Blank and colleagues (2001) studied the degree of alignment between instruction and assessments across six states using the SEC approach. As with other alignment approaches, the reviewer role was crucial to this process. Specialists were brought together for a 2-day workshop to code the assessment items and standards. At least four raters independently coded each test. Because one assessment item could potentially assess different areas of content, this procedure limited raters to matching each item with up to three topic areas by student expectation combinations. To understand the instructional content dimension, 600 teachers from 200 schools across six states completed the surveys in eighth-grade mathematics.

The results indicated that the alignment of assessment and instruction within a state was similar to the alignment of assessments across states. That is, the alignment indices derived from cross-state comparisons of tests and standards were similar to those indices derived for comparisons of tests and standards within a state. Ideally the alignment index within a state should be stronger than the index comparing that state assessment to other state standards (assuming the state standards differ significantly). Alignment of the state assessments to NAEP Grade 8 math and reading assessments were also conducted, and they found there was slightly higher alignment between state assessments and instruction within the state than there was between instruction within the state and NAEP. On the zero to one alignment index scale, across the six states the average alignment between state instruction and state assessment ranged from .23 (Grade 8 science) to .42 (Grade 4 math), and the average alignment between state instruction and the NAEP assessment ranged from .14 (Grade 8 science) to .41 (Grade 4 math). However, it should be noted that this study was conducted pre-NCLB and none of the states studied had high stakes attached to the assessments (which would probably influence the degree to which the assessments influence classroom instruction). Nevertheless, the study is a good illustration of how national assessments can be considered in alignment research. Although these indices do not provide detailed information about the nature of misalignment, understanding degrees of emphasis can be helpful on the school level (Eastman, 2008).

The involvement of teachers in the data collection process for the SEC methodology illustrates how the alignment process and results can directly impact teachers and their instruction. The SEC methodology is one way to get inside the "black box" of classroom instruction and examine these practices in the context of a large-scale study, which is necessary to evaluate the effectiveness of any reform initiative (Blank et al., 2001). To gain teachers' participation in SEC studies, it is imperative that it be voluntary and the results not be tied to any accountability measures. Additionally, teachers are given individualized results and provided with training about how to use the results (Blank et al., 2001). Results of SEC studies have been used as the basis for professional development opportunities using the in-depth curriculum data for improving instruction (Blank, 2004; Eastman, 2008).

Porter (2002) summarized the multiple benefits of implementing an SEC approach to alignment. It is an efficient process, once the coders of the assessment and standards and the teachers being surveyed are trained, and the process allows for an objective evaluation of the alignment goals. It also provides a quantitative measure of alignment that can be used to examine the effect of reform policies over time. Because this

approach maps the education pieces to a common language and then compares the results, the process can be used to compare findings across schools, districts, and states. It could also be used to evaluate NAEP–state alignment across the nation, which seems to be of interest, given the disparities that have been noted between state and NAEP assessment results (McLauglin et al., 2005).

The SEC approach has similar limitations to the other alignment approaches. The process begins with the state standards and is only as good as what they are working from. Also, the tests will measure only a sample of the content domain, whereas the standards represent the entire domain (Porter, 2002). Additionally, if the standards are not specific enough it will not be possible tightly to align the assessments (Porter, 2002). This methodology does not include the more detailed criteria beyond content and depth match, which are found in the Webb and Achieve models, and so the methodology is unable to quantify the detailed reasons behind limited alignment. Also, research is needed to understand the degree to which teachers and policymakers understand the concept maps that depict instructional coverage.

The survey process can also be somewhat complex for teachers, given the multiple ways they code their instruction (Anderson, 2002). Although response rates can be as high as 75% (Porter, 2002), the survey response rates can be dependent on how the survey is administered. Blank and colleagues (2001) found that the worst response rates were seen in those schools where teachers were given the surveys to complete on their own at their convenience and the best response rates came from those schools where the teachers gathered as a group to complete the surveys. Response rates were also higher where teachers were compensated or given professional development credit for the time it took to complete the survey. Blank and others (2001) concluded that teachers must perceive some personal value to the information they provide. It was important that the information was confidential and that teachers were provided with individual reports if requested, while ensuring the results would not be used for teacher accountability.

