Evaluating the Fit between Test Content, Instruction, and Curriculum Frameworks: A Review of Methods for Evaluating Test Alignment¹

Shuhong Li and Stephen G. Sireci²

University of Massachusetts Amherst

June 30, 2005

¹Center for Educational Assessment MCAS Validity Report No. 9. (CEA-558). Amherst, MA: University of Massachusetts, Center for Educational Assessment. This work was carried out under a contract between the University of Massachusetts Center for Educational Assessment and Measured Progress, Dover, NH, 2004-2005.
²The authors are grateful to Professor Ronald K. Hambleton for his helpful comments and editing of the final draft of the paper.
Abstract

Content validity has long been recognized as the cornerstone of a sound educational assessment, and the effort invested in achieving and evaluating content validity has never stopped. Such efforts are gaining increasing momentum due to NCLB requirements. Given the importance of ensuring that tests are aligned with curricula and the fact that there is a paucity of literature in this area, a critical analysis of new models of test alignment is needed. In this paper, we tap the available and existing literature on test alignment and evaluate the five most popular methods. We address the following questions: 1) What is test alignment? 2) Why is test alignment important? 3) What are the existing models for aligning tests to curricula? 4) What are the similarities and differences across these models? 5) What are the problems related to these models? 6) What are the implications for validity (content, curricular, consequential) and what recommendations can be made to local and state education agencies, test developers, and psychometricians? The implications of the study should be of interest to test developers, educators in local and state education agencies, and specifically, administrators in the Massachusetts Department of Education for whom this report was written. These audiences desire sound assessments that provide accurate inferences about student achievement and growth.
Alignment analysis of assessments, curricula, and standards in educational measurement is gaining increasing importance and popularity due to both psychometric and socio-political reasons. Alignment analysis is not a new phenomenon; in fact, it can be traced back to the evaluation of content validity. For example, as early as in the 1960s, analyses were performed on assessment tasks and behavioral objectives as part of the mastery-learning movement (Cohen, 1987). Alignment was said to be attained if the assessment tasks were equivalent to the instructional tasks.

Content validity has long been recognized as the cornerstone of a sound educational assessment, and the effort invested in achieving and evaluating content validity has never stopped. However, given the increased assessment demands, for example those associated with the recent No Child Left Behind (NCLB) legislation (NCLB, 2002), it is increasingly important to document the degree to which state-mandated tests are congruent with the goals of education reform movements. Earlier, Webb (1997) developed a landmark model for evaluating the alignment of tests with instruction and curricula. His work has been well received and has influenced several other models put forward by other researchers. However, the degree to which these new models improve upon traditional notions of content and curricular validity has not been studied. In addition, there is little research that critically reviews current alignment models. Therefore, the purpose of this paper is to review the existing and available
research on alignment methods and to relate these methods to traditional content validation practices where possible.

Given that the new models of test alignment only have a history of a few years, there is a paucity of research addressing their utility. In this paper, we focus on the evaluation of the five most popular test alignment methods and address the following questions: 1) What is test alignment? 2) Why is test alignment important? 3) What are the existing models for aligning tests to curricula? 4) What are the similarities and differences across these models? 5) What are the problems related to these models? 6) What are the implications for validity (content, curricular, consequential) and what recommendations can be made to local and state education agencies, test developers, and psychometricians?

To achieve these purposes, sixteen papers that focused on alignment analysis were located and reviewed. Interestingly, only three were published; the others were either conference presentations or technical reports. Eleven are empirical alignment studies. These papers and reports were critically reviewed and contrasted with the existing literature on content validity (e.g., Crocker, Miller, & Franks, 1989; Sireci, 1998a, 1998b). The work carried out previously by the University of Massachusetts for the Massachusetts Department of Education with the grade 10 ELA and math tests represents a mixture of models with the focus for the review on two aspects—the match between the content of items and the test specifications, and the coverage of the learning standards in the curriculum frameworks (Hambleton & Zhao, 2004).
Brief Description of Content Validity

