EVALUATING THE ALIGNMENT AMONG CURRICULUM, INSTRUCTION, AND ASSESSMENTS: IMPLICATIONS AND APPLICATIONS FOR RESEARCH AND PRACTICE

ANDREW T. ROACH  
Georgia State University  
BRADLEY C. NIEBLING  
Heartland Area Education Agency 11  
ALEXANDER KURZ  
Peabody College of Vanderbilt University

Alignment has been defined as the extent to which curricular expectations and assessments are in agreement and work together to provide guidance for educators’ efforts to facilitate students’ progress toward desirable academic outcomes. The Council of Chief State School Officers has identified three preferred models as frameworks for evaluating alignment: Webb’s alignment model, the Surveys of Enacted Curriculum model, and the Achieve model. Each model consists of a series of indices that summarize or describe the general match or coherence between state standards, large-scale assessments, and, in some cases, classroom instruction. This article provides an overview of these frameworks for evaluating alignment and their applications in educational practice and the research literature. After providing an introduction to the use of alignment to evaluate large-scale accountability systems, the article presents potential extensions of alignment for use with vulnerable populations (e.g., students with disabilities, preschoolers), individual students, and classroom teachers. These proposed applications can provide information for facilitating efforts to improve teachers’ classroom instruction and students’ educational achievement.

The current zeitgeist surrounding school reform and increased student achievement has created an environment where the curriculum, instruction, and assessments used in classrooms are under increased scrutiny. Accountability and large-scale assessment systems are based on a theory of action that assumes increased information about student achievement, coupled with salient incentives for increased performance (and corresponding punishments for lack of improvement), will motivate educators and produce improved student outcomes (Baker & Linn, 2002). Thus, the development and implementation of large-scale assessment and accountability programs are viewed by advocates of standards-based reform as one potential policy “lever” for improving classroom instruction and increasing equity across the educational system (Resnick, Rothman, Slattery, & Vranek, 2003). Many researchers, however, have suggested that the theory of action underlying these accountability systems may be overly simplistic in its understanding of educational improvement (Elmore, 2003; Fullan, 2003; O’Day, 2002).

Overload and fragmentation are major barriers to the successful implementation of accountability and standards-based educational reform (Fullan, 1993, 2003). The content of instructional programs, state content standards, and assessments designed to measure student achievement may contradict each other, creating increasing levels of stress and pressure for educators and students. Accountability systems posit that to ensure effective schooling, the design and implementation of three components of the educational environment—curriculum, instruction, and assessment—must be coordinated (Elliott, Braden, & White, 2001; Webb, 1997, 2002). The degree to which these components work together to facilitate student learning is often referred to as alignment.

Although developing and documenting the alignment among elements of the educational system is federally mandated, it has received scant research attention in school psychology and related fields. Alignment, however, is an area of research and applied practice that has the potential to have...
a positive impact on all students’ learning and achievement. If individuals or groups of students are not afforded opportunities to learn the skills and concepts on standards-focused assessments, they can easily become scapegoats for schools’, districts’, and states’ inability to reach the federal mandate for adequate yearly progress (AYP). “During the past half-century there has been a growing body of evidence supporting a fundamental educational truism: that what and how much students are taught is associated with, and likely influences, what and how much they learn” (Anderson, 2002, p. 255). In this article, we outline key federal policies that emphasize the importance of alignment in contributing to student achievement, three different methods of measuring alignment, and opportunities to integrate alignment into school psychology practice and research.

**Federal Policies**

Two main federal policies have direct implications for the alignment issues discussed in this article: No Child Left Behind (NCLB) and the Individuals with Disabilities Education Improvement Act (IDEIA). Both are recent reauthorizations of existing education legislation. The most recent reauthorizations have brought the two policies closer together in terms of requirements and expectations for educating students. Despite these similarities, each still has a unique purpose, both in terms of accountability for student achievement and alignment-related issues.

**No Child Left Behind.** The 2001 reauthorization of the Elementary and Secondary Education Act, now known as NCLB, placed an unprecedented emphasis on student achievement in conjunction with increased levels of accountability for the professionals that work with students in our public schools. Although an extensive review of the changes and key features of this policy is beyond the scope of this article, a brief highlighting of some of the key provisions is warranted.

Under NCLB, beginning in the 2005–2006 school year, schools are required to administer tests in reading and mathematics in grades 3 through 8 and once in high school (U.S. Department of Education, n.d.). With this increased amount of testing and the accompanying focus on test results, come requirements for more sophisticated and detailed methods of analyzing the data. For example, NCLB requires student achievement data to be disaggregated into a number of different categories, including by ethnicity, disability, and socioeconomic status. Furthermore, all students must meet state-defined criteria for proficiency by the 2013–2014 school year. Schools, districts, and states are required to demonstrate that the number of students achieving the defined levels of proficiency continues to increase until all students have reached proficiency. This is known as making AYP, and, if schools, districts, and states are not able to meet the set AYP objectives, a system of consequences can be invoked, including loss of Title I funds or restriction of local decision making and control.

Another key feature of NCLB, which does not receive as much ongoing attention as student achievement and AYP objectives, is the requirement that schools, districts, and states have a system of K–12 standards that are “aligned” with the assessments used in the state accountability system (U.S. Department of Education, n.d.). Schools, districts, and states are not considered to be in compliance with NCLB until they have demonstrated that the assessment tools used in their state accountability system have appropriate technical adequacy and are aligned with standards.

**Individuals with Disabilities Education Improvement Act.** The reauthorization of the Individuals with Disabilities Education Act (IDEA), now known as the IDEIA, brings with it some significant changes and increased accountability for schools, districts, and states for the achievement of students with disabilities. Again, an extensive review of this legislation or the changes it mandates is beyond the scope of this article.

