Setting Alternate Achievement Standards

Marianne Perie
National Center for the Improvement of Educational Assessment, Inc.
Draft: June 18, 2007

EXECUTIVE SUMMARY

Under the No Child Left Behind (NCLB) Act of 2001, states may count as Proficient up to one percent of students with significant cognitive disabilities using alternate achievement standards. However, states have struggled with the process of developing valid alternate achievement standards. The process involves deciding on an appropriate number and name for the levels, writing cohesive performance level descriptors, and choosing an appropriate methodology to set cut scores. This report discusses all of these steps and also provides guidance on how to run a sound standard-setting study. It focuses on the importance of writing detailed performance level descriptors aligned with both performance expectations and grade-level content standards while incorporating ideas specific to this population such as types of supports and contexts. It describes different methods available for setting cut scores and provides considerations for choosing methods. Finally, it includes suggestions for validity studies that can be done and details the sections that should be included in the technical documentation for alternate achievement standards. The end result of this process should be coherent alternate achievement standards consisting of performance level descriptors, cut scores, and exemplar items or sample student work for each performance level.

OVERVIEW

Under federal regulations, states may develop alternate achievement standards for students with the most significant cognitive disabilities. The *No Child Left Behind* Act (NCLB) limits the number of students that may be classified as Proficient for AYP calculations using alternate achievement standards to one percent of the population tested. The goal is to hold all students—including this one percent—to high standards. The difficulty lies in determining how high is appropriately challenging but not out of reach. As stated by Thurlow & Ysseldyke (2001) in their discussion on holding students with disabilities to the same standard of other students, "it is generally not a discussion of what content standards they are working toward, but more often a discussion of what level of performance they must meet to pass[1]." (p. 403). Since that paper was written, the discussion has moved from "passing" to developing proficiency. For too long, the field has said that less rigorous methodologies were "good enough" for this population. Now we recognize that we must find a better balance between the flexibility required of alternate assessments with a similarly high level of technical rigor required of general assessments.

[1] Note that this quote pre-dates NCLB. Although "pass" is not the same as proficient, we could substitute the term "reach Proficient" for pass and the point would remain the same.

The theory behind setting achievement standards remains the same for alternate assessment as it is for general assessment. Methods that are used to set cut scores on alternate assessments are little different from the methods used in general assessment. Both types of assessments should have well-written performance level descriptors. Yet, there are challenges to developing "alternate achievement standards" that do not exist when developing achievement standards for the general assessments. There are relatively few students participating in the assessment and the students are arguably more diverse. The assessments have greater flexibility built into them. Knowledge and skills typically are assessed within a context of independence and generalizability. Yet, the same level of rigor and standardization in the procedures is required to set valid alternate achievement standards for this 1 percent as to develop achievement standards for the remaining 99 percent.

Achievement Standards

The term "standards" has many meanings in educational discussions. We have the standards for educational and psychological testing, often called simply "the standards" And, every state must develop both content standards (also called academic content standards) and achievement standards (also called performance standards). Often these latter two types of standards are confused and the labels are misused. Content standards are statements of the knowledge and skills that students are expected to learn; achievement standards include both a minimum cutoff score and a written description to distinguish among different levels of performance. In other words, content standards are the what; achievement standards are the how much or how well.

An achievement standard typically is defined as the minimally adequate level of performance for some purpose (e.g., Kane 1994) or the level of performance that is expected of examinees (Hambleton, 2001). Under NCLB, the purpose of achievement standards is as a benchmark for evaluating school quality or effectiveness. The achievement standards typically have few consequences for students, but in the aggregate may have significant consequences for a school.

An achievement standard consists of three components: the name of the level, a written description of the level, and a minimum cutoff score. A fourth component, exemplar items or sample student work at each level, is optional but very helpful.

Similarities and Differences between Setting Achievement Standards on Alternate and General Assessments

In their paper on flexibility versus standardization in alternate assessments, Gong & Marion (2006) describe some of the differences between achievement standards[2] on the general and alternate assessments as follows:

[2] Note that these authors use the term "performance standards" rather than "achievement standards." These terms are often used interchangeably.

General assessments: All students at a particular grade level are expected to be held to the same performance/achievement standards. This almost always involves some type of standard setting process enabling raw scores (or the scale score equivalent) to be converted into performance categories. Within any given year, the only variability in how the performance standards are applied to students in the general assessment is simply a function of measurement error. Across years, additional variability in how the performance standards are applied may be introduced as a result of equating error.

Alternate assessments for students with significant cognitive disabilities: States are permitted to introduce systematic variability into the performance standards for alternate assessments because they are permitted under NCLB to establish multiple performance standards for alternate assessments. Most states employ a single set of alternate assessment performance standards at each grade level and evaluate all alternate assessment scores against these standards. This might be a case where more variability may make more educational sense, but most states have chosen to employ a common single set of standards.

Even though there is greater flexibility in setting achievement standards on the alternate assessment than on the general assessment, the basic principles for developing achievement standards remain the same for any assessment. Consider, for example, the professional standards for educational and psychological testing laid out by the American Educational Resaerch Association, the American Psychological Association, and the National Council on Measurement in Education (1999):

> *When cut scores defining pass-fail or proficiency categories are based on direct judgments about the adequacy of item or test performances or performance levels, the judgmental process should be designed so that judges can bring their knowledge and experience to bear in a reasonable way. (Standard 4.21)*
>
> *When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria. (Standard 4.20)*
>
> *Document the PLDs, selection and qualification of panelists, training provided, ratings, and variance measures (Standard 1.7, paraphrased)*
>
> *When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented. (Standard 4.19)*

Moreover, many of the steps remain the same across both assessment types. The following list shows a set of standard tasks for developing achievement standards common to all assessments

1. Write distinct, detailed performance level descriptors (PLDs) that link directly to the content standards
2. Use a documented and validated standard setting methodology
3. Convene a panel of educators and stakeholders who are familiar with the population of students tested, know the content, and represent the diversity of the state
4. Train them in the content standards, PLDs, and methodology
5. Collect individual judgments and use group discussion across rounds
6. Provide normative feedback to the committee to show the relationship between the cut scores and results
7. Aggregate judgments to determine final panel-recommended cut scores
8. Convene appropriate policymaker(s), apprise them of the recommendations, and have them legally adopt the cut scores
9. Document the PLDs, rationale and procedures for the methodology, selection of panelists, training provided, ratings, and variance measures

However, there are some aspects of developing alternate achievement standards that are specific to the alternate assessment. In part, this is due to the population. The federal government has limited the number of scores from the alternate assessment that can count towards a state's AYP Proficient targets to one percent of the population. Because the population is so small, scores may have a wider variance than those from the general assessment and there will be fewer examples of student work at each possible score point.

