EDUCATIONAL MEASUREMENT:
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2013, Vol. 32, No. 1, pp. 23–33

A Proposed Framework for Evaluating Alignment Studies

Susan L. Davis-Becker and Chad W. Buckendahl, *Alpine Testing Solutions*

Evaluating the multiple characteristics of alignment has taken a prominent role in educational assessment and accountability systems given its attention in the No Child Left Behind legislation (NCLB). Leading to this rise in popularity, alignment methodologies that examined relationships among curriculum, academic content standards, instruction, and assessments were proposed as strategies to evaluate evidence of the intended uses and interpretations of test scores. In this article, we propose a framework for evaluating alignment studies based on similar concepts that have been recommended for standard setting (Kane). This framework provides guidance to practitioners about how to identify sources of validity evidence for an alignment study and make judgments about the strength of the evidence that may impact the interpretation of the results.

Keywords: alignment, validity

Federal requirements under *No Child Left Behind* legislation (U.S. Congress, 2002) specify that state assessment programs document independent evidence of alignment between their assessments and academic content standards. Although an important concept in educational measurement, "alignment" is not explicitly mentioned in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). However, as an extension of content validation there are concepts that have similar intent. For example, *Standard* 1.6 references the need for evidence that links the assessment, the construct of interest (e.g., content standards), and the intended use of scores (e.g., achievement standards). Measurement professionals typically agree on the importance of alignment. However, given the range of alignment methods currently applied to assessments, there is a need for a common understanding of what sources of validity evidence are available that practitioners can use to evaluate the results and conclusions of an alignment study. For alignment studies, we define validity evidence as information about the people involved, the judgmental processes used, results, and conclusions of the study that can be evaluated relative to the use of the alignment information in practice. In this article, we propose a framework that practitioners can use to evaluate (i.e., make judgments about) validity evidence for an alignment study. An evaluative process like this begins with defining guidelines and expectations set forth in the professional literature and then identifying threats to validity that could impact the interpretation of the results.

Alignment research has flourished in recent years due, in part, to the federal policy requirement. This is evidenced in the number of professional conference presentations, published research studies, and professional documents (e.g., monographs, white papers) that have appeared in recent years concerning alignment methodologies, implementation, and interpretation of results (e.g., Bhola, Impara, & Buckendahl, 2003; DePascale, 2007; Frisbie, 2003; Herman & Webb, 2007; Herman, Webb, & Zuniga, 2007; Martineau, Paek, Keene, and Hirsch, 2007; Porter, 2002; Porter, Polikoff, Zeidner, & Smithson, 2008; Porter & Smithson, 2002; Webb, Herman, & Webb, 2007). However, the notion of ensuring that test content adequately represents expectations for a given domain is a longstanding concept in measurement that historically fell in the realm of content validation. Conceptually, alignment extends the scope of these inquiries to additional, explicit dimensions.

Alignment has been characterized a number of ways in published literature. Webb (1997) provided a commonly cited definition for alignment: "the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (p. 3). Bhola et al. (2003) focused more narrowly on the specific link between the domain and the assessment by operationally defining alignment as "the degree of agreement between a state's content standards for a specific subject area and the assessments used to measure student achievement of these standards" (p. 21). La Marca (2001) defined alignment as "the degree of match between test content and the subject area content identified through state academic standards" (p. 3).

Each of these definitions suggests that alignment information should be considered a key source of validity evidence for the use and interpretation of educational test scores in measuring student learning of state content and process expectations. The unified perspective of validity suggests evaluating sources of evidence based on the intended use and interpretation of test scores (e.g., Kane, 2006; Messick, 1989). As a key source of evidence in the test development process, it is important to ensure that test content (e.g., items, cognitive processes, responses, scoring guides) supports these inferences by representing a sampling of the domain of the educational program (e.g., content framework, standards, test blueprint).

Susan Davis-Becker, Alpine Testing Solutions, Inc., 51 West Center Street, Orem, UT 84057; Susan.DavisBecker@alpinetesting.com. Chad W. Buckendahl, Alpine Testing Solutions, Inc., 51 West Center Street, Orem, UT 84057; chad.buckendahl@alpinetesting.com.

The elements of curriculum, instruction, and assessment are interrelated through content and cognitive demand (Shepard, 2000). Within this representation, there are three linkages that are commonly evaluated in alignment research. Specifically, one would desire alignment information to evaluate: (1) the connection between content standards (curriculum) and instruction, (2) the connection between content standards and assessment, and (3) the connection between instruction and assessment. Guiding the meaningfulness of these connections is the intended use and interpretation of assessment results. Additional dimensions may be considered part of alignment evidence (e.g., education policies, specificity of score reports), but this illustration is a practical starting point. To establish reasonable parameters this article, we focused our treatment of alignment on the match between an assessment and content standards in terms of content and cognitive demand.

Alignment Methodologies

There are a number of strategies that have been developed for defining and evaluating alignment between assessment content and content specifications (see Martone & Sireci, 2009 for a review of several methods). Although each method has its own unique approaches for evaluating alignment, at the core are judgments about the match between components of an educational assessment system. In addition, each approach uses slightly different terminology in communicating the procedures. For consistency in this article we will refer to test content (e.g., questions, tasks, score points) as "items," the academic content expectations as "content standards," and subject matter experts making the alignment judgments as "panelists." Here we briefly discuss several commonly used alignment methodologies. For more detail, readers are referred to the primary sources.

Webb's (1997) method is a widely used, multi-dimensional evaluation process for assessing the alignment between test items and content standards. Specifically there are five dimensions considered: content, articulation across grades, equity and fairness, pedagogical implications, and systematic approaches. The dimension most often addressed is that of content which includes several attributes: categorical concurrence (general content match), depth of knowledge (cognitive demand), range of knowledge (span of knowledge within content topics), balance of representation (weight by topic), structure of knowledge (relationships among content concepts), and dispositional consonance criteria (attitudes and beliefs towards content).

