WCER Working Paper No. 2004-1
September 2004

Alignment Analysis and Standard-Setting Procedures for Alternate Assessments

Andrew T. Roach
Department of Educational Psychology/Wisconsin Center for Education Research
University of Wisconsin–Madison
atroach@wisc.edu

Stephen N. Elliott
Department of Special Education/Center for Assessment & Intervention Research
Peabody College
Vanderbilt University
steve.elliott@vanderbilt.edu

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison
http://www.wcer.wisc.edu/

The research reported in this paper was supported by the U.S. Department of Education, Office of Special Education and Rehabilitative Services, in a grant awarded to Jeffery P. Braden and Stephen N. Elliott (“Assessing One and All: An Internet Hypermedia Model for Professional Development”; Award No. H325N000064), and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

Alignment Analysis and Standard-Setting Procedures for Alternate Assessments

Andrew T. Roach and Stephen N. Elliott[1]

The purpose of this paper is to provide an overview of the methods and results of the alignment analyses and standard-setting procedures used in investigations conducted during the development of alternate assessments in two states: Wisconsin and Idaho. Although the methods used to conduct the alignment analyses and standard setting are not unique, this paper will provide insights into their application to states' alternate assessments for students with significant disabilities.

Context

The Individuals With Disabilities Education Act Amendments of 1997 (IDEA '97; 1997) clearly mandate that students with disabilities have access to the general education curriculum and instruction meeting academic standards. Specifically, one of the final regulations under IDEA '97 (34 C.F.R. § 300.347) requires that (a) all students participate in state and district-wide assessments; and (b) all students have opportunities and instruction that allow them to make progress toward state and district academic standards. Moreover, recent interpretations of the No Child Left Behind Act (NCLB; 2002) provisions requiring states to make "adequate yearly progress" (AYP) allow states to use alternate assessments to measure the proficiency of up to 1% of their students (i.e., those students with the most significant disabilities).

In many states, policymakers and practitioners are struggling to meet these requirements because (a) the skills and concepts in the state academic standards have been deemed inappropriate or irrelevant for students with significant disabilities; (b) definitions of *proficient performance* for students with significant disabilities are lacking; and (c) the development of the alternate assessment has generally been considered a special education function, precluding the involvement of general education curriculum and measurement experts.

To provide evidence of the validity of alternate assessment results as indices of students' proficiency in the academic concepts and skills outlined in the states' academic standards, policymakers and test developers must conduct two types of investigations:

1. *Alignment analyses*, which establish the alignment between curricular expectations and the assessment instrument; and

2. *Standard-setting procedures*, which determine cut scores corresponding to specified levels of performance.

[1] The authors wish to express their appreciation to Sandra Berndt and Elizabeth Compton for their assistance in completing this research.

Method

Alignment Analyses

Effective schooling is based on the coordination of three components of the educational environment: curriculum, instruction, and assessment (Elliott, Braden, & White, 2001; Webb, 1997; Webb, Horton, & O'Neal, 2002). The process of coordinating these elements—called *alignment*—is the foundation of standards-based education reform. 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 et al., 2002, p. 1). The development and implementation of large-scale assessment programs represent one approach to aligning classroom instruction with state curriculum standards.

Both Wisconsin and Idaho convened expert panels to determine whether their alternate assessments adequately measured the skills and concepts represented in their state content standards. The panel members (a) rated content standard objectives and individual alternate assessment items for depth of knowledge and (b) identified one or two content standard objectives (at each evaluated grade level) corresponding to each alternate assessment item.

The Wisconsin alignment panel ($N = 10$) consisted of special education teachers, personnel from the Wisconsin Department of Public Instruction, and graduate students who participated in a 2-day Wisconsin Alternate Assessment (WAA) Alignment Institute conducted June 13–14, 2002, at the University of Wisconsin–Madison. In Idaho, the alignment panel ($N = 11$) consisted of special education teachers from all regions of the state and personnel from the Idaho Department of Public Instruction who participated in a 3-day Idaho Alternate Assessment (IAA) Alignment Study conducted November 12–14, 2003, at the Statehouse Inn in Boise.

Both the Wisconsin and the Idaho studies used a model developed by Norman Webb (1997).[2] Webb's model provides a series of statistics that indicate the degree of alignment between the content in a state's academic standards and the content covered by the state's assessment. The Webb model was also used in previous studies conducted to determine the alignment between Wisconsin and Idaho's academic standards and their general large-scale assessments. Thus, the application of this model to the WAA and IAA was intended to provide policymakers with comparative data on the alignment of two elements of each state's assessment system—that is, both the general large-scale assessments and the alternate assessments.