Porter, Polikoff, Zeidner, and Smithson (2008) point out that all alignment methods may be of limited utility if the reliability of the SMEs' ratings cannot be established. Although there has been limited research with respect to evaluating the reliability of alignment data, Porter and colleagues (2008) used generalizability theory to estimate the reliability of SEC alignment results (at the cell level) for both tests and standards. Looking across two grades (3 and 6), two subject areas (math and English language arts), and two states, they found the SEC results for both tests and standards to have generally good reliability, although the indices were low in two situations. On further exploration, they noted the lower indices were due in part to an outlier SME. Based on their results, they recommended using at least five SMEs in SEC alignment studies, and they called for evaluation of interrater reliability whenever alignment results are reported. The use of generalizability theory to evaluate the reliability of alignment results appears promising.

Discussion of Alignment Methodologies

Bhola and colleagues (2003) provided a comprehensive overview of different alignment approaches and classified each according to the degree of complexity entailed in the model. Low complexity models defined alignment as the extent to which the items in a test match relevant content standards (or test specifications)

as judged by content experts rating the degree of match with Likert-type scale ratings. This is the approach taken in more traditional content validity-type studies (e.g., Buckendahl et al., 2000; Sireci, 1998a). In moderate complexity models, content experts decide matches both from content and cognitive perspectives and the result may be a reduction in the number of matches because of this additional constraint. This is the approach used in SEC where the standards, assessments, and instruction are aligned. High complexity models tie in additional criteria to give a broader view of alignment. Webb's (1999) approach and the Achieve approach (Rothman et al., 2002) are examples of this level of detail.

Similarities and Differences Across Methods

The Webb, Achieve, and SEC alignment methods have not yet all been applied in a single study and so the differential utility of the results they provide cannot be accurately described. However, in Table 1, we provide a description of the major aspects of each method, organized by four generic dimensions: content, cognitive, distribution, and item quality.

The Webb approach provides the most detailed quantitative results. Based on the four criteria applied, one can see what aspects of alignment are strong or weak. The Achieve methodology builds on the Webb methodology, with the addition of the source and level of challenge dimensions. These dimensions are a means to capture item quality, which was a limitation of Webb's method. However, more recent applications of Webb's methodology now include a source of challenge criterion (Webb, Alt, Ely, & Vesperman, 2005). The Achieve methodology also provides more qualitative information about overall alignment and the quality of the matches. This latter point is missing in the Webb approach where an item–objective match does not convey if the objective is only partially assessed or too vague to be assessed. In this way the specific coding in the Achieve methodology provides a bit more helpful information in terms of possible changes a state might undertake. The broader qualitative results from the Achieve method are very helpful for a specific state application but might become cumbersome if used for comparison purposes between states. The SEC methodology is the only method that considers the instructional piece of the educational process, which allows for easy comparison of assessments, standards, and instruction across states, districts, and schools. It may also be particularly useful for studying the consequences of a testing program, if comparisons are conducted and compared over time. However, this approach does not probe as deeply as the other two into the quality of the alignment. Thus, these alignment methods have different focuses and each has strengths and limitations in specific situations.

One other point of comparison to mention about these three different methods is that the SEC method is the only one to date that has provided comprehensive data regarding the reliability of the results. Porter and others (2008) used generalizability theory to evaluate the reliability of SEC results and to make recommendations regarding the minimum number of SMEs to use to facilitate reliability of the results. Their results generally supported the reliability of the SEC data analyzed and we recommend that estimates of reliability be provided for all alignment studies, regardless of the method used.