In some alternate assessment designs, states have allowed flexibility in choosing which standards or indicators the students will be assessed on. For example, in some portfolio approaches the state may list two indicators on which students are required to produce evidence and then ask the student to produce evidence for two more indicators from a list of eight possible indicators. Under this scenario, not only will there be fewer samples of student work at each possible score point, but those samples may represent different content.

Another difference is in the regulations that allow performance level descriptors to be written to a grade span rather than a grade level if the adjoining grades are sufficiently similar. That is, states can set one cut score for grades 3–4, another for grades 5–6 and another for grades 7–8. So, while there will be six Proficient cut scores for the general assessment (one each in grades 3–8) there would be only three cut scores for the alternate assessment. However, states are still required to show progression from one grade to the next within a grade band.

Finally, in all assessments, performance level descriptors are written to represent the entire range of performance within a level. For alternate assessments, the regulations

specify that these descriptors must represent the highest standard possible for this population. More so than with any other population, the emphasis with the alternate assessment for students with significant cognitive disabilities is to set the bar high, encourage teachers to teach grade-level material, and require strong academic performance to reach proficiency.

Current State Practice

The design of the assessment has major implications for developing alternate achievement standards. Different types of assessment will lend themselves to more content-specific descriptors of performance or more general applications, depending on the link between the assessment and the content standards. For example, some portfolio assessments require students to produce sample work for only two of five content standards. This design presents challenges to those writing the performance level descriptors. Likewise, the method appropriate for setting cut scores on the alternate assessment is very dependent on the assessment design. Determining whether a holistic approach, a profile approach, or an item-based approach is more appropriate will depend heavily on the format of the assessment. Consider, for example, how a portfolio assessment would require a different approach to setting cut scores than a skills checklist.

Table 1 provides a brief summary of some of the more common types of assessments used in the states. It is important to note, however, that the lines between some of these assessment types have become blurred. For instance, what one state calls a portfolio assessment may actually look more like a structured task assessment. It is important to examine the elements of the assessment thoroughly and not rely simply on its name when determining appropriate methods for setting cut scores on these alternate assessments.

Table 1. Definitions of Typical Alternate Assessment Approaches

Portfolio: A collection of student work gathered to demonstrate student performance on specific skills and knowledge, generally linked to state content standards. Portfolio contents are individualized, and may include wide ranging samples of student learning, including but not limited to actual student work, observations recorded by multiple persons on multiple occasions, test results, record reviews, or even video or audio records of student performance. The portfolio contents are scored according to predefined scoring criteria, usually through application of a scoring rubric to the varying samples of student work.
IEP Linked Body of Evidence: Similar to a portfolio approach, this is a collection of student work demonstrating student achievement on standards-based IEP goals and objectives measured against predetermined scoring criteria. This approach is similar to portfolio assessment, but may contain more focused or fewer pieces of evidence given there is generally additional IEP documentation to support scoring processes. This evidence may meet dual purposes of documentation of IEP progress and the purpose of assessment.
Performance Assessment: Direct measures of student skills or knowledge, usually in a one-on-one assessment. These can be highly structured, requiring a teacher or test administrator to give students specific items or tasks similar to pencil/paper traditional tests, or it can be a more flexible item or task that can be adjusted based on student needs. For example, the teacher and the student may work through an assessment that uses manipulatives, and the teacher observes whether the student is able to perform the assigned tasks. Generally the performance assessments used with students with significant cognitive disabilities are scored on the level of independence the student requires to respond and on the student's ability to generalize the skills, and not simply on accuracy of response. Thus, a scoring rubric is generally used to score responses similar to portfolio or body of evidence scoring.
Checklist: Lists of skills, reviewed by persons familiar with a student who observe or recall whether students are able to perform the skills and to what level. Scores reported are usually the number of skills that the student is able to successfully perform, and settings and purposes where the skill was observed.
Traditional (pencil/paper or computer) test: Traditionally constructed items requiring student responses, typically with a correct and incorrect forced-choice answer format. These can be completed independently by groups of students with teacher supervision, or they can be administered in one-on-one assessments with teacher recording of answers.

As published in Quenemoen, Thompson, & Thurlow (2003)

The National Center on Educational Outcomes (NCEO) has been conducting biannual surveys of state alternate assessment and accommodation practices since 1999. Table 2 shows the number of states using each type of alternate assessment between 1999 and 2005.[3]

Approximately half the states used a portfolio or body of evidence approach each year. However, the degree of standardization or flexibility in these approaches varies greatly among states. Some states allow teachers and IEP teams free rein over choosing which standards to assess. Others are very prescriptive in describing what evidence must be produced. Still others provide a range of indicators or tasks from which the teacher can use, and others combine a required standard or indicator with a choice standard or indicator (Gong & Marion, 2006).

[3] In 2007, the portion of the survey that examined state practices in alternate assessment was transferred to the National Alternate Assessment Center (NAAC). Those results are forthcoming.

Table 2. Alternate Assessment Approaches 2000-2005

Year	Portfolio or Body of Evidence	Rating Scale or Checklist	IEP Analysis	Other	In Development/ Revision
Regular States					
1999	28 (56%)	4 (8%)	5 (10%)	6 (12%)	7 (14%)
2001	24 (48%)	9 (18%)	3 (6%)	12 (24%)	2 (4%)
2003	23 (46%)	15 (30%)	4 (8%)	5 (10%)	3 (6%)
2005*	25 (50%)**	7 (14%)***	2 (4%)	7 (14%)	8 (16%)
Unique States					
2003	4 (44%)	0 (0%)	1 (11%)	1 (11%)	3 (33%)
2005	1 (11%)	1 (11%)	1 (11%)	0 (0%)	1 (11%)

*One state has not developed any statewide alternate assessment approaches.

**Of these 25 states, 13 use a standardized set of performance/events/tasks/skills.

***Of these 7 states, 3 require the submission of student work.

Unique states include the District of Columbia, Guam, Puerto Rico, American Samoa, Mariana Islands, and the Virgin Islands.