Webb (1997) outlined three methods for determining the alignment between the policy elements of curriculum, instruction, and assessment systems:

1. *Sequential development* involves creation and acceptance of one policy element, which subsequently serves as a blueprint for the creation of additional policy elements. For

[2] The Council of Chief School Officers (CCSSO) has identified four preferred models for states planning and conducting alignment studies. In addition to the Webb (1997) model, they are (a) the SEC (Surveys of Enacted Curriculum) model, developed by Andrew Porter and John Smithson; (b) the Achieve model, developed by Achieve, Inc.; and (c) the CBE (Council for Basic Education) model (CCSSO, 2002).

example, a state or district might develop academic standards for mathematics that provide guidance for the selection of a new performance-focused mathematics curriculum and the development of performance-based mathematics assessments.

2. *Expert review* involves the convening of a panel of content experts to review the three policy elements and determine the extent of their alignment.

3. *Document analysis* involves the coding and analysis of documents that represent the different policy elements.

By integrating these three methods, test developers and education policymakers can increase the quality of the alignment process (Webb, 1997).

Sequential development, expert review, and document analysis each contributed to the creation and validation of alternate assessment systems in Wisconsin and Idaho. 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 alternate knowledge and skills documents (called *alternate performance indicators* in Wisconsin) based on the major content area domains represented in their standards. For those students participating in an alternate curriculum and therefore an alternate assessment, the alternate knowledge and skills documents served as *downward extensions* of the state's content standards. Expert review and document analysis, conducted according to Webb's (1997) methods for determining alignment between policy elements, were used to complete the WAA and IAA alignment process.

Applying the methods previously used for analyzing the alignment of curriculum standards with large-scale assessments (Webb, 2002; Webb et al., 2002), alignment panels in each state were trained to use a collection of analytical tools and heuristics to rate assessment systems and academic standards on the following criteria:

1. Categorical concurrence;

2. Balance of representation;

3. Range-of-knowledge correspondence; and

4. Depth-of-knowledge consistency.

The first three criteria measure the correspondence between (a) the skills and concepts covered in a state's content and performance standards and (b) the skills and concepts tested by an assessment. *Categorical concurrence* indicates if 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 represented by a standard is the same as, or corresponds to, the span of knowledge represented by an assessment item or activity. *Balance of representation* provides an index of the degree to which one curriculum objective is given more emphasis on the assessment than another. Finally, *depth-of-knowledge consistency* is intended to measure the level of mastery and skill required by the performance standards and assessment items. The depth-of-knowledge criterion indicates whether what is elicited from students on an assessment is as demanding cognitively as what students are expected to know and do as stated in the model academic standards.

The Wisconsin and Idaho alignment panel members rated the correspondence between, and the depth of knowledge required by, their state's alternate assessment items and academic standards. Wisconsin used 4th-grade standards, and Idaho used 1st-, 4th-, 8th-, and 10th-grade standards. Idaho included 1st-grade reading standards because it has a state-wide reading assessment at that grade level; the alignment analysis was not conducted for 1st-grade standards in the other subject areas. The primary role of the panel members was to complete the following three tasks:

1. Rate the depth-of-knowledge level of each objective (i.e., knowledge and skills statement) in the academic standards.

2. Rate the depth-of-knowledge level of each item on the alternate assessment rating scale.

3. Identify the one or two objectives in the standards to which each alternate assessment item corresponded.

Panel members' responses were recorded on a series of coding sheets, which provided columns for (a) rating each WAA or IAA item on the depth-of-knowledge criteria and (b) indicating the objectives at each grade level corresponding to each WAA or IAA item. WAA and IAA items were presented in random order instead of by subject domain as on the actual rating scales

Before completing their ratings, panel members were trained to identify the depth-of knowledge level of objectives from the state's academic standards and alternate assessment items. This training included a review of the four general depth-of-knowledge levels outlined in Table 1. Specific descriptions of the depth-of-knowledge levels for each of the subject domains covered by the WAA and IAA were developed, using as models examples from previous alignment analyses conducted on large-scale and alternate assessments (Webb, 2002; Webb et al., 2002).

In Idaho, alignment panel members reached consensus on the depth-of-knowledge levels for one grade level of objectives in reading, language arts, and mathematics before completing their individual ratings of the IAA items using the criteria in Webb's (1997) alignment model. Working to reach consensus provided an opportunity for discussion of the criteria and "calibration" of panel members' understanding of the depth-of-knowledge rating process (Webb, 2002). For the remaining grade levels, the modal (most frequent) depth-of-knowledge ratings were assigned to each objective. When the panel members' depth-of-knowledge ratings for a particular objective did not clearly identify a mode (i.e., a bimodal distribution), the group leader worked with the panel to reach consensus on the most appropriate depth-of-knowledge rating.