This method is comprehensive as it includes criteria beyond some of the others noted here (e.g., Surveys of Enacted Curriculum, Frisbie [2003]). In an application of this method described in Webb (1999), only categorical concurrence, depth of knowledge, range of knowledge, and balance of knowledge were used in determining the alignment between assessments and content standards. Most studies utilizing this approach evaluate these same four elements. This method is popular among states engaging in alignment activities because some of the key phrases of Webb's methodology (e.g., Depth of Knowledge, Balance of Coverage) are included in NCLB's Peer Review Guidance (U.S. Department of Education, 2004) against which state assessment systems are evaluated.

A second common method, described by Porter (2002), is the Surveys of Enacted Curriculum (SEC) that employs a matrix-based design to evaluate the content and cognitive demand of items and content standards. Panelists provide judgmental weightings to estimate how each content area is represented, by cognitive level, on the assessments and also on the content standards using pre-established content taxonomies. Alignment is assessed through a mathematical comparison of the relative emphasis of each content topic by cognitive processes between the content standards and the assessment. The values that result from this process range from 0 to 1 with 1 indicating perfect alignment. Outputs of this methodology include an index of alignment with a range from 0 to 1 and graphical displays of emphasis across content areas.

A third method described by Frisbie (2003) is not as widely used, but focuses on applying alignment methods locally. This method begins with panelists evaluating the cognitive complexity of each content standard and assessment item. The next procedural step is to identify the content match between items and content standards to determine the level of fit (i.e., complete, partial, slight, no fit). Items that are noted as complete or partial fit are counted as a match. Items denoted as slightly fitting the content standards are not. Analyses of the range and depth of coverage occur after the panelists' ratings have been collected.

Finally, a methodology developed by Achieve, Inc. helps determine whether the assessment content is fully subsumed within the state content standards, whether the assessments fairly and effectively sample the content, and whether each assessment is sufficiently challenging (Rothman, Slattery, Vranek, & Resnick, 2002). This analysis of alignment includes four dimensions: content centrality (content match between assessments and content standards), performance centrality (cognitive complexity challenge between assessments and content standards), challenge (level and source of difficulty), and balance and range (match of emphasis and coverage of material between assessments and content standards).

With a range of existing alignment methodologies, some perceive there to be a "best" or "correct" alignment method, particularly in light of feedback that states have received through the NCLB Peer Review process (Davis & Buckendahl, 2007). Rather than debate the superiority of one methodology over others, practitioners could potentially support the use of any alignment methodology by evaluating the validity evidence relative to a common set of expectations. Our recommendation is similar to the perspective that some researchers take with regards to standard setting – that there are critical elements of the process and application that can be evaluated against a common set of expectations (Kane, 2001; Hambleton & Pitoniak, 2006). Thus, it is not a methodology that is valid but rather that validity evidence can be obtained from an alignment study and then evaluated relative to the intended inferences about the results and conclusions of the study.

Proposed Framework for Evaluating Alignment Studies

Kane (1994; 2001) proposed a framework to evaluate validity evidence of the processes and results of a standard setting study. Similar to the claim that standard setting decisions should not be arbitrary and capricious, judgments about the connections among content standards, instruction, assessment, and interpretation of assessment results should be systematically judged relative to the intended representation of the content domain. We are proposing an adaptation of Kane's framework for alignment studies that builds on the procedural, internal, and external validity components and extends it to include the component of utility. Table 1 summarizes our proposed alignment study evaluation framework including:

- Dimension: categories within which validity evidence can be organized and evaluated,
- Validity Evidence: examples of information about the people involved, the process used, and results of an alignment study that can be measured and documented,
- Evaluation Questions: questions that practitioners can use to evaluate the documented validity evidence, and
- Threats to validity: findings or conclusions from an evaluation of validity evidence that impact practitioners' interpretation of the results.

In this article, we expand on the framework outlined in this table by discussing the important elements within an alignment study and how practitioners can judge the results and conclusions using expectations on alignment practices grounded in the professional literature. Within each section, readers are encouraged to refer to the relevant information outlined in Table 1.

Procedural Evidence

Procedural evidence generally refers to the appropriateness of the study design, qualifications of the panelists, appropriateness of the methodology, and the process by which the method was applied in a given situation. As noted earlier, each method employs different procedures, yet includes multiple components and requirements, several of which are similar across methods. As with any evidence collection process, the sources noted here should be viewed as illustrative, but not exhaustive with respect to the types of information that practitioners may evaluate. However, they represent the key sources identified in the professional literature.

Panelists. Each method requires participation of subject matter experts (SMEs) qualified to review content-related material and make systematic judgments about content relevance, cognitive complexity, and difficulty of assessment items or tasks for a particular population. Therefore, the panelists identified to participate in an alignment study need to be familiar with a particular content area (e.g., teacher, curriculum specialist) as well as the knowledge and abilities of the population for which the assessment is intended. Part of a panelist's ability to evaluate what an item is measuring or how it relates to a set of content standards comes from their expertise and experience working with this population and understanding what knowledge and skills are required to answer a question correctly. Although a direct role within a state's educational system is not required, familiarity with the state content standards and related curricula across the panel will likely enhance the agreement among panelists because they will have an understanding of how these elements have been interpreted in practice. However, those leading the study should ensure that panelists are evaluating the "true" match of items to the content standards (by content and cognitive complexity) rather than trying to determine the test developer's intended match.

For greater credibility of the results, the panelists should also be independent of the development process as those who were involved will know the intended alignment and may not be able to make an objective judgment of whether content match exists (Webb, 1999). Bhola et al. (2003) also illustrated the importance of independence of alignment judgments from test publishers' judgments that have a vested interest in the results.