Eight states reported in 2005 that they were redesigning their alternate assessments (Thompson, et al., 2005). In fact, a large number of states have had to either revisit or redesign their alternate assessments in light of the NCLB requirements. According to the U.S. Department of Education, "An alternate assessment must be aligned with the State's content standards, must yield results separately in both reading/language arts and mathematics, and must be designed and implemented in a manner that supports use of the results as an indicator of AYP (adequate yearly progress)" (USED, 2004, p. 15).

Both the alignment requirement and the need to report results separately in reading and math has resulted in new alternate assessment designs in many states. Many states needed to constrain their portfolio assessments to better link them on grade-level academic content by providing teachers with more specific criteria to follow. Some checklist approaches were not approved by USED peer reviewers and were abandoned by states for more technically rigorous task-based approaches. Overall, it appears that task-based performance assessments are gaining popularity. Portfolio assessments remain quite popular even with the new constraints. Simple IEP analyses are no longer permitted as a form of alternate assessment. There has also been a big movement towards hybrid approaches. Checklists now have supporting bodies of evidence. Portfolios have some common tasks. Performance assessments have more structured tasks, some even use multiple-choice items scored using a rubric based on the amount of support a student needs to answer the question correctly. These changes in assessment type have direct implications for the type of standard-setting method that will be appropriate for setting cut scores.

Once the assessments have been developed or redesigned, the achievement levels must be set. Here, too, the federal government has provided guidelines. The federal regulation on alternate achievement standards §200.1 states: "For students... with the most significant cognitive disabilities, who take an alternate assessment, a State may,

through a documented and validated standards-setting process, define alternate achievement standards…"

As taken from the regulations, Peer Review Guidance Critical Element 2.1 states that "[States may] define alternate academic achievement standards, provided those standards (1) are aligned with the State's academic content standards; (2) promote access to the general curriculum; and (3) reflect professional judgment of the highest achievement standards possible."

Three primary considerations arising from these statements about defining alternate achievement standards are (1) developing appropriate performance level descriptors (PLDs), (2) conducting a formal standard-setting process, and (3) validating the achievement standards after cut scores have been adopted and implemented. States must now focus on grade-level content standards that have been made accessible to students with the most significant cognitive disabilities. The PLDs must focus on these content standards and not just refer to dimensions such as consistency, independence, or generalizability. That is, consistency is not a content standard but the degree of consistency expected may be described as it varies across content standards and across performance levels. Then, a formal "documented and validated" process must be used for setting the cut scores.

Overview of this Report

The purpose of this paper is to provide updated information on good practices for setting alternate achievement standards. These best practices will be compared to actual state practices.

The paper is organized in five major sections, four of which are devoted to a central theme in setting alternate achievement standards. The final section discusses some current issues such as setting modified achievement standards for the two percent population and setting standards for a growth model. The next sections will describe:

1. Guidance on how to develop strong performance level descriptors
2. A summary of commonly used standard-setting methods
3. A discussion of procedures that can be used to validate the alternate achievement standards
4. Best practices for documenting the development of alternate achievement standards
5. Other issues on setting achievement standards for special populations

PERFORMANCE LEVEL DESCRIPTORS

While these academic achievement standard descriptions, which will be referred to as performance level descriptors (PLDs) for the remainder of the paper, are most commonly associated with setting cut scores, they are also a useful development and reporting tool. The descriptors say in words what the cut scores mean and can help

teachers and parents interpret what their students know and can do and, potentially, what they do not know and cannot do.

Under the peer review guidance, critical element 2.5 highlights the need to ensure "*alignment between its academic content standards and the alternate academic achievement standards"* (USED, 2004). This alignment is first demonstrated in the development of the PLDs.

Well-written PLDs capture essential skills, align with state content standards, and represent the highest standard possible for this population. In addition, PLDs should clearly differentiate among levels, progress logically across levels (e.g., is Proficient appropriately higher than Basic), progress logically across grade levels (e.g., is grade 5 proficient sufficiently more advanced than grade 3 proficient), and represent knowledge and skills that can be evaluated by the assessment (e.g., don't discuss independence in the PLD if your assessment doesn't measure independence).

There are optional aspects of writing PLDs that states can use if it fits within their conceptual model for their alternate assessment program. For example, PLDs can be written for a grade span rather than a grade level. That is, if the content is sufficiently similar from one grade to the next, states may write one descriptor to cover the performance of up to three grade levels (e.g., grades 3–5). However, they must take care to demonstrate that students may still progress across grade levels within a grade span. For instance, consider a scenario where a student is assessed using the same skills checklist with the same requirement for Proficient across all three grade levels. If that student is able to perform well on 80% of the tasks at grade 3, well above the level of Proficient for that grade span, how will the state ensure that the student will be exposed to appropriately rigorous content in grades 4 and 5?

Another option is to adopt more than one set of PLDs. One common example that several states are using or considering is to develop one set of PLDs for their symbolic or emerging symbolic students and another set for their pre-symbolic students. Sometimes these distinctions are made in the design of the test. For example, a pre-symbolic student may be given a different set of performance tasks to do than a student at the symbolic level. Or perhaps a pre-symbolic student would only have to produce three pieces of evidence in his portfolio while a student at the symbolic level would produce five.

Regardless of the number of PLDs required by an alternate assessment program, developing the PLDs requires three steps:

1. Determine the number and names of the levels
2. Develop policy definitions for each level
3. Add content-specific information to develop full PLDs

These steps will be discussed in some detail in the next three sections. For additional information, see Perie (2007).

Determining the Number and Names of the Achievement Levels

The first decision policymakers must make is whether to use the same number and names for the alternate achievement standards as they use for the general assessment in their state. For example, if the general assessment uses the common nomenclature—Below Basic, Basic, Proficient, and Advanced—should the alternate assessment adopt the same names.

NCLB requires state to develop at least three levels, one for Proficient, one above and one below. Some states have four performance levels, allowing them to differentiate between students who are close to proficient and those who are well below proficient in addition to those who are proficient and above. Typically, no more than four levels are needed. Beyond this number, it becomes difficult to describe meaningful differences across levels. In addition, any particular test has a fixed amount of measurement power that depends primarily on the number and quality of the questions in the test. If there is only one cut score (giving two performance levels), a good test developer can focus most of the test's measurement power around that cut score. If there are two cut scores (giving three performance levels) the test developer has to split the available measurement power across the two cut scores, and so forth.