Table 1
Depth-of-Knowledge Levels

Level	Description
Level 1: Recall	Level 1 includes the recall of information such as a fact, definition, term, or simple procedure, as well as performance of a simple algorithm or application of a formula.
Level 2: Skill/concept	Level 2 involves some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions about how to approach a problem or activity. Keywords that distinguish a Level 2 item or task include *classify*, *organize*, *estimate*, *make observations*, *collect and display data*, and *compare data*.
Level 3: Strategic thinking	Level 3 includes items that require reasoning, planning, using evidence, and engaging at a higher level of thinking than items at Levels 1 and 2. In most instances, requiring students to explain their thinking is a Level 3 attribute. Level 3 might also require students to make conjectures or determine a solution to a problem that has multiple correct answers.
Level 4: Extended thinking	Level 4 includes items that require complex reasoning, planning, developing, and thinking, generally for an extended 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.

Note. Adapted from Webb (2002).

In Wisconsin, panel members also reached consensus on the depth-of-knowledge ratings for the objectives (i.e., performance standards) for the reading, language arts, and mathematics scales. Because of time constraints, the panels' most common depth-of-knowledge rating (i.e., the mode) was assigned to the objectives in social studies and science.

Following the "calibration" process, panel members in both states were asked to assign a depth-of-knowledge rating and corresponding objective to each assessment item on a randomly ordered list of WAA or IAA items. Panel members independently rated the depth-of-knowledge levels of individual alternate assessment items with moderate to high consistency. In both states, the average measure of intraclass correlations (Shrout & Fleiss, 1979)—which compared the ratings of alignment panel members—was consistently .80 or higher (Table 2).

Table 2
Reliability of Depth-of-Knowledge Level Ratings of Alternate Assessment Items

Subject domain	State	Number of panel members	Number of items	Alpha
Reading	Wisconsin	10	23	.95
	Idaho	11	12	.84
Language arts	Wisconsin	10	26	.94
	Idaho	11	6	.82
Mathematics	Wisconsin	10	29	.90
	Idaho	11	18	.82
Science	Wisconsin	10	21	.86
	Idaho	NA	NA	NA
Social studies	Wisconsin	10	29	.89
	Idaho	NA	NA	NA

According to Webb (2002), the alignment coding process is not designed to produce exact agreement between members of the expert panel. In fact, variance in ratings "are considered valid differences in opinion that are a result of a lack of clarity in how the objectives were written and/or the robustness of an item that may legitimately correspond to more than one objective" (p. 3).

The alignment analysis completed by the Wisconsin and Idaho panel members provided descriptive statistics for the four criteria used in Webb's alignment model (i.e., categorical concurrence, depth-of-knowledge consistency, range-of-knowledge correspondence, and balance of representation). Webb's criteria for determining alignment between assessments and curricular expectations are outlined in Table 3.

Table 3
Summary of Webb's Alignment Criteria

Criterion	Description
Categorical concurrence	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)
Range-of-knowledge correspondence	At least 50% of the objectives for a standard must correspond to at least one related assessment item (based on the ratings of alignment panel members) in order for the alignment on this criterion to be judged acceptable. 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).
Balance of representation	A balance index score is computed to judge the distribution of assessment items. The balance index "compares the proportion of items for each objective to the proportion if the items were evenly distributed among all possible objectives" (Webb et al., 2002). An index value of .7 or greater indicates that assessment items are distributed among all objectives to an acceptable degree.
Depth-of-knowledge consistency	"For consistency to exist between the assessment and the standard. . . at least 50% of the items corresponding to an objective [must] be at or above the level of knowledge of the objective" (Webb, 2002, p. 4). A test meeting this criterion would need to demand the depth of understanding and mastery of the knowledge and skills covered in the corresponding academic standards.

Note. Adapted from Webb (2002).

By focusing on the alignment between alternate assessments and academic content standards, the Wisconsin and Idaho alignment analyses provided evidence for the content and curricular validity of the WAA and IAA. Specifically, the panel's ratings provided information about the correspondence between alternate assessment items and the academic content standards. The alignment panel's responses were expected to indicate the WAA and IAA generally conformed to Webb's (1997) model for alignment of assessments and curricular expectations. In particular, it was expected that the panel's ratings would indicate that each WAA and IAA subject domain scale met the criteria for categorical concurrence, range of knowledge, and balance of representation.

On the other hand, because alternate assessments are designed for students with the most significant cognitive disabilities, the alignment panel responses were expected to indicate a low overall depth-of-knowledge rating for the WAA and IAA subject domain scales. Such a low overall depth-of-knowledge rating would represent a departure from previous alignment studies using expert panel ratings (Webb, 2002; Webb et al., 2002). Although it is desirable that depth-of-knowledge ratings for curriculum objectives and assessment items be similar, items on alternate assessments are believed to generally demand less depth of knowledge than items in the general education academic standards and on the corresponding large-scale assessment. Thus, even if alternate assessment items represent the range of concepts and skills outlined in state academic standards, these items may be presented at a lower level of complexity or prerequisite skill in order to provide access to the assessment to students with significant disabilities.