Literature on alignment methodologies and the application of alignment methodologies reports panel sizes as small as 2 (Porter, 2002) and Webb (2007) reports typical panel sizes of 5 to 8 panelists. In selecting an alignment panel there is no particular size that is best or most appropriate. Rather, when designing an alignment study one should focus on representation of different types of content experts. The panel, as a whole, should have background knowledge of all content areas covered by the assessment, the examinees that are included in the intended population, the curriculum that is the basis for examinees learning the content assessed (e.g., curriculum for a school, district, state), and if possible, the content standards to which the assessment is going to be aligned (La Marca, Redfield, Winter, Bailey, & Despriet, 2000).

When applying these expectations to their evaluation, practitioners should view with skepticism results and conclusions of alignment studies that are conducted without qualified panelists who are representative in size and diversity of stakeholders within the intended population.

Alignment method. Although there are a range of alignment methods available, we have limited our discussion in this article to some of the more prevalent methods. Bhola et al. (2003) described three classifications of alignment methodologies based on the complexity of the methods for the panelists. Methods of low complexity are primarily designed to evaluate how the exam items match to the content standards. Such ratings are likely provided by content experts and may be made on some type of scale assessing fit between item and standard. Methods of moderate complexity add to the content match by including criteria for match by cognitive process and/or level of cognitive process. With these added criteria it is likely that there will be fewer matches. Methods of high complexity consider additional criteria to match content standards and assessments such as prioritization given to different content sections (e.g., content weighting, priority standards), or conditions of administration (e.g., use of calculators).

Under NCLB requirements, an alignment method at minimum should include some type of content match as well as an evaluation of cognitive complexity. The approach for doing so will vary among the methods available and selection of a method should consider what type of information is available to conduct the alignment studies and what level/type of information is needed as a result of the alignment study. For example, some programs wish to gather alignment information on a new set of items that do not fully represent a complete test form. Therefore, methodologies that focus on gathering item-level information would be the most appropriate because such information could be analyzed with parallel information on existing items to create test forms that appropriately match the intended balance of content.

For a content match, there are two general approaches in the available literature. The first is a direct match where part of the alignment study is identifying how specific items

Table 1. Overview of Framework for Evaluating Alignment Studies

Dimension	Validity Evidence	Evaluation Questions	Threats to Validity
Procedural	• Panelists: qualifications, independence, representation, familiarity with content domain • Method: approach to content match, consideration of cognitive complexity, consideration of performance levels • Process: training of panelists, review of content domain, practice with alignment process, panelists' evaluation of process	1) Do the qualifications of the panel support a conclusion that they have requisite knowledge of the content and target examinee population to serve on an alignment panel? 2) Do the qualifications of the panel support a conclusion that they are independent of the exam development process? 3) At a minimum, did the alignment methodology consider dimensions of content and cognitive complexity match? 4) Were the training procedures consistent with expectations of the specific methodology? 5) Were training opportunities to practice with the specific methodology provided to panelists? 6) Were panelists given an opportunity to provide feedback on their experience with the methodology and the quality of the results (e.g., questionnaire, focus group)? 7) Is there sufficient evidence to mitigate any concerns of social desirability in the panelists' judgments?	• Panel members were not sufficiently familiar with content at the targeted level. • Panel members were unfamiliar with the knowledge and skills of the examinee population. • Panel members were part of the item writing team raising concerns about conflict of interest. • Alignment methodology did not consider at least content and cognitive complexity as judged dimensions. • Results of post study questionnaire suggest lack of confidence in alignment results (e.g., insufficient training, lack of confidence in judgments). • Results suggest social desirability as a factor in the panelists' judgments.
Internal	• Analysis of panelists' independent ratings • Estimates of agreement among panelists/reliability • Determination of final alignment results • Panelists' evaluation of the final results	1) Does the assessment of agreement among panel members support a general level of consistency within the panel? 2) Was agreement considered when computing the final alignment results? (e.g., independent ratings compared against an agreement criterion, group decision process)	• Agreement was not considered or evaluated in the study design or analysis of results. • Evidence of substantial panel disagreement exists.
External	• Results from separate panels • Results from application of multiple methods • Results from other applications of content analysis • Intended alignment of items to test content	1) If available, did a comparison of the results from parallel panels suggest convergence of conclusions? 2) If available, did a comparison of the results from multiple methods suggest convergence of conclusions? 3) If available, did the results of the alignment study converge with other measures of content relationships within the test? 4) Did the results of the alignment study converge with the blueprint and the intentions of the test developer?	• Results from parallel panels suggest substantial variability and result in different conclusions about alignment. • Results from multiple methods suggest substantial variability and result in different conclusions about alignment. • Comparison of results to intended alignment identified substantial differences and result in different conclusions about alignment.
Utility	• Interpretation and use of test-level results • Interpretation and use of standard-level results • Interpretation and use of item-level results	1) Were the results of the alignment study appropriately communicated to stakeholder groups? 2) Was sufficiency of measurement opportunities addressed in the results and conclusions? 3) Were items identified by the panel as not aligning identified in the report? 4) Was misalignment of items by cognitive complexity identified and reported? 5) If content is sampled across forms or years, was test form alignment evaluated and reported?	• Interpretation accompanying results misuses or over-interprets results. • Interpretation of results does not address the items found to "not align" to the content standards. • Interpretation of results does not address lack of sufficiency of measure for some reporting areas. • A substantial number of items do not align to the level of cognitive complexity specified in the content standards.

align with each content standard. This process is observed in Frisbie's (2003) method. An advantage of this approach is the richness of the information. The results of such a study include the panel's decision of which items align with a particular content standard and vice versa. The disadvantage of this approach is the time it takes it takes for the panelists to come to agreement or the decision rules needed to compute the alignment of such data. A second disadvantage of any item judgment method is that panelists are unable to draw holistic conclusions about the alignment results.