The next step is to name the levels. The terms themselves carry meaning, even without further description. The words chosen often express the values of the policymakers and thus should be selected carefully. As one example, Georgia has three levels on its general assessment, called *Does not meet the standard*, *Meets the standard,* and *Exceeds the standard*. Policymakers in that state chose to keep the same number of levels for their alternate assessment but to give them different names: *Emerging progress, Established Progress,* and *Extended Progress*. In contrast, Kentucky has the same number and names for their achievement levels in both the general and alternate assessment: *Novice, Apprentice, Proficient,* and *Distinguished*. Table 3 provides examples of naming conventions for the alternate assessments used in a handful of states.

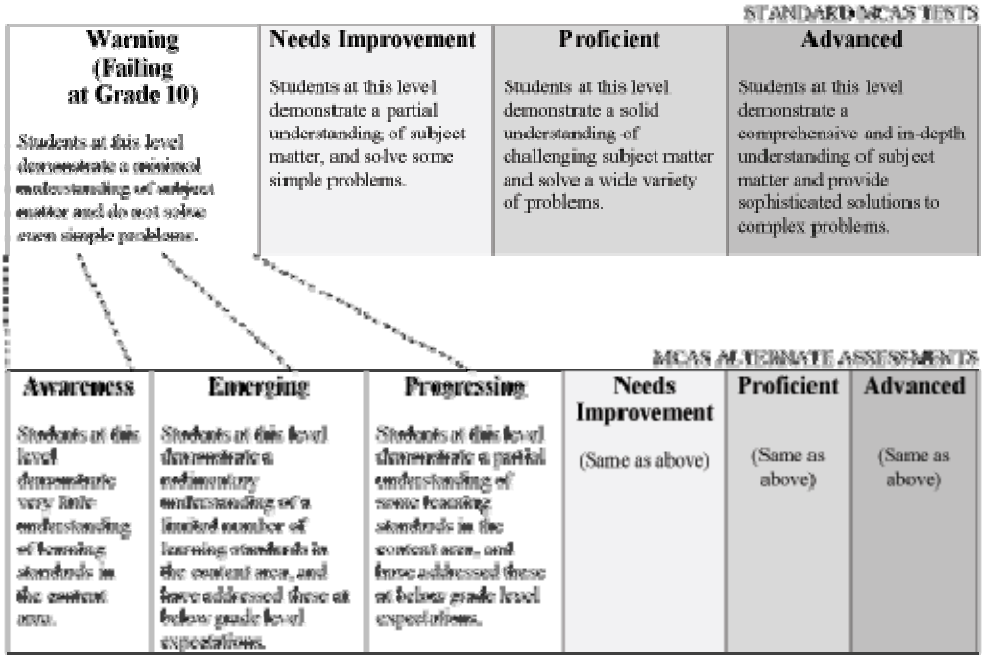
Table 3: Naming Conventions of Alternate Achievement Levels for Select States

State	Alternate Assessment Achievement Level Names
AZ	Emergent, Supported, Functional, Independent
GA	Emerging Progress, Established Progress, and Extending Progress
IA	Basic, Proficient, Advanced
IL	Attempting, Emerging, Progressing, Attaining
KY	Novice, Apprentice, Proficient, Distinguished
NH	Substantially below Proficient, Partially Proficient, Proficient, Proficient with Distinction

Other states avoid labeling the levels and simply refer to them as Level 1, Level 2, Level 3, and Level 4 (e.g., South Carolina and Washington). Massachusetts has a unique

approach. The task force recommended that performance levels be identical to performance levels on the general assessment (the Massachusetts Comprehensive Assessment System or MCAS), but that the lowest performance level, called "Warning/Failing at Grade 10" for tested students, would be sub-divided into three distinct levels in order to provide more meaningful descriptions of performance at these lower levels (Wiener, 2002). Figure 1 illustrates the performance levels and definitions used by Massachusetts to report assessment results on the standard and the alternate assessments, and the relationship between the two reporting scales.

Figure 1. MCAS Performance Levels



Once the number and names of the levels have been determined, descriptions of these levels can be written.

Developing Policy Definitions

Once the number and names of the levels have been selected, they need to be defined. A strong approach involves developing a generic policy definition for each performance level prior to drafting any performance level descriptors. Policy definitions determine how rigorous and challenging the standards will be for the assessments. They are not linked to content but are more general statements that assert a policy position on the desired level of performance or rigor intended at each level. A policy definition needs to be written for each performance level.[4] For instance, if a state is interested in setting

[4] Not all programs define the lowest level, as it includes those scores that fall below the first cut score and can range from zero points to one point less than the cut score.

just one cut score, at the proficient level, then a policy definition should be written for the proficient category. It is not necessary to write another definition for below proficient.

One key to writing strong policy definitions is to use a similar set of words that are memorable and that distinguish clearly between the performance levels. State policymakers should begin drafting the policy definitions by making a statement that is directly linked to their instructional program and goals. Their definitions should clearly state the degree of knowledge and skills expected of students at each performance level. The policy definition should apply to all subjects and grade levels and should answer the question "How good is good enough?" That is, in general terms, what is meant by proficient? This definition should be concise, 1–2 sentences, but because it is the backbone of all further writing, policymakers should carefully consider the wording. If policymakers do not write the policy definitions, they should at least approve them as these definitions reflect the level of rigor intended for each performance level.

Some examples of policy definitions are as follows:

Arizona:

E = Emergent
Student is beginning to use skill in one context with extensive support. Student cannot perform the skill without assistance. Student initiates any portion of the skill sequence but needs physical/verbal assistance to complete task.

S = Supported
Student occasionally uses the skill in one or more contexts with physical/verbal cues. Student occasionally performs the skill accurately. Student demonstrates the skill from 1–90% of the time with physical/verbal cues.

F = Functional
Student frequently uses the skill in one or more contexts with limited cues. Student frequently performs the skill accurately. Student demonstrates the skill from 91–100% of the time with physical/verbal cues or from 1–90% of the time with natural cues.

I = Independent
Student performs the skill accurately in several contexts with natural cues. Student demonstrates the skill from 91–100% of the time with natural cues.

New Hampshire

New Hampshire (a portfolio state) uses a slightly modified approach to what was discussed previously. They develop descriptions for each subject first without distinguishing performance by grade level. So the policy definitions and full performance level descriptors are combined. The grade level distinctions are made in the discussion of borderline performance for each grade span. The following are their descriptions for mathematics.