Standard Setting

On June 25 and 26, 2003, the WAA standard-setting workshop was conducted to achieve two goals: (a) to set proficiency cut scores for the WAA for students with disabilities and (b) to gather input on terminology and wording changes that would enhance the WAA rating scale and supporting materials. To achieve the same goals for the IAA, a standard-setting workshop was conducted in Idaho on July 14 and 15, 2003.

Standard setting is the process of determining the appropriate cut scores for specified levels of performance. This process requires a determination of (a) the knowledge, skills, and competencies students should be expected to understand and demonstrate at each performance level and (b) the test scores that correspond to those expectations.

The most important outcome of standard setting is not, however, the cut scores associated with proficiency levels in each content area, but rather the descriptors of what students who achieve the various performance levels typically know and are able to perform. By examining these descriptors, one can gain an understanding of the knowledge, skills, and abilities that students at various performance levels typically possess or lack. This type of information helps teachers communicate with others about a student's progress, the next year's instructional goals for the student, and the status of the student relative to the state's learning standards.

Standard setting requires a good deal of judgment, as well as a high level of confidence in that judgment. Thus, it is important to have a representative group of educators familiar with the curricular and instructional needs of students with disabilities and knowledgeable about the current alternate assessment to serve on a standard-setting committee.

There are several different approaches to establishing proficiency standards. In this case, a modified *bookmark procedure* (Lewis, Mitzel, & Green, 1996) was used to help establish the proficiency standards for the WAA and IAA. The bookmark procedure was developed by researchers at CTB/McGraw-Hill and had previously been used to establish the proficiency standards for the Wisconsin and Idaho general large-scale assessments. Standard setting using this procedure involves presenting experienced educators with a booklet presenting a set of test items, ordered from easiest to most difficult, for each content area (i.e., reading, language arts, math, science, and social studies). After carefully studying the ordered items in a booklet, the educators identify a unique cut score for a given performance level by placing a bookmark at the appropriate location in the booklet. Items preceding the bookmark represent content knowledge that all proficient students should be likely to know and demonstrate. The final cut score is computed as the mean of the number of items immediately before and after the bookmark. Although this process sounds quite simple, in fact workshop participants often expend considerable effort in reaching their final decisions about the knowledge, skills, and competencies representing proficiency.

The modified bookmark procedure entails the following steps for each of the content areas in an alternate assessment:

- Introduction to standard setting

- Review of all items on the rating scale

- Review and discussion of the current proficiency descriptors for each performance level

- Achievement of consensus by participants on the definition of proficient as measured by the alternate assessment

- Round 1: Placement of bookmarks in test booklets by each participant to indicate proficiency cut scores

- Post-Round 1: Discussion of cut scores by participants at each table

- Round 2: Achievement of consensus by each table team on bookmarks for the proficient levels of performance

- Post-Round 2: Feedback about the mean cut scores and the likely distribution of students at each level

- Round 3: Final team decisions about bookmarks for each of four levels of performance

- Post-Round 3: Feedback about the committee’s mean cut scores and their likely impact on student distributions

- Review and revision, if necessary, of the proficiency descriptors associated with each of the four levels of performance

The key materials used to conduct the standard setting were (a) a series of tables with the WAA or IAA items from each content area rank-ordered by difficulty, from easiest to hardest, and (b) a series of graphs portraying the total score distributions of students who were administered the alternate assessments. It should be noted that the student samples in the WAA and IAA research databases were rather small and may not have been representative of the entire population of students with severe disabilities. Thus, workshop participants were instructed to use caution in interpreting the graphical data on the impact of the cut-score decisions.

After reading the consensus definition of *proficient* for a content area such as reading, workshop participants used the tables of rank-ordered items to record their decisions about what knowledge and skills, as measured by the alternate assessment items, should be considered necessary to qualify as proficient. As indicated above, participants first made independent decisions, then worked with their table mates (4 or 5 other participants) to reach consensus on the number of items necessary to qualify as proficient. Once all table leaders had reported their tables' proficiency cut scores, participants were provided with graphs displaying the impact data—that is, the percentage of students in the research database likely to be considered proficient in a content area. After reaching a final decision about the cut scores for each level of performance within an area, participants were also provided comparative data for the states' general large-scale assessments.

The three-round modified bookmark procedure was followed for each of the content areas on both states' alternate assessments. This procedure resulted in cut scores and refined performance-level descriptions for each content area on the alternate assessments.