The changes that relate most directly to this discussion are those that more closely link IDEIA to NCLB. Of particular interest is the introduction of several outcome indicators for which schools,
districts, and states are required to collect data and report results (Congressional Research Service, 2005). Data from these outcome indicators are to provide information on the impact of special education services for students with disabilities. One of the indicators that most closely connects IDEIA to NCLB is that the disaggregated subgroup of students with disabilities needs to meet AYP objectives. If students with disabilities do not meet AYP growth objectives, schools, districts, and states are subject to sanctions similar to those set forth in NCLB. What sets the IDEIA sanctions apart from the NCLB sanctions is that they are not limited to schools and districts that receive Title 1 funding. Instead, any school, district, or state that receives federal special education funding through IDEIA is open to potential sanctions.

Alignment of Curriculum, Instruction, and Assessments

Although discussions of NCLB and IDEIA do not typically focus on issues around alignment, it is nevertheless a critical feature underlying these policies, particularly NCLB. One common definition of alignment is the extent “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” (Webb, 2002, p. 1). This particular definition focuses on assessments and “expectations,” the latter of which could be interpreted as standards, a written curriculum, or perhaps instructional content. For purposes of this article, we use the terms curriculum, instruction, and assessment (CIA) as the educational components being examined for degree of alignment. One of the concluding statements from the Third International Mathematics and Science Study (TIMSS) gives perhaps the most comprehensive yet concise rationale for examining CIA alignment. Schmidt, McKnight, Cogan, Jakwerth, and Houang (1999) referred to the findings from this study as indicating that the curriculum in math and science in the United States is “a mile wide and an inch deep.” The “mile wide” statement can be understood as a reference to the breadth of curricular content, or the number of different topics covered in the curricular content. The “inch deep” statement can be understood as a reference to the depth of curricular content, or the emphasis and intensity with which the instructional content is taught. In the case of TIMSS, the results indicated that the mathematics and science curricula in the United States generally focuses on a wide variety of different topics/content (i.e., extensive breadth), but does not place a lot of emphasis or instructional intensity on this curricular content (i.e., limited depth).

Webb (1997) outlined three methods for establishing the alignment among the policy elements of curriculum, instruction, and assessment systems: (a) sequential development, (b) expert review, and (c) document analysis. Sequential development involves creation and acceptance of one policy element, which subsequently serves as a “blueprint” for the creation of additional policy elements. For example, a state or district might develop content standards for mathematics that provide guidance for the selection of a new performance-focused mathematics curriculum and the development of performance-based mathematics assessments. The process of expert review involves the convening of a panel of content experts to review the policy elements and determine the extent of their “match” or alignment. Document analysis involves the coding and analysis of documents that represent the different policy elements. By integrating these three methods, test developers, instructional leaders, and educational policy makers can increase the coherence of reform efforts and accountability programs (Webb, 1997).

In light of the results of studies such as TIMSS and the passage of federal legislation such as NCLB, a number of more rigorous methods have been developed to evaluate CIA alignment. Various state agencies and professional organizations have endorsed or at least highlighted different alignment strategies. In particular, the Council of Chief State School Officers (CCSSO) has identified three preferred models as frameworks for use in designing and implementing alignment studies:
(a) Webb’s alignment model, (b) the Surveys of Enacted Curriculum (SEC) model, and (c) the Achieve model (CCSSO, 2005). The American Federation of Teachers (2001) and National Council on Accreditation in Teacher Preparation (NCATE; 2001) have also developed documents that make alignment recommendations, but neither organization has developed an actual evaluative framework.

It is important to note that descriptions of a system’s alignment are not dichotomous (i.e., in alignment vs. out of alignment). Rather, each model endorsed by CCSSO consists of a series of indices that, when considered collectively, provide a summary of the general match or coherence between state standards, large-scale assessments, and, in some cases, classroom instruction (Resnick et al., 2003). An extensive review of the model/methods of measuring alignment is beyond the scope of this article. However, these three methods—Webb’s alignment model, SEC, and the Achieve model—have received sustained attention in research literature and practice, and warrant further discussion as examples of what might be considered “best practice” options in the area of measuring alignment. Figure 1 provides an overview of the main features of each CIA alignment model.

Webb’s Alignment Model

Webb’s alignment model provides a series of statistics that indicate the correspondence between the content in the state’s academic standards and the content covered by the state assessment, and is intended to provide policy makers with comparative data on the alignment of these two elements of the state’s accountability system. Webb’s model has been used to judge the alignment between standards and general large-scale assessments for language arts, mathematics, social studies, and science in many school districts and more than 20 states. In addition, it is the only alignment analysis approach that had been adapted and applied to state alternate assessments for students with disabilities (Roach, Elliott, & Webb, 2005). The information provided by alignment analyses using Webb’s methods is typically used by state policy makers in modifying assessments, revising content standards, and verifying the extent that various policy elements are directed toward common expectations for learning.

In studies using Webb’s method, panels of educators and curriculum experts are trained to use an analytic process and heuristics to rate the alignment between states’ assessments and academic standards. The primary role of the panel members is to complete the following tasks:

1. Reach consensus on a depth-of-knowledge (DOK) level rating for each objective/benchmark in the state content standards.
2. Rate the DOK level of each assessment task or item.
3. Identify the one or two objectives/benchmarks from the content standards to which each assessment item corresponds.

Before independently completing their ratings, panel members are trained to identify the DOK level (Figure 2) for objectives from the standards and assessment items. This training includes the review of specific descriptions for the four DOK levels in each subject domain covered by an assessment. Panel members reach consensus on the appropriate DOK levels for curriculum objectives in each content domain before completing their individual ratings of assessment items in that content domain. According to Webb (2002), working as a group to reach consensus on the DOK levels for each objective provides an opportunity for discussion of the rating criteria, resulting in “calibration” of panel members’ understanding of the DOK rating process. Following this “calibration” process, panel members are asked to assign a DOK rating to each assessment item. After assigning the DOK rating for each item, panel members complete the coding process by identifying the one or two objectives from the standards that corresponded to each assessment item. The alignment coding
**Figure 1**
*Overview of major alignment models.*