Another strategy is to evaluate content match by estimating the proportion of content (i.e., items) matched to each content standard. This process is observed in the SEC methodology. The advantage of the second approach is the simplicity in the data analysis as only totals and averages must be computed. A potential disadvantage is that agreement among panelists regarding representation of content at the test level may mask substantial differences at the item level. Although some variation within alignment ratings may not represent major disagreements (e.g., ratings typically center on one content strand, but differ in terms of specific standard), there is a potential for significantly different ratings that could indicate a problem with the clarity of the item content or with panelists' interpretation of the alignment rating criteria. For example, if a state is using an assessment to provide feedback to students and teachers at the content strand level (e.g., domain score) and the alignment methodology only asks panelists for alignment information at the test level, major differences in opinion on how the assessment covers each reporting level (i.e., strand) could exist but go undetected. Disagreements at the standard level will be more tolerable as these do not represent specific reporting elements.

This disadvantage can be compounded by a procedural characteristic of alignment methods that allow for panelists to align items to multiple content standards. Webb (2007) suggests that mapping items to multiple content standards should be done sparingly. This recommendation stems from the common practice of writing selected response items to target one specific content standard. However, it is important to note that constructed-response, polytomously scored items are often developed with the intention of assessing multiple content standards and levels of cognitive complexity. In such cases, practitioners must carefully consider how they instruct panelists to deal with items that seem to align to multiple content standards (e.g., permit judgments to designate multiple content standards, limit these judgments to constructed-response items).

Process. The process for conducting an alignment study often begins with an orientation or training activity facilitated by someone with experience in alignment methodologies. This initial training should include a discussion of the purpose of the alignment study, including why the process is important to test development, and the validity of the use of test scores. Webb (1999) notes the importance of providing training before panelists are allowed to provide operational alignment judgments. Although the rationale for including multiple panelists is to obtain several perspectives from stakeholders, it is important that the panel as a whole understands the alignment process, cognitive complexity framework, and how they are to make their individual judgments.

In addition, the panel should have the opportunity to discuss the content standards from which the assessment is derived. This discussion can consist of a simple review or a more thorough analysis. Webb (1999) notes that by having raters initially start with a review of the content standards and force them to assign and then come to consensus on, the depth of knowledge (e.g., cognitive demand, cognitive complexity) of each content standard forces raters to discuss and better understand each standard. As a final part of the training, panelists should be oriented to the alignment methodology and explained how the process will work.

According to Webb (1999), training is enhanced through a practice activity whereby raters provide alignment judgments on several of the initial items and then discuss them and why they assigned specific ratings (content match and cognitive complexity). This could be repeated until appropriate levels of consensus are reached. After the training, panelists typically work independently. However, agreement among the panelists could be enhanced with through checks of consistency that are included in other points of the process. Information about the training and practice processes can be evaluated through observation or a review of the technical report that documents the activities.

A primary question in designing an alignment study is the degree to which one wishes to align the assessments to the content framework. Obviously, the level of specificity of the content framework to which panelists align the test items will affect the level of agreement that is observed. The broader the content framework, the higher the likely levels of agreement among panelists because there are fewer opportunities for disagreement. Conceptually, the level of agreement (e.g., strand versus item-level) is analogous to decision consistency where variation of the individual score may not be problematic unless there is a potential impact on the decisions based on the scores. Practitioners should rely on the intended use of the results from the alignment study to guide their evaluation. For example, if the item level results are also intended to inform test development or revision, the expectations for item level agreement increase to avoid providing misleading information.

This question of agreement leads to the next factor that practitioners should consider in their evaluation of an alignment process. Some processes determine the final operational alignment ratings through data analysis (e.g., SEC, Webb) whereas other methods include a second round that is structured as a consensus process during which the panel shares their independent ratings and discusses each item until consensus is reached (Chin, Rodeck, Buckendahl, & Foley, 2008). The issue of evaluating rater agreement is discussed more thoroughly in the internal validity section.

Popham (personal communication, January 2010) also suggested an additional methodological element for alignment studies that leverages a validation strategy observed in other facets of testing programs (e.g., scoring constructed response items). Specifically, he suggested embedding some items into the panelists' task that were known samples that could be used to evaluate potential social desirability (e.g., stringency, leniency) among panelists. In such an approach, some number of items (e.g., 5) could be included on a form of the test, some of which were intended to be matches and some that were not. The logic of this process suggests that the extent to which the panelists' judgments were consistent with the intended ratings, the more credible the results of the panelists' judgments on the assessment items. Conceptually, including items with "known" alignment properties is analogous to including work samples with "known" scores in a set of constructed responses being evaluated by human raters (Cohen & Wollack, 2006; NCES, 2011). The purpose of using items (or papers) with known properties (or scores) is to evaluate the calibration of the judges by comparing their judgments to the known properties. Directional interpretations of stringency or leniency could then be interpreted from systematic disagreements in the judgments. Although practitioners may observe some of these approaches in practice, such strategies may not be applicable for all alignment methods because of how alignment is calculated (e.g., SEC).

An important final step in data collection for an alignment study that practitioners would want to review in their broader evaluation is a process evaluation of panelists' experience to inform researchers' interpretation of the results and conclusions about alignment. Similar to other validation studies like standard setting (Kane, 1994; 2001) panelist and if applicable, observer evaluations of this process should encourage panelists to communicate their experiences about the process and the results. Evaluation questions might include:

- How confident were you with your judgments about the cognitive complexity of the items?
- How confident were you with your judgments about the content match of items to the content standards?
- How did you feel about the time allocated to making your judgments about the alignment of the items to the content standards?
- How did you feel about the consensus discussions that occurred following your initial, independent item-level judgments?

This information serves as a feedback loop for improving future studies and, more importantly, serve as validity evidence for the results of the alignment process.