Content validity is defined as the ability of a test to represent adequately the content domain it purports to measure (Sireci & Geisinger, 1992). Sireci (1998b) provides a comprehensive discussion of this concept, including the origin and evolution of content validity, the role of content validity in validity theory, the practice of content validation, and the future of content validation. Readers who want an in-depth understanding of the concept should refer to this article. He acknowledged in the very beginning of the paper that “the concept of content validity has been controversial since its inception” (p. 83). After elaborating on the origin and evolution of the concept, he defended the necessity of retaining the term content validity based on the following three reasons: 1) “it is an accurate descriptor of desirable test qualities,” 2) “the term is useful for describing the family of test construction and validation procedures pertaining to measurement of the underlying domain,” and 3) “it is a term audiences of non-psychometricians can easily understand” (p. 104).

As has been mentioned, procedures used to evaluate content validity fall into two categories: judgmental and statistical. Judgmental methods often involve subject matter experts (SMEs) who rate test items according to their relationship with the content domain that is supposed to be tested by an assessment, and an index is often provided to indicate the degree to which the SMEs believe the test is relevant to and representative of the content domain tested. Hambleton (1980, 1984) proposed an “item-objective congruence index” which reflected SMEs’ ratings along a three-point scale, and he later suggested that the SME ratings be measured on a longer Likert-type scale. Sireci and Geisinger (1992, 1995) used multidimensional scaling (MDS) and cluster analysis to
evaluate SMEs’ ratings of item similarity. Statistical methods concern those procedures in which test scores or item scores are analyzed, for example, the application of MDS, cluster analysis, factor analysis, and generalizability theory in this situation.

1990s and Beyond: Importance of Assessing Alignment of Tests to State Curricula

State-mandated testing increased dramatically in the 1990s, and NCLB explicitly requires that states select and design assessments of their own choosing, but the state assessments must be aligned with state academic standards. According to Webster’s New World College Dictionary, alignment means ‘to bring parts or components into a proper coordination; to bring into agreement or close cooperation.’ Webb (1997) defined alignment in our context as ‘the degree to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do’ (p. 4). His ‘expectations’ can be conveyed in the form of content standards, which stipulate the knowledge and skills students are expected to acquire and which the assessment is usually based on, or some other forms of framework such as recommended instructional practice. Also, as La Marca (2001) pointed out, curriculum has to be part of the alignment analysis of standards and assessments because content standards alone are simply descriptive statements about a given curriculum.

Webb further classified alignment analyses into two categories: horizontal and vertical. Horizontal alignment refers to the alignment between components at the same level of the educational system, for example, state standards and state assessments. Vertical alignment refers to the alignment between components at different levels of the system, for instance, between state assessments and instruction (1997).
Alignment is fundamental to the validity of interpretations made from educational tests because content standards are usually developed before the tests themselves and before student performance standards are set on the tests. Students’ test results are often interpreted with respect to their mastery of standards specified in curriculum frameworks. If there is poor alignment between the test and the frameworks, such interpretations will be invalid.

**Approaches for Alignment Analysis**

Based on the current practice and existing literature, Webb (1999) believed there are three approaches to alignment analysis: sequential development, expert review, and document analysis. He noted they are not the only approaches, and can be used in combination in many situations. Sequential development refers to the situation in which the expectations and assessment are aligned by design because the development of the first document lays a blueprint for the development of the subsequent documents. Expert review is adopted by some states and districts when they invite a panel of experts to review the elements and make a judgment on the alignment. Alignment can also be evaluated by coding and analyzing the documents that convey the expectations and assessments. The Third International Mathematics and Science Study (TIMSS) best exemplifies how this is carried out.
Alignment Models

Alignment analysis actually involves a number of dimensions on which the match between a set of standards and the test are evaluated. Our review found five such alignment models that are currently used by states: 1) the Webb (1997) method; 2) the La Marca method (2001); 3) the Surveys of Enacted Curriculum method (Council of Chief State School Officers, 2002); 4) the Achieve method (Achieve, 2001); and 5) the Council for Basic Education (CBE) method (CCSSO, 2002).