Level 1: Substantially below Proficient
Student demonstrates *little or no progress* in any targeted mathematics skills using the modified mathematics materials and/or activities presented. Student is *not accessing* modified mathematics materials that are linked to general education curriculum activities.

Opportunities to practice mathematics skills in various settings are *limited*. Opportunities for self determination and typical peer interaction are *rare or not present*.

Redesigned instructional supports, team supports, and/or task structure *are necessary* for this student to access modified grade-linked mathematics materials and/or activities in a manner that promotes skill progress, generalization of performance, and self determination.

Level 2: Partially Proficient
Student is demonstrating *some progress* in targeted mathematics skill(s) using the modified mathematics materials and/or activities presented. Student has *some access* to modified mathematics materials that are linked to general education mathematics curriculum activities.

Opportunities to practice mathematics skills in various settings are *somewhat limited*. Opportunities for self determination are *inconsistent*. Typical peer interactions are *inconsistent or not evident*.

Redesigned instructional supports, team supports, and/or task structure *may be necessary* for this student to access modified grade-linked mathematics materials and/or activities in a manner that promotes skill progress, generalization of performance, and self determination.

Level 3: Proficient
Student is successfully demonstrating *moderate progress* that is consistent with the intended goal(s) in targeted mathematics skill(s). Student *has access to and is using* a variety of modified mathematics materials that are linked to general education mathematics curriculum activities.

Opportunities to practice mathematics skills are offered *in varied settings*, or *consistently within a general education or other natural setting*. Opportunities for self determination and interaction with typical peers are *consistent*.

Instructional supports, team supports, and/or task structure are *adequate* for this student to access modified grade-linked mathematics materials and/or activities in a manner that promotes skill progress, generalization of performance, and self determination. Remaining areas of weakness can be addressed by the existing team.

Level 4: Proficient with Distinction

Student is successfully demonstrating *extensive progress* in targeted mathematics skills. Student *has access to and is using* a variety of modified mathematics materials that are linked to general education mathematics curriculum activities.

Opportunities to practice mathematics skills are offered in *varied settings* and include naturally embedded supports, or this student is *included fulltime in the general education classroom for mathematics*. Opportunities for interaction with typical peers and different adults are *extensive*. Opportunities for self determination are *consistent and include all required components*.

Instructional supports, team supports, and task structure *are effective* and allow this student to successfully access modified grade-linked mathematics materials and/or activities in a manner that promotes skill progress, generalization of performance, and self determination.

Writing Full Performance Level Descriptors

After the policy definitions have been completed and agreed upon, content descriptors are added to develop full performance level descriptors. Performance level descriptors (PLDs) state in words the knowledge and skills required to achieve each level of performance for a specific assessment and are linked directly to the content standards for that assessment. They should be developed prior to setting cut scores and used to inform the cut-score setting process. In addition, they can be used to provide parents, teachers, and other stakeholders with more information on what students at each level know and are able to do and what they need to know and be able to do to reach the next level.

The full PLDs should be written by committee. They should be informed by those with both content expertise and experience with students with significant cognitive disabilities. Typically, these descriptors begin with the policy definition—used to maintain a similar level of rigor across grades and subjects—and then add appropriate content descriptions linked to the grade-level content standards. Again, these descriptors should represent the highest standard possible for this population.

The steps for writing full PLDs are as follows:

1. Convene a panel of stakeholders for each subject and grade span. This panel should include both special educators and content experts. That is, to write PLDs for grades 3–5 mathematics, include both teachers who teach special education classes at the elementary level and teachers who teach elementary mathematics. District-level curriculum supervisors also may be included as content experts. If the panel is to develop descriptors for high-school students, consider inviting representatives from higher education or people from the community that might hire high school graduates. Ideally, each panel will consist of 4–6 participants.

Fewer panelists will not provide sufficiently diverse experiences and opinions; more panelists will require more time to reach consensus.

2. Provide information about the students and assessment. The content experts will need information on who the one percent population is and what their disabilities encompass. All participants will need to be briefed on the characteristics of these students, including how they communicate and the level of support needed. All participants will need to review the grade-level content standards and review the requirements of the test. A discussion about the types of accommodations allowed (if relevant) and different ways in which the students may access the curriculum is also helpful.

3. Share relevant literature about what students with disabilities can learn and do. Those on the panel who have not previously worked with students with significant cognitive disabilities will need an orientation on what these students are capable of accomplishing. They should understand what the research says about how these students develop competence in the domains of reading and mathematics (cf., Kleinert & Kearns, 2001).

4. Share sample student work. Regardless of the participant's background, they will benefit from seeing examples of student work. A sample portfolio will help them better understand the types of evidence students produce. A videotape of students taking a task-based performance test or performing skills on a checklist will give them a visual of how these students are demonstrating what they know and can do.

5. Discuss requirements for PLDs. The participants will need to be oriented to the task of writing performance level descriptors. They should be shown the policy definitions and led through a discussion of the difference in the degree of rigor across levels. Provide information on the desired format for the PLDs—one paragraph, a page of bullets, or a description separated by specific dimensions, for example.

6. Ask panel to draft PLDs as a group. Have the participants work together, going through each content standard assessed and each dimension used in a scoring rubric, and writing descriptions of the required level of performance for each level. Typically, the group starts by brainstorming simple ideas, written in bullet format. After the list of ideas has been created for each performance level, then the ideas can be connected and full descriptions can be written. The group should work toward consensus.

7. Compare PLDs across subjects and grades and revise as needed. It is important that the PLDs show coherence from one grade (or grade span) to the next. The PLD for Proficient at grade 5 should be appropriately more rigorous than the PLD for Proficient at grade 3. The look and feel of the descriptors (e.g., length)

should be similar across subjects. After all PLDs have been written, all participants should gather together and review all PLDs as a whole to ensure they communicate a consistent, coherent message about the expectations of the alternate assessment system.

8. Final revisions and adoption. These PLDs, like the PLDs for the general assessment, need to be adopted with the associated cut points by the state. (Depending on the State's governance structure, this could be approval by the state commissioner or the board of education.) These descriptors verbalize the state's policy regarding alternate achievement standards and should be reviewed carefully and modified as needed to communicate the intended message.

These adopted PLDs should be used in the standard-setting workshop. After the workshop, both the PLDs and cut scores will be available. At this point, the state policymakers may decide they want to enhance the alternate achievement standards by including examples. Once the cut scores have been determined, statisticians and content experts can identify student work that exemplifies performance at each level. This additional information can help policymakers communicate the expectations of students at each level of performance.