<table>
<thead>
<tr>
<th>Components Evaluated for Alignment</th>
<th>Webb</th>
<th>Surveys of Enacted Curriculum</th>
<th>Achieve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessments Standards</td>
<td>Assessments Standards and Curricular Materials Classroom Instruction</td>
<td>Assessments (Items and Item Sets) Standards</td>
<td></td>
</tr>
<tr>
<td>Raters or Evaluators</td>
<td>Alignment panel of 6 to 8 educators with subject area expertise</td>
<td>Individual teacher (Classroom Instruction); Alignment panel of 3 or more content area specialists</td>
<td></td>
</tr>
<tr>
<td>Alignment Evaluation Process</td>
<td>1. Panel members are trained to recognize and apply four depth-of-knowledge (DOK) levels. 2. Panel reaches consensus on DOK level ratings for objectives from content standards. 3. Panel members then independently rate the DOK level and corresponding objective from standards for each assessment item.</td>
<td>1. Teachers complete Surveys of Enacted Curriculum ratings at the end of the year. Survey includes ratings level of coverage for topics and subtopics taught and the level of cognitive demand for tasks in each topical area. 2. Panel members rate the level of coverage for topics and subtopics and cognitive demand of tasks and activities for standards, curricular materials, and assessments.</td>
<td>1. Expert panels make consensus judgments regarding the quality of the content and performance match between individual test items and their respective standards. Each item is further evaluated regarding the source of its difficulty. 2. Panels then judge whether entire item sets assess the respective standards with a comparable emphasis and range of expectations. Each set of items is further evaluated regarding the grade-level appropriateness for its span of difficulty.</td>
</tr>
</tbody>
</table>

**Breadth criteria**
- Categorical Concurrence
- Range of Knowledge
- Balance of Representation

**Depth criteria**
- DOK Consistency
- Cognitive demand categories Emphasis ratings within cognitive demand

The process is not designed to produce exact agreement between members of the expert panel. In fact, variance in ratings may represent valid differences in opinion that reflect a lack of clarity in the descriptions of objectives in the content standards or the robustness of assessment items that could reasonably correspond to more than one curricular objective (Webb, 2002).

Following completion of the expert panels’ ratings, analyses of the ratings result in descriptive statistics for each of the four criteria included in Webb’s alignment model. Analyses of the
Figure 2

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1: Recall</td>
<td>This level includes the recall of information such as a fact, definition, term, or simple procedure, as well as performing a simple algorithm or applying a formula.</td>
</tr>
<tr>
<td>Level 2: Skill/Concept</td>
<td>This level includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach a problem or activity. Key words that distinguish a Level 2 item or task include “classify,” “organize,” “estimate,” “make observations,” “collect and display data,” and “compare data.”</td>
</tr>
<tr>
<td>Level 3: Strategic Thinking</td>
<td>This level includes items that require reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3 attribute. Students might also be required to make conjectures or determine a solution to a problem with multiple correct answers at this level.</td>
</tr>
<tr>
<td>Level 4: Extended Thinking</td>
<td>This level includes items that require complex reasoning, planning, developing, and thinking most likely over an extended period of time. At Level 4, the cognitive demands of the task should be high, and the work should be very complex. Students should be required to make connections both within and between subject domains. Level 4 activities include designing and conducting experiments, making connections between a finding and related concepts, combining and synthesizing ideas into new concepts, and critiquing literary pieces and experimental designs.</td>
</tr>
</tbody>
</table>

Panel members’ responses provide information on assessments’ attainment of the following alignment criteria: (a) categorical concurrence, (b) balance of representation, (c) range-of-knowledge correspondence, and (d) DOK consistency. The first three criteria measure the correspondence between skills and concepts covered by the state’s content standards and objectives (i.e., performance standards) and the skills and concepts tested by an assessment. Categorical concurrence indicates whether the same or consistent categories of content appear in both the content standards and the assessment items. Range-of-knowledge correspondence indicates whether the span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need in order to correctly answer collection of assessment items or activities. Balance of representation provides an index of the degree to which one curriculum objective or benchmark is given more emphasis on the assessment than another. Conversely, DOK consistency is intended to represent the level of complexity required to respond to the objectives and assessment items. The DOK criterion indicates whether what is elicited from students on an assessment is as complex for the content area as what students are expected to know and do, as stated in the state’s content standards. An advantage of the Webb model is the inclusion of specific levels for each criterion (Figure 3) to help educators and policy makers determine the adequacy of the alignment between content standards and assessments.

The Survey of Enacted Curriculum

Porter and Smithson (2000, 2001, 2002; Porter, 1991, 2002) have developed a set of procedures for conducting content analyses, and subsequent data analyses, of different sets of educational materials (e.g., assessments, standards, textbooks). The primary features of the SEC methodology include (a) a common language framework for examining the content of CIA, (b) a single alignment statistic, and (c) graphical output of CIA content. Currently, the SEC tools are being used in at least 17 states, including major projects in at least 7 school districts.
Figure 3


<table>
<thead>
<tr>
<th>Criteria</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categorical Concurrency</td>
<td>An assessment must have at least six items measuring content for each standard in order to demonstrate an acceptable categorical concurrence between the standard and the assessment. “The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students’ mastery of content on that subscale. . . . Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63” (Webb, 2002).</td>
</tr>
<tr>
<td>Range of Knowledge</td>
<td>At least 50% of the objectives for a standard corresponded with at least one related WAA item based on the ratings of Alignment Institute panel members. This criterion is based on the assumption that an assessment should test students’ understanding or mastery of the majority of the knowledge (i.e., more than half the objectives) represented by any given standard (Webb, 2002).</td>
</tr>
<tr>
<td>Balance of Representation</td>
<td>A balance index score was computed to judge the distribution of assessment items. “The balance index compares the proportion of items for each objective to proportion if the assessment items were evenly distributed among all possible objectives” (Webb, Horton, &amp; O’Neal, 2002). An index value of 0.7 or greater indicated that WAA items are distributed among all objectives to an acceptable degree without forming a monominal or binominal distribution of assessment items on objectives under a standard.</td>
</tr>
<tr>
<td>Depth of Knowledge</td>
<td>“For consistency between the assessment and standard... at least 50% of the items corresponding to an objective had to be at or above the level of knowledge of the objective” (Webb, 2002, p. 4). Meeting this criterion suggests a test demands the adequate depth of understanding and sufficient mastery of the knowledge and skills covered in the corresponding academic standards.</td>
</tr>
</tbody>
</table>