Webb’s model. Norman Webb of the Wisconsin Center for Education Research developed his alignment model in cooperation with Council of Chief State School Officers (CCSSO) and states. This model has been used in alignment studies for language arts, mathematics, science, and social studies carried out in more than 10 states (CCSSO, 2002). Webb’s model has five criteria: 1) Content Focus, 2) Articulation across Grades and Ages, 3) Equity and Fairness, 4) Pedagogical Implications, and 5) System Applicability.

Content focus has four sub-categories: Content Depth of Knowledge, Categorical Concurrence, Range of Knowledge, and Balance of Representation. Content Depth of Knowledge refers to the agreement between the cognitive demands of the standards and those of the assessment. Webb suggested four levels of cognitive complexity (recall, skill/concept, strategic thinking, and extended thinking) that can be modified to adapt to different situations. Appropriately chosen and trained experts rated both the cognitive demand of the test items and that of the standards. There is not a cut-and-dried standard to indicate this consistency, but Webb suggested a common sense one – to be reasonably sure students were assessed at the appropriate cognitive level, at least 50% of the test
items needed to be at the same, or higher, cognitive levels as those described in the standard.

Categorical Concurrence requires that the same or consistent categories of content appear in both expectations and assessments. To be exact, he suggested that to reach Categorical Concurrence at least six items correspond to the standard. The rationale for using the number six as the criterion was based on Subkoviak (1998) and some assumptions such as: there are only two levels of mastery, the passing score for mastery determination is four out of six, and the six items when considered together produce a score having adequate reliability. However, he agreed this number could vary according to different situations.

The ideal requirement for the Range of Knowledge is that the full range of expected knowledge and skills in standards be measured by the assessment. Because of practical constraints like time limitations, costs, students’ fatigue etc., this criterion suggests that the assessment must reflect a reasonable amount of the content described in the standards. His criterion for this category requires that at least 50% of the objectives in the standards have at least one matching item.

Balance of Representation means similar importance should be given to different content topics, or items should be evenly distributed across objectives. Based on the assumption that standards can have more than one level of specificity and that the most specific statements are comprehensive and equally weighted, he produced an index to measure the extent to which this criterion was met. In calculating the index, one objective in the standard that has at least one related assessment item is called a ‘hit.’ The index is derived by considering the difference in the proportion of objectives and the proportion
of ‘hits’ assigned to the objective. If the value of the index is close to one, most of the objectives are measured by the same number of items, and the assessment is balanced. If the index is close to zero, then either few objectives are being measured, or the distribution of items across objectives is concentrated on only one or two objectives. He suggests that an index should be greater than 0.7 in order for an assessment to be considered aligned to the standards it is based on in terms of balance of representation.

By Articulation across Grades and Ages, Webb means that students’ knowledge of subject matter grows over time, so both assessment and standards must be rooted in a common view of how students develop, and how best to help them learn at different developmental stages. This common view is based on cognitive soundness and cumulative growth in content knowledge.

The Equity and Fairness criterion is designed to make sure that the aligned assessment and its corresponding standards give every student a reasonable opportunity to demonstrate the expected level of learning for each of them. As Webb (1997) said, “an aligned assessment will provide opportunities for students who are unsuccessful on one form to demonstrate their understanding by taking an alternative form or taking the same form under different conditions as long as validity and reliability are maintained” (p. 27). He also pointed out that for the assessment to be fair and equitable, time is required for patterns to form in order to decipher how system components are working in concert with each other.

Pedagogical Implications include the criteria of engagement of students; effective classroom practices; and the use of technology, materials, and tools. In an example, Webb said that if standards indicate that students should learn to use calculators
routinely, then the curriculum should provide adequate opportunity for students to use them in this manner, and an aligned assessment should allow students to use calculators effectively to derive correct answers.