Results

The WAA and IAA alignment panels' responses indicate that these alternate assessments are generally well aligned with the skills and knowledge represented by the Wisconsin and Idaho academic standards. In fact, the performance of these alternate assessments on the four criteria making up Webb's (1997) alignment model met or exceeded the performance of many states' general education assessments. By comparison, in a recent alignment analysis of alternate knowledge and skills documents from 42 states, 60% of the special education experts surveyed indicated that most states had not adequately assessed the general education curriculum standards with their alternate assessments (Browder et al., 2002).

Alignment Analyses

Wisconsin

The results of the WAA alignment analysis indicate that the WAA language arts and science scales achieved categorical concurrence for less than 50% of academic standards (Table 4). Although this result is less than optimal, it is important to emphasize that meeting the categorical concurrence criterion only indicates that there are sufficient items to create subscales within a particular academic area. Because the WAA reports only total scale scores for each subject domain, meeting this criterion was desirable but not necessary to determine the validity and usability of the assessment.

The range-of-knowledge criterion was met for the reading and language arts scales. According to the panel members' ratings, 100% of the reading and language arts objectives had a corresponding WAA item. The range-of-knowledge criterion was also met for the mathematics, social studies, and science scales, although the panel members' ratings indicate the WAA items only weakly met the criterion for the majority of standards. This result is attributable to the numerous academic standards for these subject domains and the relative brevity of the WAA subject domain scales. For example, the low levels of range-of-knowledge consistency between some social studies standards and the WAA social studies scale reflect the numerous objectives for those standards. Although the panel members' ratings indicated multiple items on the WAA social studies scale corresponded to the standards, the range of item hits was not expansive enough to strongly meet the range-of-knowledge criterion.

Table 4
Alignment Indices for WAA Subject Domain Scales

Subject domain	No. of academic standards	Categorical concurrence (% of academic standards acceptable*)	Range of knowledge (% of academic standards acceptable*)	Balance of representation (% of academic standards acceptable)	Depth of knowledge (% of academic standards acceptable*)
Reading	1	100%	100%	100%	100%
Language arts	5	40%	100%	100%	40%
Mathematics	6	50%	66%	100%	83%
Science	8	13%	75%	100%	75%
Social studies	5	60%	80%	100%	80%

*Includes standards with weak categorical concurrence, range of knowledge, and depth of knowledge.

The balance of representation was rated as acceptable for all the subject domain scales. This result is attributable to the concise format of the WAA rating scale in comparison to many individually administered standardized tests. The limited number of items for each subject domain scale demanded that the scale developers evenly distribute items among the objectives. The panel members' ratings confirmed that the item development process resulted in a well-balanced scale for assessing students' performance.

The WAA rating scale was not expected to demonstrate acceptable depth-of-knowledge consistency using Webb's alignment procedures; in fact, meeting the depth-of-knowledge criterion could be considered an indication that some WAA items were too difficult for the population of students for whom the test was developed. The results of the WAA Alignment Institute, however, indicated a generally acceptable level of depth-of-knowledge consistency between the WAA and the majority of academic standards in reading, mathematics, and social studies. There are multiple plausible explanations for this unexpected result: (a) the wording of the WAA items is general enough to allow for more complex interpretations of the tasks; (b) panel members felt that the items tapped the same skills and knowledge expected in the objectives in a way that made them accessible to students with severe disabilities; and (c) the skills and concepts expected in the state's academic standards primarily focus on recall and simple application of knowledge.

Idaho

Analysis of the results from the IAA Alignment Institute indicates reading/receptive communication and language/expressive communication standards achieved an acceptable level of categorical concurrence (Table 5). Standards in these areas were judged by panel members to have at least six corresponding items on the IAA scale.

The IAA mathematics scale achieved categorical concurrence for approximately 50% of mathematics standards at each grade level. Although this result is less than optimal, as noted earlier attaining the categorical concurrence criterion only indicates that there are sufficient items to create subscales within a particular academic area. Because the IAA reports only total scale scores for the overall subject domain of mathematics, meeting this criterion was desirable but not necessary to determine the validity and usability of the assessment.

Table 5
Alignment Indices for IAA Subject Domain Scales

Subject domain	Grade	No. of academic standards	Categorical concurrence (% of academic standards acceptable*)	Range of knowledge (% of academic standards acceptable**)**	Balance of representation (% of academic standards acceptable)	Depth of knowledge (% of academic standards acceptable*)
Reading/ receptive communication	1 4 8 10	3 3 3 3	100% 100% 100% 100%	67% 67% 67% 67%	100% 100% 100% 100%	100% 100% 100% 100%
Language/ expressive communication	4 8 10	2 2 2	100% 100% 100%	100% 50% 50%	100% 100% 100%	100% 100% 100%
Mathematics	4 8 10	7 7 7	57% 43% 43%	57% 43% 43%	100% 100% 100%	86% 86% 71%

*Includes standards with weak categorical concurrence, range of knowledge, and depth of knowledge.