Common Language Framework. One of the key features of the SEC tools is that content is described within a common language framework, regardless of what is being examined (e.g., instructional content, assessments, standards), which allows for direct comparisons between the CIA elements (Porter, 2002). The framework offers a list of general categories under which a series of more discrete topics is organized. For example, the subtopics of “add and subtract whole numbers” and “multiply whole numbers” would be listed under the larger topical category of “Operations.” At the same time, SEC provides a list of descriptors of cognitive demands or expectations (e.g., memorize, solve nonroutine problems) that students might engage in for each topic and subtopic listed. A two-dimensional topic-by-cognitive demand framework results from examining the intersection of topics covered and the cognitive expectations for student performance with those topics (Table 1 and Figure 4).

SEC also includes strategies for evaluating emphasis of various content topics in curriculum, instruction, and assessment. For example, emphasis is captured on the teacher survey by asking teachers to report the level of coverage for each topic (i.e., how much time is spent on each topic) and the relative emphasis for each cognitive demand (i.e., amount of time spent on each cognitive demand). Conversely, for assessments, emphasis could be captured by the points allocated to each test item. For each subcomponent described (e.g., test item, standard, teacher survey topic-by-cognitive demand intersection point), the proportion of emphasis is calculated relative to the whole of that element. Put another way, the sum of the ratings for any CIA element will be 1.0. Therefore, each unit analyzed (e.g., test items, standard) is a proportion of the whole (Table 1).
Table 1
Generic Surveys of Enacted Curriculum Content Matrix with Proportions of Emphasis

<table>
<thead>
<tr>
<th>General Topic</th>
<th>Category 1</th>
<th>Category 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtopic 1</td>
<td>0.25</td>
<td>0.50</td>
</tr>
<tr>
<td>Subtopic 2</td>
<td>0</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Alignment Statistic. Although the SEC methodology and tools described previously address the content analysis of one CIA component in isolation, these methods and tools do not compare the alignment between two components by themselves. Two content matrices must be compared to each other to examine alignment. This comparison results in a statistic known as an alignment index (AI). At its simplest conceptual level, the AI can be considered the sum of all the cell-by-cell intersection points (Porter, 2002). Using this method, the smaller of the two scores is used as the number in the equation. Although Porter has referred to and tested other methods of calculating the AI, the method used in this example is the one most typically used and referred to in alignment studies using the SEC because it is readily understood and calculated. Within the SEC framework, alignment is best understood as a matter of degree, rather than an absolute. Although two CIA components could be in “perfect” alignment (i.e., an AI score of 1.0), a more typical result is a score between 0 (no match) and 1.0 (perfect alignment).

Visual Display of Data. Porter and Smithson (2001; Porter, 2002; have developed a unique method of visually displaying alignment results, namely, the use of topographical maps (Figure 5). Displaying the content coverage and emphasis data in this manner yields a number of benefits. First, it allows for a visual summary of where different points of coverage and emphasis are within a CIA component. Second, when comparing two content maps to each other, it allows for a visual analysis of where the points of alignment, or lack of alignment, are between two CIA components. Third, the visual presentation of the topic-by-cognitive demand intersections is a relatively easy way for researchers and practitioners alike to interpret the CIA content data (Porter, 2002).

Data Collection Procedures. SEC data collection varies depending on which CIA components are being analyzed. The collection of instructional data is done via the SEC teacher surveys on which teachers report on the instructional content taught in their classrooms. Teachers typically report on the content of their instruction at the end of the school year. Within the common topic-by-cognitive demand framework, teachers are asked to rate the emphasis placed on different topics and cognitive demand.

Figure 4
Surveys of Enacted Curriculum descriptions of level of coverage and cognitive demand.

<table>
<thead>
<tr>
<th>Category of Emphasis</th>
<th>Scale of Emphasis Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of Coverage</td>
<td>(a) None/not covered; (b) slight coverage (less than one class/lesson); (c) moderate coverage (one to five classes/lessons); (d) sustained coverage (more than five classes/lessons)</td>
</tr>
<tr>
<td>Cognitive Demand</td>
<td>(a) No emphasis; (b) slight emphasis (less than 25% of time spent on this topic); (c) moderate emphasis (accounts for 25%–33% of time spent on this topic); (d) sustained emphasis (accounts for more than 33% of time spent on this topic)</td>
</tr>
</tbody>
</table>
demands. What results are data that can be converted to proportions that describe the amount of time spent at each topic-by-cognitive demand intersection point or cell, and in conjunction with assessment, standards, or instructional support material content data to calculate an AI.

When collecting data for assessments, standards, and instructional support materials, a panel is convened to do the content analysis. Panel members are trained on the common language framework, the method of distinguishing relative emphasis for the materials that are being coded, and how to accurately complete the coding process. Then, panel members independently complete the content analyses of the documents and submit their rating materials.

The rules for coding the data for assessment instrument content are relatively straightforward. First, the amount of points available for the entire assessment tool, as well as each individual item, must be determined. Once this is determined, the common language framework of the SEC is used.

---

**FIGURE 5.** Analysis of alignment for fourth-grade curriculum, performance standards, and the Wisconsin Knowledge and Concepts Exam. *Note.* From “Alignment of State Testing Programs NAEP and Reports of Teacher Practice in Mathematics and Science in Grades 4 and 8,” by A. C. Porter and J. L. Smithson, 2000, paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Adapted with permission of the authors.
to indicate which topic-by-cognitive demands are covered by each test item. The number of topic-by-cognitive demand combinations is then distributed proportionately across the points available for that item. For example, if a test item is worth one point, and there are two topic-by-cognitive demand combinations associated with that test item, then each topic-by-cognitive demand combination for that item is assigned half of the value of the item, in this case, 50% or 0.50. The sum of the topic-by-cognitive demand combinations, when divided by the total number of points, will equal 1.0. Therefore, the relative emphasis of each topic-by-cognitive demand combination is calculated by dividing that proportion by the total number of points available on the test.