In summary, the procedural sources of evidence suggest expectations that should be met when designing and conducting an alignment study. Panelists should be experts in the content area, knowledgeable of the target examinee population, representative of the eligible population, and independent of test development activities. The alignment method should include specific considerations for cognitive complexity and content when evaluating the match between assessment items and content standards. Finally, the process should include a training that covers the methodology to be used along with an opportunity for panelists to practice the rating strategy and conclude with an evaluation activity where panelists can provide their thoughts on the process. Table 1 provides some specific questions that can be used to collect evidence that practitioners can then use to evaluate the strength of different sources of procedural validity evidence identified in this section. Any potential threats to validity that are identified from this consideration of the strength or weakness of evidence should then mitigate the interpretation of the results and conclusions of the alignment study.

We next discuss sources of internal evidence that practitioners should consider in their evaluation of the results and conclusions of alignment studies.

Internal Evidence

We define internal evidence as the consistency of judgments that produce the results from the alignment study—similar to Kane (1994; 2001). Although there are numerous factors that can lead to some differences in the independent ratings, one would expect a substantial level of agreement among panelists' judgments (e.g., cognitive level, aligned content standard) to suggest there is a common perspective on the representation of content between the test and a state's content standards. Internal evidence is typically provided as some indicator of agreement among the panelists. However, to obtain this support for the results requires consideration in the design and execution of the study, followed by an analysis of the results, and finally an interpretation of the agreement in light of the study design.

In alignment studies, there are several procedural elements that can influence the observed level of agreement and will impact conclusions about the results. Thus, these characteristics should be considered in the design of the study. First, as mentioned in the procedural section, the grain size of the content level judgment to which items are aligned and the way in which alignment is determined will certainly influence the agreement in panelists' ratings (Bhola et al., 2003; Koretz & Hamilton, 2006). For example, if panelists are asked to evaluate alignment of items to a framework of 15 content standards, they are likely to have a higher level of agreement in their ratings than if they are identifying matches among 30 content standards. In the latter scenario, the probability of chance agreement is lower than in the former.

Second, the clarity of the content framework and the content standards is important to help differentiate content standards from one another. Vague or overlapping content among standards will lead to lower levels of agreement and threaten inferences about the representation of the content. If there is a problem with a lack of specificity in the content standards, alignment results will be confounded. That is, it will be impossible to determine if results indicating a lack of alignment are caused by a genuine mismatch between the items and content standards or by unclear/overlapping content standards.

Third, one must consider the panelists and, more specifically, diversity among panelists in education, experience, familiarity with assessments, assessment items, and content standards and how this range of experience and expertise will influence the results and conclusions about alignment. It is important that the panel includes a diverse group that is representative of the users of these assessments and content standards; however, panelists' experience with the curriculum, assessments, and content standards will influence how they interpret content and identify links. Webb (2007) reports that a greater number of panelists lead to increased reliability. This assertion is based on a computational estimate of reliability where greater variance brought about by more ratings leads to higher estimates of reliability. However, this does not mean there will be an overall greater level of agreement among the panel if it is larger. When selecting panelists, one would want the panel to be large enough to represent sufficient diversity among relevant stakeholder groups and have a clear majority in the case of some divergent panelists. However, the panel should not be too large that it cannot be managed by a panel leader or facilitator. These multiple considerations lead to the range in the recommended panel sizes noted above.

Fourth, we must also think about agreement in decisions about cognitive complexity. Understanding and evaluating cognitive processes can be very difficult. In some alignment studies, the panelists are well-trained educators who have had experience with the concept of cognitive complexity (e.g., Bloom's taxonomy, Webb's Depth of Knowledge) and find the task comprehendible, but in others where panelists may not be familiar with such terms, agreement may be quite low. Similar to aligning test items to content standards, higher agreement would be expected when there are fewer and more well-defined levels of cognitive complexity.

The design of the alignment study also has a direct influence on how agreement can be calculated. Some alignment methodologies suggest collecting independent panelist ratings which are combined after the study, possibly using a decision rule, to determine the final panel judgment (e.g., 50% of the panel must agree at the predetermined level of judgment, 67% of the panel must agree, 75% of the panel must agree). Other applications require the panel to come to agreement on the alignment judgments after having the opportunity to make initial independent judgments (e.g., Badgett, Davis, & Buckendahl, 2008). Each approach has specific advantages that practitioners evaluating alignment studies can consider.

There are two major advantages to using an independent approach where group consensus is not required. First, the makeup of the panel and the social dynamic of its members can influence the outcome. Specifically, different personalities or dynamics within the panel may influence the outcome and one panelist's opinion may be heard more than others. This may legitimately result in some panelists' dissenting judgments being overlooked or not heard by the group as a whole if the dynamic is strong enough. Using an independent approach allows each panelist's judgments to be counted equally. Second, from a practical perspective, the independent approach may take less time than the consensus process depending on the level disagreement and resulting discussion within the panel. Chin et al. (2008) suggested that post-study alignment analysis of independent ratings when applying a decision rule that specified a minimum percentage agreement may result in similar alignment conclusions to those resulting from panel consensus discussions. If so, the additional time required for consensus discussions may not add significant value to the process.

When using an independent approach, agreement would be based on the panelists' independent ratings and the method of computation will depend on the alignment method used. For example, Porter et al. (2008) computed reliability by comparing the percent of content aligned to each cell within the content matrix by a team of raters using Generlizability theory (e.g., Shavelson & Webb, 1991). Other alignment methodologies may lend themselves to computing reliability as agreement among item-level judgments (e.g., cognitive complexity rating, specific standard-level content match). In these instances, agreement can be estimated through percent agreement or intraclass correlation of panelists' independent ratings. In particular methods where an independent approach can be used, agreement also directly influences the results. For example, if the agreement rule is not met (e.g., 55% is observed when 67% is required for alignment) the results would indicate that a particular item is not aligned to the content. Therefore, such decision rules can have a substantial impact on the final alignment decisions (Webb et al., 2007).

A consensus process also provides several advantages. First, the results of the study (alignment decisions) provide the test developer with a specific mapping of all items to content standards rather than just assurances that the items are aligned in a general sense. This is an important source of information that will be beneficial when evaluating external validity evidence of the results of the alignment process. In addition, this information can be used when items on the test form need to be replaced, retired, or removed from use. The test developer could utilize this mapping information to find alternative items that are aligned to the same content standards or to estimate sufficiency depending on the purpose of the test (Norman & Buckendahl, 2008).