In reviewing the PLDs, be sure that they clearly distinguish content from one performance level to the next. There has been a temptation to vary the descriptors simply in terms of the level of complexity and/or support needed. While these are important components, consider the interaction between the content and the level of support needed. For instance, a proficient student at grade 3 reading may be able to identify the main character in a text with minimal support but may need more support to use supporting detail from the text to describe that character's motivation. The same consideration should be given to the interaction between complexity and performance. Transfer or generalizability will also interact with the specific content standards. For instance, identifying the main point in a five-sentence fictional story may be less difficult than identifying the main point in a poem or a newspaper article. An example of a full performance level descriptor is shown on the following page.

PERFORMANCE LEVEL DESCRIPTORS
English Language Arts Grade 3

Below Basic ( )	Basic ( )	Proficient ( )	Advanced ( )
Provided supports such as assistive technology, adaptations, and/or modifications, and a skill that is reduced in complexity the student demonstrates inaccurate or minimal knowledge of English Language Arts content as outlined in the following: Language Development • Identify prefixes and suffixes • Identify unfamiliar words or words with multiple meanings Informational Text • Identify purpose of simple text • Identify stated fact or opinion • Identify stated cause or effect • Locate information on a graphic representation • Locate specific information from text. Literary Text • Identify events/characters/author of a story • Identify story elements and events • Identify characters • Identify types of literature • Match moral to its fable • Identify rhymes • Identify a poem	Provided supports such as assistive technology, adaptations, and/or modifications, and a skill that is reduced in complexity the student demonstrates limited/basic yet accurate knowledge of English language arts content as outlined in the following: Language Development • Identify words with prefixes and suffixes • Identify words with multiple meanings Informational Text • Identify purpose of simple text • Identify fact or opinion • Identify stated cause or effect • Locate information on a graphic representation • Locate specific information from text. Literary Text • Identify events/characters/author of a story • Identify story elements and events • Identify characters • Identify types of literature • Match moral to its fable • Identify rhymes • Identify a poem	Provided supports such as assistive technology, adaptations, and/or modifications, and a skill that is reduced in complexity the student demonstrates understanding of English language arts content as outlined in the following: Language Development • Use affixes to change the meaning of a root word • Use context cues to complete a cloze sentence Informational Text • Identify purpose or main points • Distinguish between fact and opinion • Identify stated cause and effect relationships • Answer questions about graphic representations • Locate specific information from text (e.g., letters, memos, directories, menus, schedules, pamphlets, search engines, signs, manuals, instructions, recipes, labels, forms). Literary Text • Classify events/characters as having happened in author's life or not • Identify story elements and events of the story • Identify character's traits, relationships and feelings • Classify literature using structural elements • Identify morals of fables • Define patterns of sounds or rhythm patterns in poetry • Identify a poem	Provided supports such as assistive technology, adaptations, and/or modifications, and a skill that is reduced in complexity the student demonstrate understanding and application of English language arts content as outlined in the following: Language Development • Analyze the meaning of unfamiliar words using root words and affixes. • Analyze context cues to determine the correct meaning of a word with multiple meanings. Informational Text • Identify purpose or main points and summarize supporting details • Distinguish fact from opinion • Identify cause and effect relationships(stated and implied) • Interpret information in graphic representations • Locate specific information from text (e.g., letters, memos, directories, menus, schedules, pamphlets, search engines, signs, manuals, instructions, recipes, labels, forms). Literary Text • Compare characters or events in a story to author's life experiences • Understand how story elements influence the events of the story, using specific examples from the text. • Describe character's traits, relationships, and feelings supported with text • Compare/contrast forms of literature using structural elements • Compare morals of fables • Recognize similarities of sounds in words and rhythmic patterns in poetry • Identify characteristics or structural elements of poetry

METHODS FOR SETTING CUT SCORES

This section will provide a brief overview of the steps that should be included in any standard-setting study. Next, it contains a summary of the commonly-used methods and then discusses how to choose a method. Finally, it summarizes the latest survey results on the methods states are using to set alternate achievement standards.

Steps for All Methods

Regardless of the method chosen, all standard-setting methods follow essentially the same steps. For a more detailed explanation of all of the tasks see Cizek & Bunch (2007) or Zieky, Perie, & Livingston (forthcoming).

A list of all steps would include:

1. Choose standard setting method
2. Convene panel of stakeholders
3. Review content standards and assessment
4. Discuss PLDs and borderline student
5. Train on and practice methodology
6. Run 2-3 rounds with individual judgment, feedback, and group discussion
7. Provide normative data based on operational test results
8. Summarize results of “provisional” cut scores
9. Smooth data across grades/subjects
10. Formally adopt cut scores

There is a subsequent section on choosing a standard setting method, so there is no need to spend time on that here. Convening a panel of stakeholders requires finding at least 10–15 panelists per subject area and grade span to agree to come to a 2–3 day standard setting workshop. It is important that the panel be comprised of both special educators and content experts. Collectively, the panel should represent expertise both in working with students with significant cognitive disabilities and in the subject area being assessed.

Within each standard-setting workshop, the facilitator should spend time reviewing the content standards and the assessment. Often providing samples of student work will help the panelists better understand the nature of the assessment. Significant time should be devoted to discussing the performance level descriptors and discussing what it means to be at the bottom of that performance level. Some refer to this discussion as the discussion of the “borderline candidate” or the “target student.” We can also think of this examinee as the student who is just barely Proficient (or Basic or Distinguished, etc.). The questions these panelists should be discussing include “What will this barely Proficient student know? What skills will she have? How will she show her level of knowledge and skill?” This discussion is essential to the process of setting cut scores, as the panelists must agree what it means in words to score at the very bottom of a

performance level before they can determine the most appropriate minimum score (cut score) for that level.

Once the panelists are grounded in the content, assessment practices, and expectations for each performance level, they will need to be trained on the standard-setting methods they will be using. It is essential that adequate time is spent on both training the panelists to use the methodology and allowing them a chance to practice that methodology. If they will be rating portfolios, provide them extra portfolios to practice rating before they start round 1. If they will be providing judgments on items, have additional (e.g., field test) items available for them to practice making judgments on. At the end of the practice session, the facilitator should ask the panelists to complete an evaluation form to assess their understanding of the task and readiness to proceed. If the panelists are at all unclear on their task, additional training should be provided.