TABLE 1

A comparison of the three most popular alignment approaches

Dimension	Webb	Achieve	SEC
Content	Categorical concurrence: compare standards and assessments Goal: six items per broad content standard	Confirm test blueprint then analyze content centrality Able to capture standards that are too broadly written to be completely assessed	Topic coding: items, standards, and instructional content are mapped to a common content language and organized into logical groupings of topics to allow for comparison of instructional content emphasized in standards, assessments, and instruction
Cognitive levels	Depth-of-knowledge consistency: cognitive demand comparison between objectives and test Cognitive levels: recall, skill and/or concept, strategic thinking, extended thinking Goal: At least 50% of the items matched to an objective at or above the cognitive level of that objective	Performance centrality: cognitive demand comparison between objectives and tests, coded after content match; focuses on verbs used in the standard versus what the item requires Cognitive levels: assigns a level of demand, ranging from 1–4, to each item Level of challenge: captures whether the collection of items mapped to a given standard is appropriately challenging	Expectations for student performance: cognitive demand comparison of items, standards, and instructional focus Cognitive levels: memorize facts; perform procedures; demonstrate understanding; conjecture, generalize, prove, and solve nonroutine problems; and make connections

(continued)

TABLE 1 (continued)

Dimension	Webb	Achieve	SEC
Distribution	Balance of represent: how evenly items are distributed across objectives within a standard Range of knowledge: percentage of objectives within a standard that are measured Goal: all objectives measured by at least two items and at least 50% of the objectives within a standard are measured by at least one item.	Balance: relative importance the test items give to content and performances should be proportionately similar to what is stated in the standards Range: percent of the objectives within a standard measured by at least one item. Captures which objectives within a standard seem to be over- or underassessed, redundant items, and how the set of items measures what content reviewers think is important for that level	NA
Item quality	Source of challenge: ensure items are fairly constructed and do not trick students; Examines reading passages, prompts, rubrics, and anchor papers for grade level appropriateness	Source of challenge: ensure items are fairly constructed and are not designed to trick students; also examines reading passages and prompts, rubrics and anchor papers for writing samples for grade level appropriateness	NA
Sample/ application	http://wat.wceruw.org/index.aspx	http://www.achieve.org/node/280	http://seconline.wceruw.org/secWebHome.htm

Note. SEC = Surveys of Enacted Curriculum.

Importance of Subject Matter Experts (SMEs)

All of the alignment methods depend on SMEs to rate the different components of alignment. In selecting these expert reviewers, all approaches emphasize the importance of knowledgeable SMEs who are familiar with the standards, assessments, and instructional components. It is also critical that the SMEs are familiar with the knowledge and skill levels of the tested population (Sireci, 1998b).

Different types of experts may also bring different views of the content to their analyses that affect their ratings. Herman and colleagues (2007) found that teachers rated items differently as compared to higher education faculty in terms of the depth of knowledge required and the dimensionality of the items. Buckendahl et al. (2000) also found that test publishers ratings can be significantly different from expert reviewers. Additionally, SMEs may be influenced by the fact that they are told the categories that the items, standards, or instructional content must fit into and are constrained by these definitions. Furthermore, they can be influenced by social desirability of what they think is expected, leniency to find a match, and guessing (Sireci, 1998b).

A Comparison of Matching and Alignment Approaches

D'Agostino and others (2008) conducted an interesting study in which they used an experimental design to compare "matching" and "rating" methods for evaluating alignment. The matching method was similar to a traditional content validity approach (having SMEs match items to objectives) and the rating method was similar to a low-complexity alignment approach. The study involved items from a statewide high school math test. They randomly assigned SMEs to one of two rating conditions. In the first condition, SMEs matched each test item to the standard it measured. The matching task allowed for secondary and tertiary matches, if necessary. In the second condition, the SMEs provided three ratings for each item: (a) content alignment, (b) cognitive (intellectual) alignment, and (c) "overall match." The ratings were based on a 3-point scale (*consistent*, *somewhat consistent*, and *not consistent*) for each rating. Although they found similar conclusions across methods for about three fourths of the items, they noted important differences across the methods for the remaining items. About 20% of the items had high alignment ratings, but low proportions of SMEs who matched the item to its specified objective. About half of these items involved graphs. They concluded that the rating method was more time efficient (taking about three fourths of the time it took the SMEs to match items), but the matching method was more comprehensive. As they stated,