System Applicability demands that though system components should seek to encourage high expectations for student performance, they also need to form the basis for a program that is realistic and manageable in the real world.

Bhola et al. (2003) noted that Webb’s conceptualization of alignment is more comprehensive and extends far beyond that of other models, and therefore it can be adapted for use in many more contexts. It should be noted that Webb (1999) provided illustrations of aligning four states’ assessment systems with their content standards by adopting only his Content Focus criterion.

La Marca et al. (2000) also described a model for aligning assessment systems to state content standards that are built specifically around the requirements of Title I (United States Department of Education, 1999). Their model also has five interrelated dimensions: Content Match, Depth Match, Emphasis, Performance Match, and Accessibility. The first two correspond respectively to the Content Depth of Knowledge and Categorical Concurrence subcategories of Webb’s Content Focus criterion. Emphasis refers to the weight given to what is tested and how well this matches the weight assigned in the standards. This differs from Webb’s Balance of Representation subcategory in the Content Focus criterion where the same weight is given to different standards. This difference is caused by the fact that in some states, each standard possesses the same weight, while in other states standards are weighted differently. Performance Match deals with the degree to which items on the assessment
allow students to demonstrate the kinds of knowledge and skills intended by the standards, or assessment items should be congruent with the types of student performance described in the standards as well as the methods by which the content of these standards was delivered in the classroom. Accessibility means the assessment should include items that will allow students at all levels of proficiency to have an opportunity to demonstrate their level of knowledge.

**Achieve model.** Beginning in 1998, Achieve Inc. (2001) has been using a protocol of six criteria for determining the alignment of assessments to standards: Accuracy of the Test Blueprint, Content Centrality, Performance Centrality, Challenge, Balance, and Range. This model has been used by a number of states.

Accuracy of the Test Blueprint makes sure that each test item corresponds to at least one standard, which is important because a test blueprint usually forms the basis for score reports. Content Centrality examines the degree of alignment between the content of the test items and the content of the related standard. Performance Centrality refers to the cognitive demand of the task. These two correspond respectively to the Content Depth of Knowledge and Categorical Concurrence subcategories of Webb’s Content Focus criterion. The Challenge criterion checks two factors to make sure the items are fair: source of challenge and level of challenge. Regarding source of challenge, reviewers check to see if the challenge in an item comes from content in the standards that students must know in order to answer the question correctly. With respect to level of challenge, reviewers determine whether the sets of assessment items span an appropriate range of difficulty for students in the target grade level. Balance requires that the same weight be given to the skills or knowledge in the assessment as is given in the standards.
Achieve’s balance criterion is different from Webb’s Balance of Representation subcategory in the Content Focus criterion in which the same weight is given to different standards, while with the Achieve model, most items will be allocated to assessing the more important content and performances described by the standards. It is often hard to assess the whole standard with a single assessment, so Range evaluates the degree to which the assessment samples the standard.

**SEC model.** The Surveys of Enacted Curriculum (SEC) alignment methodology was developed by Andrew Porter and John Smithson from the Wisconsin Center for Education Research. Development and application of the model was supported by CCSSO through grants from the National Science Foundation and through a state collaborative project. The SEC alignment methodology has been field tested and demonstrated with 11 states and four large urban districts. The SEC model produces 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 content matrix has two dimensions for categorizing subject content: Content Topics and Cognitive Demand (or expectations for student performance). The dimension of cognitive demand has five categories for mathematics: (a) memorize, (b) perform procedures, (c) communicate understanding, (d) generalize/prove, and (e) solve non-routine problems. It employs a matrix of two dimensions: content topic and cognitive complexity match (CCSSO, 2002).

**CBE model.** This is a model used by the Council of Basic Education, and has been executed in several projects, including a standards and assessment alignment study for a local school system, several states, and the standards alignment described in CBE’s
"Great Expectations" report on the analysis of content and rigor in English language arts and mathematics in all 50 states (CCSSO, 2002).