When the panelists are ready to proceed, round 1 of the standard setting process can begin. Typically, there are three rounds of standard setting. The first round consists of independent judgments or ratings. The second round provides feedback on the range of judgments or ratings and a chance for group discussion before a second set of judgments or ratings. The third round includes the same information as the second but also provides the judges with normative data. This normative data is typically in the form of consequence or impact data, meaning that given the average (or median) cut score from round 2, the distribution of actual student scores can be calculated. Then the panelists can be told the percentage of students who would fall into each category. If they feel the distribution looks improperly skewed (e.g., too many students scoring above Proficient, or too many scoring below basic), they can adjust their ratings so that the impact data better match their expectations.

At the end of the standard setting workshop, an additional evaluation form should be provided to gather information about the panelists' understanding of the task, comfort level with the results, and overall impression of the process. The results of the judgments or ratings should be summarized in a manner that shows the level of variance across panelists and across rounds. The provisional cut score can be the mean, trimmed mean, or median of the individual judgments.

It will be important to examine these provisional cut scores across subjects and grades along with the student impact data. Avoid strong discrepancies in percentages of students at proficient or above across grades. For example, if 60 percent of students are Proficient or above at grade 3, 65 at grade 4, 40 percent at grade 5, and 70 percent at grade 6, questions might be raised about the severity of the cut score at grade 5. Some smoothing might be required. In addition, it is important to consider the distribution of scores across the alternate achievement standards compared to the distribution of scores across the performance levels for the general assessment. That is, if 50% of all students scored Proficient or higher on the general assessment, and 85% of students with significant cognitive disabilities scored Proficient or higher,

policymakers may want to consider if the alternate achievement standards are truly holding these students to a high enough standard.

Both the state department of education and the state policy board (commissioner or state board of education) should review the provisional cut scores and make adjustments as needed that will result in a consistent set of achievement standards for the alternate assessment program. For more information on how to achieve this consistency, see Perie (2006). The Department of Education's peer review of assessment systems requires that the full alternate achievement standards—both the PLDs and the cut scores—be adopted formally by the state.

Specific Methods

This section provides brief descriptions of each of the methodologies typically used to set cut scores on alternate assessments. These descriptions were taken largely from Zieky, Perie, and Livingston (forthcoming) and further discussion can be found in that report. If more detail on specific procedures for implementing the methods and analyzing the results is needed, see the individual chapters on each method in Cizek & Bunch (2007).

We divide the methodologies into three types based on the judgments required. As panelists are evaluating evidence to determine the cut score, they can judge the items, the rubrics, or the students themselves or the work they produce.

Methods Requiring Judgment of Items

The first type of methodology examined here requires panelists to make judgments about individual items. Typically, the judgment they make involves hypothesizing how a student who just barely met the description for Proficient (or Basic or any other category name) would perform on an item. These methods are typically applied in tests that are primarily multiple-choice, with some constructed-response items, or to checklists where each skill can be treated as an item.

Angoff

People most often refer to the "modified Angoff." However, there is no real Angoff method, so anything anyone does is a modified Angoff. Back in 1952, Ledyard Tucker had a discussion with William Angoff about having committees speculate whether a minimally acceptable candidate would respond correctly or incorrectly to each item. Later, Angoff referenced this discussion in his book on scaling and equating and added a footnote suggesting a modification to Tucker's approach in which committee members would estimate the probability of a correct response by a minimally acceptable candidate (Angoff, 1971). Tucker's suggestion is now most commonly known as the Yes/No method and Angoff's footnote has become the "modified Angoff" approach. In addition, there now exists the "extended Angoff" which follows similar logic for open-ended items.

Yes/No Method

This method, initially suggested by Angoff and Tucker (Angoff, 1971) was resurrected in the 1990s and is now commonly referred to as the Yes/No method (Impara & Plake, 1997). In this method, all items must be scored right/wrong. The panelists examine the items and determine whether the target student is more likely to answer the question correctly or incorrectly. If they think the target student is likely to get the answer right, they record a 1, and if they think the target student is likely to get the answer wrong, they record a 0. Then, all the 1's are summed to obtain the total cut score for that performance level. A group performance level can be calculated by calculating the mean or trimmed mean of all the panelist cut scores.

This method is only appropriate for tests in which every item is scored right/wrong as in a skills checklist approach. Then the question for the panelists becomes "is the target student likely to complete this skill?" for every skill on the checklist. Summing the number of times the panel says "yes" results in the cut score. Iowa used the yes/no method recently to set cut scores on its skills checklist.

"Modified" Angoff

In effect, all versions of an Angoff standard setting are modified versions as all Angoff ever recommended was to ask qualified panelists one question. So, adding more than one round, providing feedback, even training panelists is a modification to the original suggestion. Thus, there is no one "modified Angoff" methodology, although the following discussion represents the common elements in most modified Angoff methodologies (cf, Zieky, Perie, & Livingston, forthcoming).

In this method, standard-setting panelists are asked to state the probability that the target student would answer each item correctly. Summing the expected scores for all the items produces the expected score on the whole test. Therefore, you can find the expected score for a borderline test taker if you know, for each question on the test, the probability that a borderline test taker will answer correctly. In this way, each participant's judgments lead to an estimate of the cut score. The group's estimate of the cut score is calculated by averaging the individual participants' cut scores, using the mean or the trimmed mean.

Angoff has the advantage of being the most researched standard-setting methodology. It is widely used in certification/licensure testing. There are few circumstances in which it will be appropriate for alternate assessments, as few assessments use multiple-choice formats. However, it can be considered for an assessment that uses a skills checklist approach.

Extended Angoff

The extended Angoff approach is intended for open-ended items scored using a rubric. The Angoff methods described previously are limited to questions scored right/wrong or

0/1. However, these methods can be extended to work with performance tasks or other open-ended items that are given more than one point per question. In this method, panelists determine which rubric point is most closely identified with the performance of the target student, such as the student who just barely meets the definition for proficient (Hambleton & Plake, 1995). For each performance task or open-ended item, the panelists would select the minimum rubric point they believe the target student would achieve. These estimates are then summed to obtain the cut score for each panelist and then averaged across panelists for the recommended cut score. The Extended Angoff method can be combined with either the Yes/No or modified Angoff method for a test that mixes questions that are scored right-wrong and questions that are scored using a rubric.

This method is appropriate for some task-based approaches where student work is not readily available to evaluate. Pennsylvania used this method to set cut scores on their task-based assessment.