> For the most part, matching and rating are not substitute methods. Rating appears to be most suitable for confirming test specifications, whereas matching can be used to confirm specifications or explore other possible [item–objective] connections that were not included in the test specifications . . . . Results from a rating method will not inform [us] if certain items are aligned with other [objectives] than initially presumed . . . . Matching can provide both types of information. It can indicate the degree to which the SMEs chose the [specified objective] or identify the best matching [objective] . . . . Matching also has the capacity to reveal if there are several plausible [objective] matches for an item, or if there are only tenuous connections to the standards. (pp. 19–20)

One reason they gave for the improved information from the matching task is that the likelihood of social desirability is lessened because the SMEs are not informed of the objectives the items are presumed to measure. Although this study does not directly compare the three major alignment methods reviewed here, it illustrates the different types of information that can be provided depending on the tasks required of the SMEs. The matching method is more similar to a Webb or SEC approach, whereas the Achieve method uses both rating and matching tasks.

Challenges in Evaluating Alignment

Alignment research can be difficult to conduct for several reasons. First, all content standards for a state cannot typically be assessed through large-scale standardized assessments. Webb (1977) supported broadly defined assessments to include classroom, district, and statewide assessments to capture a broader view of content standards. However, that does not seem practical and was not done in any of the alignment studies we reviewed. All of the alignment studies used statewide, standardized assessments as their comparison, which is most in line with the expectations laid out in NCLB. A second difficulty is that standards may be written at multiple levels (e.g., objectives, topics, strands) and tests may be written to align with standards at the highest level (e.g., number sense, algebra), but the alignment study may use a more detailed level for the standard comparison (e.g., specific item–objective matching; Ananda, 2003a). Also, standards may be written to different levels of specificity and may be written so generally that many different types of content are incorporated so that determining a match is difficult (Rothman et al., 2002).

Inconsistent interpretation of the standards across SMEs is a fourth area of difficulty in conducting alignment studies. Webb (1997) provided an example of this problem with the phrase "demonstrate a range of strategies" and discussed how it was difficult to interpret and therefore to assess. This point can be addressed in the training of the expert reviewers by determining a set protocol about the level and types of matches that are acceptable. A related problem is items may measure multiple content standards, which can result in error among expert judgments (La Marca et al., 2000). Finally, some standards may not be easily assessed and may be redundant within a level, or tests may be designed to assess multiple grade levels. For these reasons, perfect alignment is never expected (Ananda, 2003a).

Given the range of criteria used in an alignment study, states need to be clear about their alignment goals. For example, some states might not value the goal of the assessments having a balanced distribution of items across objectives within a standard and may want greater emphasis within specific areas (Ananda, 2003b). Most states will want to ensure their tests adequately measure the intended strands or objectives, and so a traditional content validity study that focuses on this congruence or on the dimensions of alignment models that look at this congruence may suffice.

Conclusions

Alignment is a means for understanding the degree to which different components of an educational system work together to support a common goal. In this age of accountability, it is important that state organizations, districts, and schools support each other to send a consistent message to teachers and students about what is required. Alignment research is one method to demonstrate this consistency of message or to understand what changes need to be addressed to ensure

every student has the opportunity to learn the content on which they are assessed, and to demonstrate his or her proficiency. Furthermore, to meet the expectations of alignment under NCLB, states will need to conduct independent analyses of the alignment between their tests and state standards, and if any gaps are discovered, they will need to take corrective action.