The results of the IAA Alignment Institute indicate the range-of-knowledge criterion was partially met for the reading/receptive communication scale. According to the panel members' ratings, the IAA consistently met the range-of-knowledge criterion for listening and viewing standards at all four grade levels (1st, 4th, 8th, and 10th). The range-of-knowledge criterion was not met for the reading standard at any grade level. The percentage of reading objectives that had a corresponding IAA item ranged from 27% (1st grade) to 34% (8th grade). It is important to note, however, that across all four grade levels, approximately twice as many reading objectives as listening and viewing objectives were "hit" (i.e., rated as corresponding with an IAA item).

The results on the range-of-knowledge criterion for the IAA reading/receptive communication scale are most likely attributable to two factors. First, Idaho's academic standards contain an exceptionally large number of objectives for each of the reading standards. For example, the first-grade reading standards include 25 objectives for Performance Standard 680.1 (*Read a variety of traditional and electronic materials for information and understanding*). Because of the relative brevity of the IAA reading/receptive communication scale, there are simply not enough items to achieve the range-of-knowledge criterion for such an expansive number of objectives. Moreover, many of the IAA reading/receptive communication items focus on pre-reading and receptive communication skills that may be more clearly aligned to objectives in the listening and viewing standards at each grade level than to the objectives in the reading standards.

The results also indicate the range-of-knowledge criterion was partially met for the language/expressive communication scale. Although the range-of-knowledge criterion was met for the 4th-grade writing standard, it was not met for the writing standard at the 8th- and 10th-grade levels. This result is attributable to the additional goals and the additional corresponding writing objectives at those grades (seven goals at 8th and 10th grade vs. three at 4th grade). Although the panel members' ratings indicated that IAA items corresponded with between six and seven writing objectives at these grade levels, the range of item "hits" was not expansive enough (23% at 8th grade and 31% at 10th grade) to meet the range-of-knowledge criterion.

The range-of-knowledge criterion was only partially met for mathematics standards at each grade level. The basic arithmetic, estimation, and accurate computation standard received the most hits and was the only standard with consistently acceptable range-of-knowledge across all rated grade levels. This result may reflect a perception on the part of alignment panel members that many IAA mathematics items focus on the basic arithmetic knowledge and skills. The range-of-knowledge criterion was not met for the algebra, geometry, and statistics, probability, and data analysis standards at any grade level. This result may reflect the difficulty of writing a variety of accessible IAA items that reflect the complexity and range of skills and knowledge in these areas.

The balance of representation between the IAA subject domain scales and the Idaho standards was rated as acceptable across all subjects and grade levels. This result is attributable to the concise format of the IAA rating scale in comparison to many individually administered standardized tests. The limited number of items for each subject domain scale demanded that the assessment developers evenly distribute items among the objectives. The panel members' ratings confirmed that the item development process resulted in a well-balanced scale for assessing students' performance.

As noted earlier, although it is generally desirable that depth-of-knowledge ratings for curriculum objectives and assessment items be similar, many alternate assessments items demand less depth of knowledge than items on the general education academic standards and on the corresponding large-scale assessment. IAA rating scale items represent the range of concepts and skills outlined in Idaho's academic standards, but these items are presented at a lower level of complexity that allows access for students with significant disabilities. Therefore, the IAA was not expected to demonstrate acceptable depth-of-knowledge consistency. The acceptance of a low overall depth-of-knowledge rating would represent a departure from previous alignment studies using expert panel ratings (Webb, 2002; Webb et al., 2002).

The results of the IAA Alignment Institute, however, indicate a generally acceptable level of depth-of-knowledge consistency for the reading/receptive communication and language/expressive communication scales at all grade levels (see Table 5). This result may be attributable to (a) the inclusion of a variety of performance indicators for each IAA item, allowing for more sophisticated performance on the part of students; and (b) the wording and content of Idaho's academic standards, which raters may have perceived as generally focusing on skills and application rather than more extended problem solving.

The results also indicate a generally acceptable level of depth-of-knowledge consistency for the mathematics scale. The depth-of-knowledge consistency was rated weak to unacceptable for the geometry standards across grade levels, perhaps suggesting the need to include more sophisticated and complex performance indicators for IAA items that address geometry objectives.