Table 2. Sample Panelist Judgments Using Number of Items

Standard	Panelist	Items Aligned	Total Number of Items
1	A	1, 3, 15, 22, 35, 40	6
1	B	5, 7, 17, 21, 35, 37	6
2	A	7, 12, 26, 39, 43	5
2	B	2, 12, 23, 39, 43	5

The second advantage to requiring consensus for a panel for final results is verification that the panelists are in agreement on the specific content standard to which each item is aligned. Judging agreement by number of items aligned to a specific content standard may hide substantial disagreement among panelists regarding content matches. For example, Table 2 provides a sample of results from an alignment study where panelists independently evaluated the content alignment of items to standards. The information provided includes the specific items aligned by two panelists to two different standards. Under a numerical agreement model (e.g., agreement is evaluated by comparing number of items aligned to each standard), the two panelists are in "agreement" on the number of items identified as aligning to each standard but have important differences in their ratings. For example, in the first standard, there is only 17% agreement in specific items and in the second standard there is 60% agreement in the specific items aligned. Under an alignment model where agreement is determined by number of items aligned, these two panelists would be in "agreement" on the alignment of these two content standards. Therefore, analyzing agreement at a summary level (e.g., test or content area) could mask substantial disagreement at the item level. Evidence of item-level disagreement can provide valuable information to practitioners and feedback to test developers for improvement of item writing on defining content standards.

A third advantage of using a consensus approach is that professional consensus-building conversations may increase the confidence among panelists and researchers with respect to the panelists' alignment decisions. This follows from standard setting where researchers have advocated that multiple rounds be conducted so that each panelist can make independent judgments and in later rounds learn how their opinions compare to their fellow panelists (e.g., Hambleton, 2001; Kane, 1994).

When panelists are required to arrive at a group decision, agreement is estimated and interpreted differently. For example, Chin et al. (2008) evaluated agreement among panelists' ratings that served as the foundation for the group discussion at different stages in the iterative judgmental process. Not surprisingly, panelists' individual judgments became more consistent after multiple opportunities of consensus judgment discussions. Such empirical evidence could be supplemented through use of written process evaluation questions on this topic. Although agreement among the panelists is the primary source of internal evidence, a lack of confidence among panelists regarding the consensus results may serve to discount those findings.

In summary, to meet this evaluative criterion, an alignment study should provide some estimation of agreement among panelists along with an appropriate interpretation of this estimate. Again, the reader is referred to Table 1 for questions, the types of evidence, and threats to validity that can aid in evaluating internal evidence. Any substantive level of disagreement among panelists should be noted and prompt further investigation to determine if there are specific factors that contributed to the disagreement. For example, Herman et al. (2007) considered rater agreement within three facets using Webb's method: number of items assigned to each standard, percent of topics within one standard that had at least one item assigned, and number of items assigned to two content standards versus one. The results indicated the panelists had very different opinions and different collections of panelists produced different results. In conclusion, the authors stated, "Agreement in the assignment of ratings can be considered an indicator of the extent to which common understandings are shared. Similarly, lack of agreement on the relationship between content topics and items suggest that educators operate with diverse definitions of the meaning of standards in terms of content and Depth of Knowledge expectations" (p. 122).

External Evidence

In Kane's (1994; 2001) framework, collecting external validity evidence often extends beyond the boundaries of the current study, yet is sought to evaluate the results through connections to results of similar studies or other types of information. One strategy that Kane (1994; 2001) suggests is the use of multiple independent studies to triangulate the results (e.g., Green, Trimble, & Lewis, 2003) and a similar approach can be applied to alignment studies. This could be done using the results of multiple alignment methods (e.g., Webb's method, SEC) to compare the potential range of alignment conclusions. In addition, one could compare the alignment judgments from the independent panel to those from the test developers (e.g., Buckendahl, Plake, Impara, & Irwin, 2000) or from two independent panels to evaluate agreement.

In standard setting, Kane (2001) reminds us that no single comparison is definitive but results that are consistent across multiple sources provide important validity evidence for the use of the study findings. In using multiple methods, the question becomes—what do non-convergent results mean? Who is correct? Ultimately, as with any judgmental process, there is no "gold standard" for alignment judgments. Because alignment data are collected to evaluate the intended use and interpretation of test scores, evidence similar to the particular use of the test might be given more weight in determining the final alignment. For example, alignment results from a panel working in the particular educational system where the content standards dictate curricular content and instructional practice or where the test is administered might have a better perspective than a group of similarly trained educators working in a different educational setting. Similarly, when comparing multiple methods one should consider the level of agreement required by the method to count an item/content standard match as "alignment."

External validity evidence requires additional effort to collect and evaluate. More often than not, programs do not have the time or resources to collect this source. However, practitioners are still encouraged to consider the importance and value of this type of evidence when designing alignment studies. In addition, when this type of information is available, practitioners have the responsibility to evaluate this information and interpret differences (see Table 1 for additional guidance on the types of questions, related sources of evidence, and specific threats to validity).

Utility Evidence

As an extension of Kane's (1994; 2001) evaluation framework, we propose the addition of a fourth source from which one can identify and evaluate validity evidence, specifically, *Utility*. This dimension pertains to the interpretation and use of the alignment study results to inform test development practice and policy. The complexity of alignment studies discussed previously (e.g., unit of alignment, panelist agreement, etc.) should all be incorporated in the interpretation of the results. In addition, with respect to interpretation of alignment study results, several questions arise related to the intended use of the test scores (e.g., Webb, 2007). Because we acknowledge that utility will have different meaning for different programs, in this section we start the discussion of this source of evidence by identifying some of these questions and suggest strategies for addressing them.