All three of the methodologies we reviewed start with the basic evaluation of the alignment of the content and cognitive complexity of standards and assessments. The SEC methodology also includes an instructional component. This methodology is very useful if the goal is to study the enacted curriculum. Also, given the common language framework that is used in the mapping process, the SEC model allows for alignment analyses across textbooks, professional development tools, and many other aspects of the educational process. This methodology can be helpful to understand both the content and cognitive emphases across a wide range of the educational process. The Webb and Achieve methodologies add criteria to understand better the breadth and range of comparison between the standards and the assessments. The Webb approach is useful to understand a degree of alignment. Across a variety of dimensions, the Webb methodology provides clear guidelines about acceptable levels of alignment. This information can serve as helpful information to determine what next steps are needed in the process of revising the assessments and/or the standards. The Achieve methodology builds on many components within the Webb approach but also includes an overarching view of the sets of items to look at the broader quality of an assessment relative to the standards on which it is based. Unlike the Webb approach, however, the Achieve approach does not offer clear guidelines as to acceptable levels of alignment relative to each dimension. An application of the Achieve methodology results in a very comprehensive and informative report that details each aspect of the alignment process. The report offers specific suggestions about what changes are required to improve the assessment and the standards.

When deciding between these three alignment approaches, it is important to understand the financial, time, and personnel resources available, as well as the ultimate goals of the research. However it is accomplished, alignment research should be viewed as an ongoing process to continually understand how the assessment, the standards, and the instruction support each other to deliver a consistent message to students about what is expected.

Through NCLB, student assessments have become a dominant feature of the educational process. An important component of the effectiveness of NCLB is the use of assessments to improve instruction. Teachers need to understand the value of the assessments, how the assessments relate to what they should be teaching, and how to make changes in their approach based on the results they see. Teachers' involvement in alignment research is one way to help teachers become more familiar with the assessments and the standards on which they are based. In fact, as Martone (2007) demonstrated, alignment studies can be valuable professional development activities for teachers and curriculum developers. By evaluating test items and their congruence to state-defined benchmarks, participants in alignment studies are forced to become intimately familiar with state standards and the assessments. This increased familiarity could have positive effects on instruction. By participating in an alignment study, teachers can apply what they are learning through the alignment process in their classroom.

NCLB has generated many studies of state assessment–state standards alignment, but we would like to see more studies, not required by NCLB, that actually look at the degree to which the assessments, standards, and *instruction* are aligned. Such studies, perhaps using the SEC method, could provide valuable information regarding how state-mandated curriculum frameworks and assessments have impacted instruction, particularly if the studies are conducted over time.

Alignment research represents an exciting and powerful means for bringing different parts of the educational system together in a systematic and efficient way. Although the process may be costly, as it is dependent on expert reviewers and takes time, the results send a powerful message about the quality of assessments, standards, and instruction, and what might need to be improved. As states, districts, or schools consider instructional changes, they should be aware of the valuable information different types of alignment studies may provide, as well as the potential benefits to teachers who participate in an alignment study. All three of the alignment methods we reviewed in this article have their advantages and disadvantages, but each can provide important information to educators. It is important for educators to carefully consider the types of information they most need before selecting an alignment method.

More research on determining acceptable levels of alignment is needed. As noted above, an important aspect of alignment research is how the results are used. Future research could explore the results of an application of two or more of these methods to the same data set by the same participants. Analyzing the results, the views of the participants on the process, and how the results are used and interpreted by the stakeholders could inform future alignment processes as well as decisions about acceptable dimensions of alignment. An application of multiple alignment methods with the same population could help us better understand the underlying differences in how the methodologies are applied and used, the reliability of the results of each procedure, and the similarities and differences in the types of information provided by different methods.

Notes

This report was funded in part through a subcontract with the University of Nebraska (project/grant S-900-000136) as part of the Comprehensive Evaluation of the National Assessment of Educational Progress (funded by the U.S. Department of Education's Planning and Program Studies Service). The authors are grateful for this support. We are also grateful for the constructive feedback provided by Chad Buckendahl, Leah Kaira, Jay Noell, and members of the Technical Work Group overseeing this project, including John Dossey, Stephen Elliott, Michael Kane, Cindy Paredes-Ziker, and Jean Slattery on an earlier version of this report. Finally, we are grateful to the editor and three anonymous reviewers who made many helpful suggestions on our original submission.

1. In this article, Norman L. Webb's initials do not appear with his many citations and references or with the discussion of his Webb model. The one reference and accompanying citations by author Noreen M. Webb use her initials.