Standard Setting

Participants in the standard-setting process in both Wisconsin and Idaho were able to reach consensus on definitions of proficient performance for students with significant disabilities. Using these definitions and the standard-setting procedure described earlier, participants in both states developed cut scores to indicate levels of performance on the alternate assessments corresponding with proficiency. Examination of the impact of these cut scores showed that the percentage of students judged proficient as measured by each state's alternate assessment was similar to the percentage of the general student population judged proficient as measured by each state's large-scale assessment.

Wisconsin

The results of the 2-day WAA standard-setting workshop are displayed in Tables 6 and 7, providing a summary of the key outcomes with regard to cut scores, impact, and labels for the various performance levels. Table 6 provides an integrated summary of the cut scores and impact data for the WAA in reading, language arts, mathematics, science, and social studies.

Table 6

Summary of Recommended WAA Proficiency Standard Cut Scores

Content area	Final-round cut scores and percentage student impact estimate			
	PS[a] Minimal	PS Basic	PS Proficient	PS Advanced
Reading (score range 0–69)	0–4 8.9 %	5–20 23.6%	21–52 50.4 %	53–69 17.1%
Language arts (score range 0–78)	0–5 4.6%	6–23 18.5%	24–54 59.3%	55–78 17.6%
Mathematics (score range 0–87)	0–3 9%	4–21 26%	22–64 50%	65–87 15%
Science (score range 0–72)	0–4 21%	5–20 32%	21–53 44%	54–72 3%
Social studies (score range 0–87)	0–4 10%	5–25 30%	26–67 55%	68–87 5%

[a]PS = Prerequisite skill.

Table 7 provides a simplification of the impact that the cut scores will have on the percentage of students declared to be proficient on the WAA. Students performing at either the *prerequisite skill (PS) proficient* or the *PS advanced* level are considered "proficient as measured by the WAA." Table 7 illustrates similar data for the Wisconsin Knowledge and Concepts Examinations (WKCE), based on the 2003 WKCE results. Using the percentage of students who met proficiency as an indicator of rigor, the WAA is as rigorous as—or perhaps more rigorous than—the performance standards for the general education assessment.

Table 7

Comparison of the Percentage of Students Above the Proficient-Level Cut Score on the WAA to the Percentage of Students Above the Proficient-Level Cut Score on the WKCE

Content area	% students above proficient-level cut score on WAA	% students above proficient-level cut score on WKCE '03
Reading	67.5%	80%
Language arts	76.9%	78%
Mathematics	65%	71%
Science	47%	77%
Social studies	60%	90%

Idaho

The Idaho standard-setting workshop focused on Grades 4, 8, and 10—Idaho's "benchmark" grades for meeting the federal requirement for statewide assessment at the elementary, middle, and high school levels. Tables 8 and 9 provide a summary of the main outcomes with regard to cut scores, impact, and labels for the various performance levels. Table 8 provides an integrated summary of the cut scores for the IAA in reading, language arts, and mathematics. The scores reported in Table 8 are raw total scores for the knowledge and skills items assessed on each of the IAA content area scales. Given that the number of items varies for the reading (12 items), language arts (6 items), and mathematics (18 items) scales, the possible range of total scores differs for each scale. The possible score range for reading is 12–192; for language arts, it is 6–96; and for mathematics, it is 18–288.

Table 8
Proficiency Score Ranges for the IAA Proficiency Levels

	Reading	*Language arts*	*Mathematics*
Advanced IAA			
4th	126–192	54–96	147–288
8th	126–192	63–96	183–288
10th	132–192	75–96	195–288
Proficient IAA			
4th	68–125	30–53	58–146
8th	68–125	30–62	82–182
10th	76–131	38–74	106–194
Basic IAA			
4th	24–67	11–29	28–57
8th	25–67	13–29	32–81
10th	26–75	16–37	36–105
Below Basic IAA			
4th	12–23	6–10	18–27
8th	12–24	6–12	18–31
10th	12–25	6–15	18–35

Table 9 provides a simplification of the impact that the cut scores will have on the percentage of students declared to be proficient on the IAA. Students performing at either the *proficient IAA* or the *advanced IAA* level are considered to be "proficient as measured by the IAA" for federal accountability reports. Table 9 illustrates similar proficiency-level impact data for the Idaho Standards Achievement Test (ISAT), based on 2003 test results for 4th, 8th, and 10th grades across the entire state. Using the percentage of students who met proficiency as an indicator of rigor, the IAA is as rigorous as—or perhaps more rigorous than—the performance standards for the general education assessment.