Test level results. What does it mean to say an assessment is "aligned"? Koretz and Hamilton (2006) noted, "There are currently no criteria for determining what level of alignment is acceptable. Perfect correspondence between tests and standards is infeasible because tests typically require sampling from the domain of interest" (p. 556). Should this imply that a certain proportion of the test is aligned to the content standards? Should it imply that all content standards have sufficient coverage in the test form? Should it imply some level of agreement on the content and cognitive complexity ratings among the independent panel?

As we have described, the various alignment methods may produce different results. The SEC method produces an index of alignment that ranges from .0 to 1.0 with values closer to 1.0 indicating greater alignment whereas other methods (e.g., Frisbie, 2003; Webb, 1997) provide summary information on the alignment between content standards and items with conclusions about the results of different alignment dimensions. How does one interpret each type of information? Porter (2002) notes "there is still no easy way to think about how big the alignment index value must be to be considered 'good'. The index does not have a straightforward interpretation like the proportion of common content between say, standards and assessment" (p. 6).

At a minimum, an alignment process should result in some overall estimation of the alignment of an operational test form to the content standards. Whether this is an index, a percentage, or summary statistic will depend on the method. The important element is that a test user can interpret this statistic based on his or her expectation. At minimum, this summary reporting should include elements of both content and cognitive demand because an item cannot be aligned on only one dimension.

Item level results. How does a test user deal with items judged as "not aligned" to any content standards? Should these items be revised, deleted from the test, un-scored, or ignored? Some alignment methodologies specifically identify such items as not aligned whereas others only provide a general indication of alignment for the entire exam and thus such items are not identified as needing action. Test sponsors and developers should have a policy and plan for dealing with any items that are identified as not aligned to the content standards. If these items only represent a small proportion of the assessment, the plan may be to flag these items in the bank for revision before they are used operationally on any other test forms. If a significant portion of items on an assessment are identified as not aligning to the test content, the test sponsor or developer should consider replacing these items with others to achieve a sufficient level of alignment for the operational test. For custom tests, it is reasonable to expect that 100% of items are subsumed within a defined content domain.

Similarly, how does one interpret items that are deemed to be aligned to more than one content standard? Koretz and Hamilton (2006) note that most methods do not provide guidance for interpretation of items aligned to multiple content standards. In the SEC method, the item points are divided into the content areas to which they are aligned. In some cases, the design of an assessment will include items mapped to multiple content standards. Review of the observed alignment results to the intended alignment as external evidence would identify any of these occurrences that match the intention of the assessment design. Any remaining items mapped to multiple content standards should be reviewed at the test-level. In some cases, such items are aligned to the same two content standards, which indicate a lack of clarity or distinction between the two content standards.

It is important for alignment researchers to develop a decision rule for how to interpret these items when analyzing results and forming conclusions about overall alignment. One strategy would be to identify an item as two measurement opportunities if it is aligned to multiple content standards. Practitioners can ask the panelists to consider the inference(s) that can be made if a student answers the item correctly. Specifically, would a correct response indicate that the student has demonstrated the primary skill indicated by standard A and the primary skill indicated by standard B? With this guidance included as part of the alignment training procedures and reiterated during operational ratings, it would be appropriate to identify these items as providing two measurement opportunities. However, panelists in other alignment applications may not be given such specific guidance and test developers would need to provide a test user with some guidance on how to interpret these results.

Standard level results. How does one deal with content standards to which there are no or only a few items aligned? Should content standards not meeting some type of preset criteria be excluded from the assessment? Excluded from reporting? Are methods that sample different content across years acceptable if the alignment evidence is still persuasive? To what extent does the number of reported achievement levels influence the minimum number of items needed to evaluate alignment?

Frisbie (2003) suggests that each standard have a minimum of five items to be sufficiently covered, whereas Webb (2007) recommends six items. The minimum level of coverage will depend on how the results of the assessment are reported and interpreted. If a state chooses to report and evaluate strand-level or standard-level results, a more stringent criterion should be in place for a minimum number of aligned items. If a state limits reporting to the test level, less stringent criteria may be reasonable to support the larger sampling frame. However, in this latter case, it is important that the test assesses the breadth of content that is intended. Substantial gaps in the content coverage should require revision of the operational test forms.

Dissemination. Beyond meeting any requirements for alignment (e.g., NCLB requirement), how are the results of the alignment study used by test developers and test users? Because validity is not an all or nothing concept, but rather an evolving property of the use and interpretation of test scores, alignment evidence should not be reported and then put on the shelf. Rather, the information should be used as part of the continuous feedback for improvement of the test and for enriching the information available in the item bank.

Each method provides different types of information but the utility of the results stems from the richness of the information and what test users can take from the information. As an additional step, a test developer may choose to share the information with educators, administrators, or policymakers. Although an ideal practice in the spirit of full disclosure, providing detailed information should be done with caution. Typically, alignment results may only be provided for one operational test form of several that are used in a given administration year. As noted above, an assessment is a sampling of the content and it would be inappropriate to provide alignment results as the only sampling possible for a given assessment.

In summary, utility evidence comes from specific features of how the results of alignment studies are reported and used. The results should provide some type of summary information that can aid readers in understand how the overall assessment aligns to the content standards. In addition, specific features that should be addressed include items that are not aligned with the content standards (if item-level data is collected) and content areas which are not sufficiently covered by the assessment. Practitioners should document these results with accompanying interpretation of what these results mean for test score use.

Use of the Evaluative Framework

By design, this framework is intended to provide practitioners with evaluative questions, identify sources of evidence that could be used to respond to these questions and identify related threats to validity for each criterion. In this article, we described the key elements of an alignment study, how practitioners can evaluate these specific elements and why information pertaining to these elements can serve as evidence of the validity of the intended uses of the results. Table 1 served as an organizer for our discussion. Within each of the four dimensions, practitioners can use the validity evidence examples as a way to organize information pertaining to the people involved, the processes used, results and conclusions of an alignment study. This information can then be used to address the related evaluation questions. In answering these questions, practitioners should consider the guidance and recommendations laid out in this article to understand the given strengths and weakness of the alignment study to provide an overall judgment.