The same identification number system used to code assessment items is used for coding standards as well. Panel members are given a copy of the content-specific standards (e.g., reading, math, science) to examine. For each standard, panel members indicate which topic-by-cognitive demand combinations are covered in the standard. Once the standards have been coded, the total number of topic-by-cognitive demand combinations is summed. The proportion of emphasis is calculated by dividing the number of “hits” each topic-by-cognitive demand combination has by the total number of topic-by-cognitive demand combinations. The sum of all these proportions will equal 1.0.

**The Achieve Model**

The Achieve model was originally developed in 1998 at the Learning Research and Development Center at the University of Pittsburgh. Achieve, Inc., an independent and bipartisan organization created by governors and chief executive officers, now offers alignment studies and additional services to states and districts in the area of reading/language arts, math, and science. The model uses an assessment-to-standards alignment protocol to provide a qualitative as well as quantitative analysis of alignment featuring judgments about the quality and rigor of individual test items and sets of items (i.e., entire test). According to Achieve, Inc., their model has been used in 14 states to analyze the overall quality of tests and alignment of assessments with state standards.

**Alignment Protocol.** The Achieve alignment protocol requires both low- and high-level inferences from its reviewers. To prevent excessive training time, particularly for the high inference judgments at the item set level, Achieve staffs its panels with experienced and knowledgeable reviewers, including classroom teachers, curriculum specialists, and subject-matter experts. Similar to the previously discussed models, expert judgment is guided along certain criteria, in this case (a) content centrality, (b) performance centrality, (c) challenge, (d) balance, and (e) range (CCSSO, 2002). Sample items and previously analyzed tests are used to familiarize panel members with the various rating criteria and train them in the application process. As part of the preparatory training, reviewers are also asked to take the assessment in question in its entirety to provide them with an authentic understanding of the test’s demand and focus. Panels consist of three or more reviewers whose decision-making process is discussion driven and consensus based. Generally speaking, the alignment protocol guides participants in gathering information about (a) the test’s breadth in relation to standards (i.e., test measures only the content and skills expressed in the standards), (b) the test’s overall sampling of content and skills (i.e., test samples content and skills evenly and adequately), and (c) the test’s level of difficulty (i.e., test reflects those levels of thinking prescribed by the standards) (Resnick et al., 2004).

**Alignment Procedure.** The alignment protocol is first applied to each test item individually and then to the test as a whole. At the item level, alignment analysis begins with the creation of a “blueprint” mapping each test item to a corresponding standard. The initial blueprint is typically established by the state or test contractor and subsequently confirmed by a senior reviewer who
flags confusing or nonmatching items for further analysis (i.e., reassignment, elimination). Items are always matched to the most specific outcome criteria put forth in the standard (e.g., objectives). Once the blueprint has been confirmed, alignment between test items and standards is analyzed by several panels across the following criteria: content centrality, performance centrality, and source of challenge. When analyzing content centrality, reviewers evaluate standard specificity and face validity of matching content between item and standard by ascribing one of the following scores: 2 = clearly consistent (i.e., item assesses the exact content articulated in the standard/objective); 1A = not specific enough (i.e., standard is too broad to confidently judge item alignment); 1B = somewhat consistent (i.e., item samples only part of a standard); and 0 = inconsistent (i.e., item only marginally assesses what is prescribed by the standard) (Resnick et al., 2003). After having judged the quality of content match between the item and corresponding standard, reviewers turn their attention to the degree of match regarding the required types of performances. Performance centrality, akin to the SEC’s cognitive demand criterion, requires judges to score whether test items demand the same type of performance task as the related standard (e.g., “identify,” “analyze”). Performance centrality is also scored on a 4-point scale according to 2 = clearly consistent (i.e., same type and number of cognitive tasks); 1A = not specific enough (i.e., standard/objective too broad); 1B = somewhat consistent (i.e., item samples only part of the cognitive demands expressed in the standard/objective); and 0 = inconsistent (i.e., cognitive demands of test item and standard/objective do not match). The next criterion, source of challenge, asks reviewers to indicate whether a test item’s difficulty is due to appropriate or inappropriate sources of challenge. A score of 1 is given when a test item’s difficulty is appropriately located in the subject matter and performance demanded by the corresponding standard. Items whose difficulty stems from extraneous sources such as inappropriate grade-level language, misleading graphics, or unfair assumptions about a student’s background knowledge are to receive a score of 0 (i.e., inappropriate source of challenge).

After judging alignment at the item level, reviewers continue to evaluate the test as a whole via the range, balance, and level of challenge criterion. At this point, the protocol guides reviewers to assess whether the test samples the standards comprehensively and with appropriate emphases across standards and objectives. Range is expressed as the proportion of objectives assessed by at least one test item and thus represents a basic indicator of overall coverage. Under the balance criterion, all items matching a standard are considered a set. Reviewers are asked to make qualitative judgments considering whether a set reflects the corresponding standard’s emphases on content and skills along two questions: (a) what objectives in a standard seem to be overassessed, and (b) what objectives in a standard seem to be underassessed or not assessed at all (Resnick et al., 2003). Reviewers evaluate each question from two perspectives: (a) their reading of the standards and (b) their personal judgment of what is most relevant for the particular grade level in question. Balance judgments fall into one of four categories: good, appropriate, fair, or poor. Since these judgments are qualitative, the different categories are not delineated by some kind of balance percentage or index. In general, well-balanced sets are expected to dedicate more items to the corresponding standard’s most crucial content and performances.