Table 9
Comparison of the Percentage of Students Above the Proficient-Level Cut Score on the IAA to the Percentage of Students Above the Proficient-Level Cut Score on the ISAT

Content area	% students above proficient-level cut score on the IAA	% students above proficient-level cut score on the ISAT '03
Reading		
4th	61%	75%
8th	56%	74%
10th	68%	75%
Language arts		
4th	60%	80%
8th	55%	71%
10th	61%	74%
Mathematics		
4th	61%	77%
8th	50%	53%
10th	61%	72%

Conclusions

This paper provides an overview of the initial applications of a nationally recognized alignment procedure (Webb, 2002) and a modification of a widely used standard-setting procedure (Lewis et al., 1996) to two states' alternate assessments. The results suggest that Webb's alignment model and a modified bookmarking procedure can be meaningfully applied to alternate assessments, providing educators and policymakers with a tool for gathering evidence on the validity of their states' inclusive assessment and accountability systems.

Moreover, use of these strategies can provide evidence of states' compliance with the requirements for alternate assessments outlined in IDEA '97 and the NCLB. Results of alignment analyses can be used to establish the correspondence between states' content standards and assessment instruments, and standard-setting procedures can assist policymakers in determining cut scores that correspond to specified levels of performance.

According to data collected by the National Center on Educational Outcomes (NCEO), many students with disabilities traditionally have been excluded from state and district-wide assessment and accountability systems (Ysseldyke et al., 1998). The exclusion of these students has been unfortunate because it is impossible to measure the overall effectiveness of instructional and school reform efforts without considering the performance of *all* students. Recent federal legislation (IDEA '97; NCLB) addresses this situation by requiring the inclusion of all students, including those with the most significant disabilities, in the assessment and reporting of school-, district-, and state-level performance.

Although the development and implementation of standards-focused alternate assessments represent a promising strategy for increasing the inclusion and achievement of students with significant disabilities, they are not without challenges or risks. Ysseldyke, Thurlow, and Shin (1995) suggested that higher standards and accountability can be viewed as both an opportunity and a burden for students with disabilities. If students with disabilities are not afforded the same (or perhaps greater) opportunities to learn and demonstrate their proficiency on the skills and concepts tested on alternate assessments, they can easily become scapegoats for schools', districts', and states' inability to reach NCLB mandates for adequate yearly progress. As Browder, Fallin, Davis, and Karvonen (2003) suggested in a recent review of research on alternate assessments, additional research is needed on the student, teacher, classroom, and system variables that influence alternate assessment performance, so that students with severe disabilities are not "left behind."

References

Browder, D., Flowers, C., Ahlgrim-Delzill, L., Karvonen, M., Spooner, F., & Algozzine, R. (2002, April). *Curricular implications of alternate assessments.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Browder, D. M., Fallin, K., Davis, S., & Karvonen, M. (2003). Consideration of what may influence student outcomes on alternate assessment. *Education and Training in Developmental Disabilities, 38,* 255–270.

Council of Chief State School Officers. (2002). *Models for alignment analysis and assistance to states.* Retrieved September 15, 2004, from http://www.ccsso.org/content/pdfs/AlignmentModels.pdf

Elliott, S. N., Braden, J. B., & White, J. L. (2001). *Assessing one and all: Educational accountability for students with disabilities.* Arlington, VA: Council for Exceptional Children.

Individuals With Disabilities Education Act Amendments of 1997, Pub. L. No. 105-17, 111 Stat. 37 (codified as amended at 20 U.S.C. § 1400 *et seq*.).

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures utilizing behavioral anchoring*. Symposium conducted at the meeting of the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Shrout, P. E., & Fliess, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison: University of Wisconsin–Madison, National Institute for Science Education. Retrieved September 17, 2004, from http://www.wcer.wisc.edu/nise/Publications/Research_Monographs/WEBB/WebbALL.doc

Webb, N. L. (2002, April). *An analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved September 15, 2004, from http://facstaff.wcer.wisc.edu/normw/AERA%202002/Alignment%20Analysis%20three%20states%20Math%20Final%2031502.pdf

Webb, N. L., Horton, M., & O'Neal, S. (2002, April). *An analysis of the alignment between language arts standards and assessments for four states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. Retrieved

September 15, 2004, from http://facstaff.wcer.wisc.edu/normw/AERA%202002/Alignment%20Analysis%20Language%20Arts%20%20Four%20States%2031202.pdf

Ysseldyke, J. E., Krentz, J., Elliott, J., Thurlow, M., Erickson, R., & Moore, M. (1998). *NCEO framework for educational accountability*. Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes. Retrieved September 17, 2004, from http://education.umn.edu/nceo/OnlinePubs/Framework/FrameworkText.html

Ysseldyke, J., Thurlow, M., & Shin, H. (1995). *Opportunity-to-learn standards* (Policy Directions No. 4). Minneapolis, MN: National Center on Educational Outcomes. Retrieved September 17, 2004, from http://education.umn.edu/NCEO/OnlinePubs/Policy4.html