2. The term "consequential validity" has been proposed to describe the evaluation of testing consequences, but this term is controversial. Readers interested in this debate are referred to the two special issues of *Educational Measurement: Issues and Practice* on this topic that appeared in 1997 (Volume 16, number 2) and 1998 (Volume 17, number 2).

3. This criterion was suggested by Popham (1992) and supported by Sireci (1998a). D'Agostino and colleagues (2008) used a criterion of 50%.

4. Thus, evaluating the specific standards covered over time is necessary to ensure important standards are not being neglected.

5. For example, use logical reasoning and mathematical knowledge to obtain and justify correct solutions (Achieve, Inc., 2006).

6. For example, use measures of central tendency (mean, median, mode) and spread (range, quartiles) to summarize data, draw inferences, make predictions, and justify conclusions (Achieve, Inc., 2006).

7. For example, evaluate functions to generate a graph (but the items do not involve a graph; Achieve, Inc., 2006). Although the 1C rating was not present in the earlier description of the methodology (Resnick et al., 2004; Rothman et al., 2002), it was added in the Achieve, Inc., 2006 study and provides a helpful additional point of analysis.

8. Webb recently included source of challenge as one of his alignment dimensions, although it is captured only through reviewer comments (Webb, Alt, Ely, & Vesperman, 2005).

References

Achieve, Inc. (2006). *An alignment analysis of Washington State's college readiness mathematics standards with various local placement tests*. Cambridge, MA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Ananda, S. (2003a). Achieving alignment. *Leadership, 33*(1), 18–22.

Ananda, S. (2003b). *Rethinking issues of alignment under No Child Left Behind*. San Francisco, CA: WestEd.

Anderson, L. W. (2002). Curricular alignment: A re-examination. *Theory Into Practice, 41*(4), 255–260.

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.

Blank, R. K. (2004, April). *Findings on alignment of instruction using enacted curriculum data: Results from urban schools*. Paper presented at the annual meeting of American Educational Research Association, San Diego, CA.

Blank, R. K., Porter, A. C., & Smithson, J. L. (2001). *New tools for analyzing teaching, curriculum and standards in mathematics and science*. Washington, DC: Council of Chief State School Officers.

Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. M. (2000, April). *Alignment of standardized achievement tests to state content standards: A comparison of publishers' and teachers' perspectives*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*(4), 19–27.

Cohen, S. (1987). Instructional alignment: Searching for a magic bullet. *Educational Researcher*, *16*(8), 16–20.

Council of Chief State School Officers. (2002). *Models for alignment analysis and assistance to states*. Retrieved August 28, 2005, from www.ccsso.org/content/pdfs/AlignmentModels.pdf

Council of Chief State School Officers. (2004). *Coding procedures for curriculum content analyses*. Retrieved March 3, 2009, from www.ccsso.org/content/pdfs/CodingProcedures.pdf

Crocker, L. M., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*, 179–194.

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

Eastman, C. (2008). *Case study: A district using SEC: A vital part of the comprehensive data analysis process*. Retrieved March 2, 2009, from the Council of Chief State School Officers website at http://www.ccsso.org/projects/surveys_of_enacted_curriculum/SEC_Resources/#guides

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

Johnson, H. (2005). *South Dakota assessment letter*. Retrieved March 13, 2009, from http://www.ed.gov/admins/lead/account/nclbfinalassess/sd.html

Leffler, J. C., Carr, M., Griffin, L., & Gates, C. (2005). *Alignment of Montana state standards with state assessments*. Portland, OR: Northwest Regional Educational Laboratory.

La Marca, P. M. (2001). *Alignment of standards and assessments as an accountability criterion*. ERIC Development Team. (ERIC Document Reproduction Service No. ED458288)

La Marca, P. M., Redfield, D., Winter, P. C., & Despriet, L. (2000). *State standards and state assessment systems: A guide to alignment. Series on standards and assessments*. Washington, DC: Council of Chief State School Officers.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Martone, A. (2007). *Exploring the impact of teachers' involvement in an assessment-standards alignment study*. Unpublished doctoral dissertation, University of Massachusetts Amherst.