Sets of items are further evaluated on their level of challenge, a global judgment on the test’s overall difficulty according to assessed concepts and cognitive demands placed on students. Reviewers make qualitative judgments regarding the cognitive demands of an entire set in relation to the demands specified in the matching standard as well as item skewing toward more or less challenging concepts, types, or parts of objectives. The level of challenge for sets is rated as being easy, medium, or hard. As noted by Resnick et al. (2003), tests should ideally feature items that run the whole gamut of challenge from simple tasks on accessible concepts to complex tasks on challenging concepts. A short written evaluation on each set’s level of challenge concludes the alignment protocol process.
Extensions of Alignment for Research and Practice in School Psychology

As demonstrated in the previous sections, multiple researchers have developed strategies for examining the coordination of various policy elements—curriculum, instruction, and assessments—posited as being necessary to establish effective educational programs. These methods for understanding the overlap or links among these components can provide us with an index of the “power” of educational programs to produce changes in student achievement. In the three-tiered intervention model underlying the response to intervention (RTI) movement, classroom curriculum and instruction are generally conceptualized as the primary or universal level of intervention. When classroom curriculum and instruction are well aligned with desired outcomes (as measured by an achievement test or other assessment), the majority of students might reasonably be expected to make AYP. When these components are poorly aligned with expectations, students will be less likely to demonstrate their proficiency on outcomes-focused assessments.

The concept of schooling as an intervention is also endorsed by Deno (2005) in his description of the problem-solving model:

> Education is viewed as a deliberately conceived “intervention” in children’s lives. The intervention of schooling is created by society to produce specific developmental outcomes. While members of society may disagree on those outcomes, there should be no question that the primary purpose of school is to intervene in children’s lives to produce those outcomes. (p. 14)

Alignment, therefore, represents a promising framework for analyzing the extent to which components of the educational system are coordinated, and its measurement has the potential to provide empirical evidence of the potential of classroom instruction to influence student achievement. Traditionally, alignment research has been limited to systems-level analyses of K–12 accountability programs as described previously. The underlying logic of alignment, however, may be applied to components of the educational programs provided for more vulnerable student populations (e.g., students with disabilities, preschoolers) and to individual students’ curriculum and instruction, providing information that can facilitate efforts to improve their educational achievement. In this section of the article, we present potential extensions of the concept of alignment that have the potential to influence both school psychology practice and research in this area.

Alignment Analyses of Accountability Systems for Unique Populations: Alternate Assessments and Early Childhood Accountability

Although large-scale assessments and high-stakes accountability have become “standard practice” in the United States’ K–12 general educational system, programs for students with significant disabilities and early childhood education were initially excluded from these reporting systems. Recent federal initiatives (e.g., the Government Performance and Results Act, IDEA, NCLB) have created a situation in which these programs must develop assessment and accountability systems to document the progress of the students they serve in achieving agreed-on educational outcomes. Conducting alignment analyses is an important component in providing evidence for the validity of the results from these newly developed assessments as indices of students’ progress in attaining the concepts and skills outlined in state standards.

Early Childhood Accountability. The Head Start National Reporting System (NRS) is one prominent example of an early childhood accountability system. Based on the Bush Administration’s Good Start, Grow Smart initiative and provisions of the Head Start Act, the NRS is intended to provide important information about young children’s progress in attaining the expectations described in...
the Head Start Child Outcomes Framework (2000). The Outcomes Framework is a set of early learning standards developed to guide curriculum planning and instruction in the United States’ Head Start programs. The NRS is a 15- to 20-minute assessment of early literacy and numeracy skills administered to all 4- and 5-year-old children participating in the nation’s Head Start programs. Items on the NRS assessment were drawn from other existing assessments developed and validated for use with preschool-age children.

The creation and implementation of the NRS has been a source of disagreement and controversy, with many early childhood educators and researchers questioning the validity of the assessment as an indicator of child progress and program functioning, and expressing concerns over possible “narrowing of the curriculum” as teachers focus on the skills and concepts assessed by the NRS (Hirsh-Pasek, Kochanoff, Newcombe, & deVilliers, 2005; Meisels & Atkins-Burnett, 2004). The completion of an alignment analysis would provide important information about the content validity of NRS results (i.e., “Does the NRS adequately measure the important skills and concepts outlined in the Head Start Child Outcomes Framework?”).

To date, none of the alignment models typically used to evaluate K–12 general education standards and the large-scale assessments have been modified for application to the early learning standards and assessments being developed and implemented at either the national (e.g., Head Start NRS) or state level. Kagan and Scott-Little (2004) developed a coding procedure that has been used to examine the breadth and depth of state early childhood learning standards and assessments typically used to evaluate young children’s learning. They combine breadth and depth in their coding system by using a set of 36 indicators that are organized into five domains (physical and motor, social and emotional, approaches toward learning, language and communication, and cognition and general knowledge) in terms of complexity or developmental challenge (Scott-Little, 2005). The coding in these studies, however, was done by the researchers themselves rather than the panel of educators that is typical of most alignment analyses (Scott-Little, 2005).

Development of a rigorous alignment analysis procedure for early childhood accountability systems will require studies that (a) result in modification of the existing heuristics used in K–12 alignment studies for use with pre-K assessments, and (b) pilot the use of modified alignment analysis procedures to examine the content validity of early childhood accountability systems (e.g., the Head Start NRS). It is conceivable that the language and examples included in the descriptions of the depth and breadth criteria used in the Webb, SEC, or Achieve methods could be reconceptualized for use with early learning standards and assessments. Structured collaboration and conversation among early childhood researchers, state/national policy leaders, and experienced practitioners could result in a set of revised heuristics that would guide investigations of alignment of early childhood accountability systems.

Development and implementation of these procedures have the potential to provide content-related evidence for the validity of results of the NRS and other early childhood accountability measures in use at the state and federal levels. Collection of this evidence is essential to demonstrate performance on these assessments provides an index of student progress toward and program effectiveness in teaching the important skills and concepts outlined in state and national early childhood standards.