Applying this framework to evaluate different types of alignment studies will likely reveal the substantial differences that can exist due several factors including the purpose of the test, the alignment methodology selected, and the individuals involved. Rather than creating a framework to compare one application to another, we encourage practitioners to focus on identifying any threats to validity that might affect how the results of the study should be interpreted or used. These are specifically noted as individual threats to validity for each identified criterion in Table 1).

Summary

With the increased importance of alignment studies as a primary source of validity evidence for educational assessment and accountability programs, there is a need for an evaluation framework that can be systematically applied to alignment research. This is particularly important to avoid policymakers and the field from succumbing to dogma for any given methodology. Because both are important judgmental activities in the validation process, connections can naturally be drawn between alignment and standard setting studies. From standard setting guidance (e.g., Kane, 1994; 2001), we have developed the proposed framework for evaluating specific sources of validity evidence to identify threats to validity that would impact the interpretation of the results.

In designing this proposed evaluation framework for alignment studies, we intended for practitioners to consider each component in the design of the alignment study as well as in the reporting of the results. Similar to standard setting, certain elements of validity evidence are more difficult to obtain and may require more specific planning and additional resources beyond what is available within the structure of many studies. Practitioners should strive to collect diverse sources of evidence that represents the key components of the alignment process. Although we acknowledge that it may be difficult to fully address each of the four areas outlined in this proposed framework, some evidence is needed to inform conclusions drawn from results. This is particularly important for communicating results to the diverse stakeholder groups using the results (e.g., technical advisory committees, content specialists, test developers, policymakers). Conversely, evidence is needed to ensure that there are limited or no threats to validity that would influence how the results of the alignment study are interpreted.

Currently, many alignment studies address some of the elements outlined in this paper but do not fully meet the expectations. With regard to *procedural* elements, the requirements of NCLB as dictated through the peer review process include consideration of many of the procedural elements (e.g., independent and representative panel, consideration of cognitive complexity). Similarly, reporting requirements dictate that practitioners consider agreement either through the design of the alignment study (e.g., consensus) or through the computation of the results (e.g., agreement criterion), which provides some *internal* evidence. However, estimation and interpretation of rater agreement statistics may not often be included as part of the report from an alignment study. *External* evidence of validity is likely rarely obtained as to do so requires additional resources (e.g., additional panel to replicate method, replication with an alternate method). In some cases, a state may have the information available to compare the independent alignment results to a test publisher's intended alignment. Finally, some alignment practitioners are considering the *utility* issue by providing guidance on the interpretation of detailed results. However, more frequently we see only high-level results reported that address the question of test-level alignment requirements.

One goal of this proposed framework is to serve as a starting point for further discussions about how to more systematically evaluate the category of judgmental validation research that includes alignment studies and the sources of evidence needed to inform that evaluation. A framework that can be used to evaluate alignment studies will spur further alignment research rather than limiting conversations.

References

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.

Badgett, B. A., Davis, S. L., & Buckendahl, C. W. (2008). *Alignment in higher education: Matching assessments to program goals of an online university*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

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.

Buckendahl, C. W., Plake, B. S., Impara, J. C., & Irwin, P. I. (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.

Chin, T., Rodeck, E. M., Buckendahl, C. W., & Foley, B. P. (2008, March). *The impact of consensus on alignment judgments*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Cohen, A., & Wollack, J. (2006). Test administration, security, scoring, and reporting. In R. Brennan (Ed.). *Educational measurement* (4th ed., pp. 356–386) Westport CT: American Council on Education and Praeger.

Davis, S. L., & Buckendahl, C. W. (2007, April). *Evaluating NCLB's peer review process: A comparison of state compliance decisions*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

DePascale, C. (2007, June). *Alignment: Their's not to reason why*. Paper presented at the National Conference on Large Scale Assessment, Nashville, TN.

Frisbie, D. A. (2003). *Checking the alignment of an assessment tool and a set of content standards*. Iowa Technical Adequacy Project (ITAP). Iowa City, IA: University of Iowa.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice*, *22*(1), 22–32.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education and Praeger.

Herman, J., & Webb, N. (Eds.). (2007). Alignment methodologies [Special Issue]. *Applied Measurement in Education*, *20*(1), 1–135.

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

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*, 425–461.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.) *Setting* *performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Koretz, D. M., & Hamilton, L.S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education and Praeger.

La Marca, P. M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, *7*(21). Retrieved March 13, 2008 from http://PAREonline.net/getvn.asp?v=7&n=21

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

Martineau, J., Paek, P., Keene, J., & Hirsch, T. (2007). Integrated, comprehensive alignment as a foundation for measuring student progress. *Educational Measurement: Issues and Practice*, 26(1), 28–35.

Martone, A., & Sireci, S. (2009). Evaluation alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*, 1332–1361.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

NCES. (2011). National Assessment of Educational Progress (NAEP): NAEP item scoring process. Retrieved May 23, 2011 from http://nces.ed.gov/nationsreportcard/contracts/item_score.asp #ensuring

Norman, R. L., & Buckendahl, C. W. (2008). Determining sufficient measurement opportunities when using multiple cut scores. *Educational Measurement: Issues and Practice*, *27*(1), 37–46.

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

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

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). *Benchmarking and alignment of standards and testing*. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Shavelson, R., & Webb, N. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(1), 4–14.

U.S. Congress. (2002). *Public Law 107–110: No Child Left Behind Act of 2001*. Accessed May 15, 2010 from http://www.ed.gov/policy/elsec/leg/esea02/index.html.

U.S. Department of Education. (2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001*. Washington, SC: Office of Elementary and Secondary Education.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 18). Madison, WI: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, *20*(1), 7–26.

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.