This model has four criteria: (1) Content, (2) Content Balance, (3) Rigor, and (4) Item Response Type. Content has a similar meaning to that used in the previously described models, and Rigor refers to higher order thinking skills. These two coincide with Content Depth of Knowledge and Categorical Concurrence subcategories of Webb’s Content Focus criterion respectively. Content Balance looks at the number and distribution of items used to assess the various standards, and it roughly coincides with Balance of Representation in Webb’s Content Focus criterion. Item Response Type examines if the item requires a select or supply response type.

**Similarities and Differences Among the Models**

Since the models described above have similar goals, there are several similarities among them, but there are also important differences. The biggest difference is the number of criteria each uses to evaluate alignment. The users of the models need to notice that even though some of them might have the same number of criteria, these criteria might not mean the same thing. As far as commonalities are concerned, generally, they all combine qualitative expert judgments and quantified coding of standards and assessments, with their focus on judgmental calls. That is, none of them can be represented by a mathematical formula which does the alignment for you directly. Rather, they all involve an appropriate number of appropriately chosen, appropriately trained, experienced and knowledgeable reviewers who bring their knowledge to bear in evaluating alignment. A specific number of reviewers has not been empirically determined, but Webb (1999) suggested at least three. The reviewers should include
content-area experts as well as those who know the assessment very well and the standards on which the assessment is based. In order to ensure reliable and stable results, the reviewers should be extensively trained with appropriate materials like anchor items, and their work should be appropriately coordinated. For example, monitoring checks of inter-reviewer agreement can be incorporated at different points until there is sufficient evidence that they are interpreting the standards and assessment items in the same way.

Another obvious commonality these methods share is the important role the content plays in all the models, although different names for this criterion might be assumed in different models. Bhola et al. (2003) classified the current applied alignment models into three categories: low complexity, moderate complexity, and high complexity.

- Low complexity models are those that employ only the content-related criteria. They not only yield content evidence that allows appropriate interpretation and use of test scores, but also represent a logical starting point for conducting studies to align assessments with content standards. They define alignment as the extent to which the items on a test match relevant content standards (or test specifications).

- The moderate complexity models are those in which the assessment items are checked on the dual perspectives of content match and cognitive complexity match.

- High complexity models are those that include more criteria than those in the second category. Among the five models reviewed by us, the SEC model falls into the second category, while the rest belong to high complexity models.
Issues in Applying the Models

Bhola et al. (2003) had a good summary of some issues related to applying these models. To be specific, they listed the following three problems:

1) Problems associated with specificity of the alignment criteria. That is, the problems that occur when items have no standards at the corresponding level of specificity. For example, a problem arises with components of an assessment system that are designed to be interpreted holistically rather than focused on particular content standards as in standardized norm-referenced tests. They also noted in their paper that different states’ standards within the same content area are likely to have quite different structural requirements and also quite different levels regarding depth of knowledge. Thus states that seem to have comparable standards may be similar at only a superficial level, and therefore alignment of one state’s standards to an assessment does not mean that the same assessment would align in the same way with another state’s standards, even tough the standards seem, on the surface, to be comparable.

2) A problem also arises with the need to classify students into performance categories based on levels of proficiency, especially when more than two performance categories are involved. According to the authors, the first reason for this is that each performance category within which examinees will be classified adds to the number of items required in the assessment to facilitate making reliable classification decisions at different points along the proficiency continuum. La Marca et al. (2003) suggested ways to judge if the criterion of Accessibility is met in the context of alignment, but they failed to indicate estimated minimum numbers of items needed to make appropriate classification decisions for multiple performance categories. Webb (1999), on the other
hand, suggested having at least six items related to a standard to classify students into two levels with minimal classification error. However, this suggestion has to be based on several assumptions and simply multiplying by a constant such as the number of performance categories will not provide the answer to the minimum number of items needed for multiple performance categories. Thus, it is necessary for an assessment to have enough items and for the items in the assessment to span a wide enough range of difficulty to permit students along the whole proficiency continuum to demonstrate their ability. Another reason the authors gave is especially applicable to commercially available standardized tests when they include several subtests. Some of the subtests are narrowly focused (like spelling) while others are broadly focused (like science). A narrowly focused subtest may have enough items across the range of difficulty to permit classification of examinees into categories, while not providing adequate coverage of the standards. The broadly focused one may do well in covering the full range of standards, but the number of items for each standard may be insufficient for classifying students into more than two performance categories.