School psychologists have an important role to play in this work. Because early childhood assessments are often curriculum based and context embedded, the collection of observational data and work samples will be necessary to demonstrate the alignment between curriculum, instruction, and assessment. By drawing on their assessment and evaluation skills, school psychologists can assist educators in documenting how instruction in their classrooms is “linked” to early childhood standards and assessment tasks.
Alternate Assessments. The reauthorized IDEA clearly mandates that students with disabilities have access to the general education curriculum and instruction focused on meeting academic standards. Specifically, IDEA requires that (a) all students participate in state- and districtwide assessments; and (b) all students have opportunities and instruction that allow them to make progress toward state and district academic standards. Recent interpretations of the NCLB (U.S. Department of Education, 2005) provisions requiring states to demonstrate AYP allow the use of alternate assessments based on alternate achievement standards to measure the proficiency of up to 1% of the students with the most significant disabilities (typically students with profound developmental disabilities). In addition, another 2% of the total student population may demonstrate their proficiency on alternate assessments based on modified achievement standards (although the policy on the modified alternate assessment had not been finalized prior to submission of this article).

Some states (e.g., Mississippi, Idaho) have applied the model developed by Webb as part of the validation process for their alternate assessments. In an effort to make their academic content standards meaningful and relevant to the needs of all students—including those with significant disabilities—both states used sequential development to create extended content standards documents based on the major subject area domains represented in the content standards. These more accessible modifications of the states’ grade-level standards subsequently served as the framework for developing the states’ alternate assessment items.

Applying the methods previously used for analyzing the alignment of curriculum standards with large-scale assessments, alignment panels in each state were trained to use a collection of analytic tools and heuristics to rate the alignment of assessment systems and academic content standards (Roach & Elliott, 2004; Roach et al., 2005). The Mississippi and Idaho alignment panel members rated the correspondence between, and the depth of knowledge required by, their state’s alternate assessment items and objectives on the states’ grade-level content standards. In both states, the alignment panels’ responses indicate that the alternate assessments are generally well aligned with the skills and knowledge represented by the states’ grade-level content standards. In fact, the performance of these alternate assessments on the four criteria comprising Webb’s (1997) alignment model met or exceeded the performance of many states’ general education assessments.

Missing from these efforts is the identification of an acceptable index of the alignment between the curricular experiences of students with disabilities and the state’s content standards. Both the IDEIA and NCLB mandate increased access to the general education curriculum for students with disabilities. Although students in special education have the legal right to individually referenced curriculum and instruction, outcomes linked to the general education program have become the optimal target. As a result, physical presence in mainstreamed settings does not meet the spirit of the law; instead students with disabilities must have instruction and accommodations that promote their progress, no matter how modest, toward the educational expectations of the larger student population (Pugach & Warger, 2001).

Development of an approach for evaluating this form of alignment is essential to (a) demonstrating progress toward the mandate to provide students access to the general curriculum and (b) meet the proposed regulations for alternate assessments based on modified achievement standards (i.e., the 2% AA). Specifically, proposed Reg. 200.1(e)(2) would require individualized education plan teams to determine if a student “is receiving instruction in the grade-level curriculum for the subjects in which the student is being assessed. . . . Students are not assessed based on modified achievement standards if they have not had the opportunity to learn grade-level content” (U.S. Department of Education, 2005, pp. 7, 10). Currently, no strategy has been identified by the U.S. Department of Education to meet these proposed criteria. One option, however, would be modification of the SEC tools and heuristics for use in evaluating individual students’ curricular experiences and opportunity
to learn the skills and concepts on states’ grade-level content standards. An initial attempt to develop a modified approach to evaluating curricular alignment was reported in a recent investigation of the influence of curricular access on students’ alternate assessment performance (Roach & Elliott, 2006).

Alignment Analyses, Opportunity to Learn, and the Role of Academic Skill Level

With the accountability requirements found in both NCLB and the IDEIA, it has been increasingly clear that all students must display “proficiency” and progress in reading, mathematics, and science. With these requirements come a number of questions, including (a) do all students have an adequate opportunity to learn the content on which they will be assessed?, and (b) do educators have adequate data from which to make decisions about differentiating both content and instructional delivery?

Opportunity to Learn. The discussion of opportunity to learn (OTL) in the classroom logically begins with defining what constitutes a legitimate OTL. Unfortunately, this is a complicated task, as evidenced by a wide variety of definitions, standards, and methods of measurement currently in the literature (Ysseldyke, Thurlow, & Shin, 1995). It does, however, appear across this array of opinion and research that OTL can be represented by the content overlap between instruction and what is tested. This idea has been extended beyond the simple overlap in content to represent more positive attempts to provide all students with learning opportunities that are appropriate given their academic and behavioral skills (Porter, 1995). Ideally, this would include using content standards, or at the very least the results of tests intended to measure the standards, to guide the content of instruction (Yoon & Resnick, 1998). The methods and materials that constitute the SEC (e.g., Porter & Smithson, 2002) represent one potential approach to examining the degree to which students have received the OTL the curricular content.

Using the SEC allows us to obtain a quantitative and qualitative description of a teacher’s instruction, as well as to calculate the degree of alignment between classroom curriculum and either standards or large-scale assessment instruments. This alignment statistic can be used to quantify the content of instruction, thereby facilitating a direct quantitative examination of how the content of instruction impacts academic achievement. For example, Gamoran, Porter, Smithson, and White (1997) found that higher degrees of alignment between the content of classroom curriculum and a standardized test of mathematics were predictive of gains in achievement for low-achieving, low-income high school students as measured by the standardized test. Overall, it appears that there is a strong correlation between the degree to which students are exposed to tested content and performance on tests of academic achievement (McPartland & Schneider, 1996). In other words, students perform better on achievement tests when they have had the opportunity to learn the content on the test.

Academic Skill Level. Deeply embedded in the discussion and application of standards-based reform efforts is the inclusion of students with disabilities in accountability systems (McDonnell, McLaughlin, & Morison, 1997). In other words, districts and schools are required to administer the same assessment to a group of students who not only have had the varying levels of OTL, but who also have a wide variety of academic skills and understanding. At a classroom level, teachers are asked to help students meet rigorous academic expectations while differentiating instructional content and delivery to meet the diverse needs of their students (e.g., Tomlinson, 1999). To take this issue a step further, one might ask: “Are equal levels of OTL adequate to meet the individual student needs within a group of diverse learners?”