McGehee, J. J., & Griffith, L. K. (2001). Large-scale assessments combined with curriculum alignment: Agents of change. *Theory into Practice, 40*(2), 137–144.

McLaughlin, D., de Mello, V. B., Blankenship, C., Chaney, K., Hikawa, H., Rojas, D., William, P., & Wolman, M. (2005, February). *Comparison between NAEP and state mathematics assessment results: 2003*. Final report Volume II: Appendix D State Profiles. Washington, DC: American Institutes for Research.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*, 285–301.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3–14.

Porter, A. C. (2006). Curriculum assessment. In J. Green, G. Camilli, & P. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 141–160). Washington, DC: American Educational Research Association.

Porter, A. C., & Smithson, J. L. (2001). *Defining, developing, and using curriculum indicators*. CPRE Research Report Series (No. RR-048). Philadelphia, PA: Consortium for Policy Research in Education.

Porter, A. C., & Smithson, J. L. (2002). *Alignment of assessments, standards and instruction using curriculum indicator data*. Paper presented at the Annual Meeting of American Educational Research Association, New Orleans, LA.

Porter, A. C., Polikoff, M., Zeidner, T., & Smithson, J. (2008). The quality of content analyses of state student achievement tests and content standards. *Educational Measurement: Issues and Practice, 27*(4), 2–14.

Porter, A. C., Smithson, J. L., Blank, R. K., & Zeidner, T. (2007). Alignment as a teacher variable. *Applied Measurement in Education, 20*, 27–51.

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

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

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*, 158–176.

Rothman, R. (2003). *Imperfect matches: The alignment of standards and tests*. National Research Council.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing* (CSE Technical Report No. CSE-TR-566). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Sireci, S. G. (1998a). The construct of content validity. *Social Indicators Research, 45*, 83–117.

Sireci, S. G. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299–321.

Sireci, S. G. (in press). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications*. Charlotte, NC: Information Age.

Sireci, S. G., Robin, F., Meara, K., Rogers, H. J., & Swaninathan, H. (2000). An external evaluation of the 1996 Grade 8 NAEP Science Framework. In N. Raju, J. W. Pelligrino, M. W. Bertenthal, K. J. Mitchell, & L. R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 74–100). Washington, DC: National Academies Press.

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice, 10*(4), 7–11.

Smith, M. S., & O'Day, J. (1991). Systemic school reform. In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the Politics of Education Association* (pp. 233–267). Bristol, PA: Taylor & Francis.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

U.S. Department of Education. (2002). No Child Left Behind Act of 2001, Pub. L. No. 107–110. 115 Stat. 1425 (2002). Retrieved on July 30. 2008, from http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf

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

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

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

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

Webb, N. L., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2005). *The Web alignment tool: Development, refinement, and dissemination*. Washington, DC: Council of Chief State School Officers.

Webb, N. L., Alt, M., Ely, R., & Vesperman, B. (2005). *Web alignment tool (WAT): Training manual draft 1.1*. Retrieved March 17, 2006, from http://www.wcer.wisc.edu/WAT/Training%20Manual%202.1%20Draft%20091205.doc

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

Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *Journal of Negro Education, 62*, 288–310.

Authors

ANDREA MARTONE completed her doctorate in the Teacher Education and School Improvement program at the University of Massachusetts Amherst in 2007 and received her B.A. in economics and sociology from Amherst College and her M.S.T. from Fordham University. She is currently an assistant professor of teacher education at the College of Saint Rose in Albany, NY (martoned@strose.edu). Her research interests include teacher education, standardized and classroom based assessment, and methods of improvement for schools labeled as underperforming.

STEPHEN G. SIRECI is a professor of education and director of the Center for Educational Assessment, 156 Hills South, University of Massachusetts, Amherst, MA 01003; e-mail: Sireci@acad.umass.edu. His areas of specialization include test development, test evaluation, validity theory and practice, and cross-lingual assessment. He is a Fellow of the American Educational Research Association and a Fellow of Division 5 of the American Educational Research Association. He is also the coeditor of the *International Journal of Testing*.