3) The third problem concerns the training of reviewers. According to the paper, one of the most difficult aspects of training, especially when judging commercially available tests, is finding a sufficient number of high-quality test questions to use in the training, due to practical reasons such as test security. Another difficult element in the training is reducing the tendency of the participants to be overly generous in the number of matches that they find. We anticipate this problem in the alignment that is being proposed for the Massachusetts Department of Education in 2004-2005.
Other Issues

After carefully scrutinizing the methods, their applications, and some other related materials, we concur with the authors on the problems listed above. Take problem one for example, we did find there are inconsistencies in the nomenclature used to articulate standards across states. Also, states may have statements of different levels of specificity within standards, and Webb (1997) proposed that the primary goal of alignment is to make sure that assessments are aligned with the more general statements of outcome. We also realized that some problems are inherent in the methodology, such as problem two listed above (see pages 16 and 17). At the same time, we think the following issues related to the models are also worth noticing:

1) The models share some common criteria, especially content and cognitive criteria which are the bases for the alignment analysis for all the models. However, different names are assumed for these same two aspects in different models, which makes is a bit confusing for users.

2) Some models may have too many criteria to be practical. Webb (1997) himself noted that reaching full agreement between expectations and assessments on all criteria is extremely difficult, and that trade offs must be made because real constraints exist for any educational entity, including resources, time, and legal authority. Only the subcategories of the Content Focus criterion were applied in his empirical alignment studies.

3) By saying some models have too many criteria to be practical, and content and cognitive criteria are the basis for all alignment analysis, we don’t mean to imply the rest are not important. Rather, they can be taken better care of by other procedures along the
test development process. For example, in explaining his Equity and Fairness criterion, Webb (1997) quoted The South Carolina mathematics Framework as saying that “To be fair to all, assessments must be sensitive to cultural, racial, and gender differences” (p.110). This can be dealt with in the test development process by sensitivity review (Ramsey, 1993, Sireci & Mullane, 1994). He also noted that the wording in the item may affect student performance. This again can be dealt with in a Differential Item Function (DIF) study with Mantel-Haenszel or logistic regression procedures that are especially designed for this purpose.

Conclusions

Implications for Validity (content, curricular, consequential)

Webb (1999) held that alignment is intimately related to test “validity,” and that important and useful distinctions can be made between the two concepts. He quoted Cronbach (1971) as saying that validity refers to the appropriateness of inferences made from information produced by an assessment, while he believed alignment refers to “how well all policy elements in a system work together to guide instruction and, ultimately, student learning” (p.4.). However, he further stated that of the many different types of validity, alignment corresponds most closely with content validity and consequential validity.

Judging from its function, we think alignment analysis is obviously a matter of validity, and that it corresponds to content validity, curriculum validity, and consequential validity. As mentioned in the beginning of the paper, all five models either focus completely on or devote most attention to the content aspect; that is, their main task is to make sure that the content covered by the assessment is congruent with what is
stipulated in the standards. Conceptually, the content validity of the assessment is demonstrated by the extent to which the questions on the tests reflect the standards, and it is ensured by a carefully and thoughtfully planned test construction process that is guided by test specifications or a test blueprint. The basic components of a test blueprint are the specifications of cognitive processes and the description of content to be covered by the test. Construct and content domain definitions included in assessment specifications generally identify the range of content, or some decision rule for generating assessment activities, to assure full coverage of content (Sireci, 1998). So what content validity deals with is the exact basis and most important part of alignment analysis. In this sense, analyzing the alignment of the assessment and the standards is an important way of providing evidence of content validity. Also, the methods employed in evaluating alignment and content validity (expert judgment) are very close – both involve subject matter experts (SMEs) to rate test items according to their relationship with the content domain that is supposed to be tested by an assessment.