Despite its importance, the interaction between OTL and academic skill level on assessment outcomes is an area that has not been examined in great detail. Some evidence regarding this
interaction can be inferred from a study conducted by Gamoran et al. (1997). Specifically, Gamoran and colleagues examined gains in achievement test scores between high school students in remedial, transition, and college preparatory mathematics classes. They found that students in the remedial courses, which primarily consisted of low-achieving, low-income students, were exposed to less rigorous curriculum, resulting in smaller gains in achievement scores on average when compared to their peers in transition or college preparatory courses.

The concept of gains in achievement has been embraced in NCLB and the IDEIA and is especially apparent in the NCLB provisions for schools and districts to measure the AYP for all subgroups of students in their school, including students with disabilities. The convergence of these federal policies, coupled with the growing body of evidence that both opportunity to learn and academic skill levels impact the students’ subsequent achievement, highlights the growing importance of having reliable and valid data on the alignment between content standards, accountability assessments, and instructional content. The qualitative, quantitative, and graphical data that results from application of the alignment tools (e.g., SEC or Achieve) can provide helpful information to guide instructional decision making and support educational achievement for all students.

Alignment: Beyond Systems-Level Policy Analysis

Although relatively new as a specific field of study, alignment research and practice has quickly risen in importance. It has received a prominent position in federal legislation, most recently in NCLB. A number of tools and procedures have been developed to either facilitate alignment among curriculum, instruction, and assessment (e.g., curriculum mapping), or to measure the degree of alignment among those elements (e.g., Webb’s tools, SEC, Achieve). Despite these advances, much work remains to advance the application of alignment concepts and tools. As leaders in areas such as assessment and instructional consultation, school psychologists are in a unique position to become leaders in using alignment tools and data as well, in both research and practice.

For example, there are a number of questions and issues that remain to be examined, beyond those already mentioned here previously, that could have potentially large impacts on student learning and educator practices. There is little clarity in terms of “how much” alignment is adequate to facilitate learning and improved outcomes for students. This issue becomes more complicated when taking into consideration the multiple methods of measuring the alignment among curriculum, instruction, and assessment. With the possible exception of the SEC, there is no coherent set of tools that provides a consistent set of tools and processes that could result in a reliable and valid degree of alignment calculation for all three components of the educational system (i.e., curriculum, instruction, assessment). More work is needed to develop consistent sets of tools and processes that allow researchers and practitioners to not only measure the degree of CIA alignment, but also to engage in ongoing practices that result in increased alignment among the components.

Another area related to the issues discussed previously is producing classroom- or student-level changes in instructional content and practices. Specifically, how can alignment data be used to change what occurs in classrooms? To date, the only available research on this topic suggests that teachers have been able to increase the degree of alignment between their instruction and either assessments or content standards by addressing more topics (J. Smithson, personal communication, July 22, 2005). Although this may result in an increase in one aspect of alignment (i.e., breadth), it is conceivable that the topics taught may not be given the appropriate emphasis (i.e., depth) to generate significant changes in learning and outcomes for students. Refining and implementing rigorous CIA alignment methods could improve the feedback mechanisms that are currently in place to share information with teachers about what they are teaching.
In addition, alignment among a state’s content standards, classroom instruction, and large-scale assessments may set the conditions for high student achievement, but alignment does not cause high student achievement. That is, alignment represents a necessary but not sufficient ingredient in the recipe for greater student achievement. Mere exposure to a well-aligned curriculum, for instance, does not guarantee student engagement in that curriculum. Alignment data may tell us something about the students’ opportunity to learn certain skills and concepts, but whether they actually realize these opportunities and to what degree is a different matter.

One aspect absent from current conceptualizations of alignment is the degree to which students are engaged in the enacted curriculum. It may seem apparent that unengaged students will not be able to translate their exposure to rigorous standards-based instruction into higher achievement on assessments. However, the exact interactions between engagement, alignment, and achievement have yet to be examined in greater detail. Student engagement and its relation to student achievement has received sustained attention over the years (e.g., Brophy & Good, 1986; DiPerna, Volpe, & Elliott, 2001; McWilliam & Bailey, 1995; Walberg, Fraser, & Welch, 1986). Considerations regarding student engagement are thus able to draw from a rich research basis. Researchers and practitioners are encouraged to consider further how aspects of student engagement may be integrated into the concept of alignment.

Although heated debates on topics such as the utility and feasibility of RTI models for identifying students with significant learning needs are likely to dominate the professional literature and research agenda for the foreseeable future, there is at least one area where we believe all school psychologists are in agreement: all students must have the opportunity to learn the core knowledge and academic skills. As we describe in this article, although there are tools that can help us answer this question, we have little evidence to suggest that these tools are used systematically to determine whether individuals or groups of students have had an adequate opportunity to learn. Nevertheless, there is promise in using these tools to help us respond to an important opportunity to learn questions.

A few years ago, there was a call for the field of school psychology to begin shifting our attention and efforts to solving the “big” instructional problems faced by teachers and students (e.g., Shapiro, 2000; Ysseldyke, 2000). Although it will not be the sole component in these efforts, there is great potential in the concepts and tools of alignment to facilitate answers to many instructional questions faced by educators and students. Questions such as “To what degree are content standards being taught in the classroom?”, “How much time is spent on different topics?”, and “What types of things are students expected to do to demonstrate the knowledge they are being taught (i.e., cognitive demands)?”

It is unconscionable for school psychologists and other educators to give lip service to access to relevant, rigorous content and equitable opportunity to learn, without taking the opportunity to use tools at our disposal to help us determine the extent to which these situations occur in our schools. By undertaking the development and use of tools that provide information on the alignment among content standards, instruction, and assessments, school psychologists can move from assuming students are getting the chance to learn to truly facilitating those opportunities. Alignment concepts and tools can provide a framework for examining the content of instruction and assessments in an observable, measurable, and reliable manner, resulting in important information for facilitating student achievement.

REFERENCES


Psychology in the Schools DOI: 10.1002/pits