By the same token, analyzing the alignment between the assessment and instruction/curriculum is one way of providing evidence of curricular validity for that assessment. Also, the degree of alignment between the assessment and instruction/curriculum directly affects the degree to which valid and meaningful inferences about student learning can be made from assessment data. Some important decisions are made based on these inferences; therefore, the alignment analysis also has consequential validity.

Recommendations (to states, test developers, psychometricians)
Important as it is, the present state of alignment is still weak, with sound methodologies for examining and documenting it just recently emerging (Herman et al, 2003). Our review of the literature has yielded a number of recommendations for helping to promote a better alignment between assessment and standards.

1) The first thing to bear in mind is that it is pointless to talk about alignment between the assessment and standards if the assessment itself is not psychometrically sound, or the standards are vague or not carefully developed. Also, classroom instruction must be based on the standards listed in the curriculum frameworks.

2) Alignment can be achieved through sound standards and assessment development practices. It should not be treated as a completely post-hoc activity to take place after the assessment is built. Instead, consideration should be given to alignment from the beginning and throughout the process of building the assessment.

Let us review the test development procedures briefly: First, an appropriately selected expert panel will determine the appropriate elements of the standards that will be tested. Then a test specifications table or test blueprint will be built, indicating the content and cognitive levels of the test, the percentage of the test items that will be dedicated to each standard or each content and cognitive level. The item writers, after being trained appropriately, will write items to measure skills and knowledge at these content and cognitive levels. If student performance on the test is classified into performance levels, performance standard setting should be performed appropriately to get the cut scores. Then either classical test theory (CTT) or item response theory (IRT) framework can be used to pick items to address the problem of how to choose an appropriate range of item difficulties. The assessment items can then be reviewed by another expert panel for
sensitivity, and bias issues. It is then field tested, and the data on all items are analyzed. A DIF study should be performed to make sure the test is fair to all examinees. If the standards on which the assessment is based is sound, the assessment thus built will meet all the criteria for the alignment models.

3) It is highly desirable to have a reasonable test development sequence. That is, the assessment development should proceed from the establishment of the learning standards to the establishment of curriculum to the development of test items. In this way, the elements will be aligned with each other naturally. According to Webb (1999), one of the reasons it is hard to judge alignment between assessment and expectations is that both expectations and assessments are frequently expressed in multiple pieces of documents and therefore this makes it hard to assemble a complete picture.

4) When choosing a model to work with, it is important to understand that the more criteria the model has and the more complex the model is, the less likely the assessment will be found to be closely aligned with the standards. It is also evident that alignment models that use the same number of criteria will not necessarily yield the same alignment results, so the definition of the criteria must be carefully considered.

5) In applying the chosen model, an appropriate cross-section of reviewers must be chosen and given extensive training. As Webb (1999) pointed out in his research, issues that arose from an analysis of the process indicated that reviewers could benefit from more training at the beginning of the institute. He also noted that reviewers needed more clarification of the four depth-of-knowledge levels and more explicit rules for assigning an assessment item to more than one statement of expectation.
6) The alignment should be treated as a dynamic rather than a static activity, just as the content validity of tests must be examined every year when new tests are developed and the program matures (Sireci, 1998).

Alignment of content standards, instruction, and assessments is a core validity issue. Each of the five operational methods reviewed shows that evaluating alignment is in a large part providing evidence for content validity. The models are generally sound and have produced credible results according to the empirical studies in which they are used. However, remaining issues must be solved in order to get more scientific and reliable alignment results to assist education leaders in making informed decisions about how to improve their capacity for determining the alignment.
References