Mean Estimation Method

A variant of the Extended Angoff that has been used most commonly in NAEP standard-setting studies is the mean estimation method (Loomis & Bourque, 2001). It has also been used in many state assessments with open-ended items. Instead of asking participants to select the rubric value most closely identified with the target student, simply ask the panelists to estimate the average score that a large group of target students would obtain on the question. This average score does not have to be an integer. For example, if a particular is scored on a scale from 1-4, one panelist might estimate that a group of target students would obtain an average score of 2.8. Another panelist might estimate the average score of target students to be 3.2, and so forth. As in the previous Angoff methods, sum the estimated scores across items for each participant to get each participant's cut score. Then use the mean or trimmed mean to calculate the group's estimate of the cut score.

There is some evidence that this method produces less biased cut scores than the original extended Angoff method (also called the item score string estimation method in the NAEP studies), which tended to bias the estimates of the lowest levels down and the highest levels up (Loomis & Bourque, 2001, p. 194–195) meaning too few students would be categorized in both the lowest and highest achievement levels.

This variant can be used for the same type of assessments as the original extended Angoff approach, namely, task-based assessments scored using a rubric.

Bookmark

The Bookmark method was developed to be used with tests that are scored using Item Response Theory (IRT; Mitzel, et al., 2001). It is now one of the most widely used standard-setting methods for state K-12 assessments for the general population.

The participant is given a special test booklet called an *Ordered Item Booklet* that displays the questions in order of difficulty from easy to hard. The participant's task is to place a bookmark at the spot that separates the questions into two groups—a group of easier questions that the borderline test taker would probably answer correctly (with *probably* defined as a chance of at least 2 out of 3 or .67), and a group of harder questions that the borderline test taker would probably not answer correctly (i.e., the test taker would have a probability of less than .67 of answering correctly.)

The bookmark placement can then be converted into an expected test score using IRT software based on the ability of a test taker who has a 2/3 probability of correctly answering a question of that difficulty. To use this method as it was designed, you must have a test that was calibrated using IRT. You also must have one form or one pool of items that all students take. The majority of items should be scored right/wrong, but an advantage of this method is that it can combine multiple-choice and open-ended items into one set of judgments for panelists. So, this method could be used with an alternate assessment with tasks scored right/wrong combined with tasks that are scored on a rubric. For example, some performance tasks or checklists may have certain items that a student either gets or does not get, while others may be scored polytomously depending on how much support a student needs in order to get the item right. Colorado used this method to set cut scores on its alternate assessment.

ID Matching

The Item Descriptor Matching method, also called ID Matching, also requires a test that was calibrated using IRT (Ferrara, Perie, & Johnson, 2007). In ID Matching, the participants associate each test question in the Ordered Item Booklet with a performance level, using the performance level descriptors and their own determination of what a test taker must know and be able to do to answer the question correctly. Participants match the knowledge and skills required to answer each question to the knowledge and skills required to be in a performance level. For example, if a question requires a test taker to determine the average temperature over a fixed period of time, and the ability to perform mathematical computations such as averaging is part of the performance level descriptor for the Advanced Level and not for any lower level, the participant would match the question with the Advanced level.

The participants write the initial of the selected performance level (e.g., Basic, Proficient, or Advanced*)* next to each question. After matching each question to a performance level, the participants analyze the pattern of matches to determine where the knowledge and skills seem to change from one performance level to the next. Then they draw a cutline between two questions at the point that best represents the borderline between the lower performance level and the next higher performance level. The cut score is calculated as in the Bookmark method. South Carolina and New Mexico both used this method to set cut scores on their task-based alternate assessment.

This method has similar advantages and uses as the Bookmark method. However, keep in mind that both of these methods assume that all students follow the same learning trajectories. That is, regardless of disability type, the student will have a decreasing chance of success on items that are further into the ordered item booklet. You must believe this assumption is true for your alternate assessment prior to using either the Bookmark or ID Matching procedure.

Judgments Based on Score Profiles

The next set of methods ask judges to focus on the scores themselves. Most methods focus on weighting the importance of various subscores, such as those given across dimensions. Others focus on the whole scale. Still others examine profiles of performance across both items and sections of the assessment. All these methods focus on how the items are scored, rather than on the cognitive demands of the items themselves. In most cases, these methods are used with assessments that are scored using rubrics.

Reasoned Judgment

In the Reasoned Judgment method, a score scale is divided into the appropriate number of categories in some way determined by a panel. The method of dividing the scale could be into equal parts or placing the division in such a way as to have a larger number of scores in the middle level and smaller on either end (Roeber, 2002). For example, if scores range from 0 to 32 and four performance levels are needed, a panel may choose to divide the scale equally so that level 1 is 0-8 points, level 2 is 9-16 points, level 3 is 17-24 points, and level 4 is 25-32 points. Or, they could choose to weight the middle categories and make level 1 equal 0-5 points, level 2 equal 6-16 points, level 3 equal 17-27 points, and level 4 equal 28-32 points. This method can be done relatively quickly and with minimal preparation.

This method works best when panelists are given specific criteria to evaluate in making their judgments. For example, in Massachusetts, panelists were provided with all possible score combinations on the portfolio assessment (Wiener, 2002). Each portfolio was scored on a 1–4 scale on each of three dimensions. Thus there were 64 different score combinations possible. Panelists used a consensus approach to determine what performance level was demonstrated by each possible score combination. For example, a 4 on complexity, a 2 on skills, and a 2 on independence was rated “emerging” while a 2 on complexity, 3 on skills, and 3 on independence was rating “progressing.” In this case, performance levels were not set based on total score but on the combination of scores on each dimension. In many portfolio assessments, including those used in Massachusetts, the portfolios include evidence from more than one content strand. In this particular case, a performance level was determined for each strand, and then the levels were averaged over all strands to determine the predominant performance level. Thus the score report showed the ratings on each dimension for each content strand, the performance level on each content strand, and the overall performance level.

Judgmental Policy Capturing

In this method, the panelists review the various components of an overall assessment (which might be quite similar or quite dissimilar) and determine which of the components are more important than others. This might suggest weighting one type of item more important than another, or might be to weight one type of evidence (e.g., performance measures) as more important than another (e.g., a checklist) (Jaeger, 1995). Originally, this technique was implemented by asking participants to rate each sample of student work as Basic, Proficient, or Advanced (or whatever performance levels were being used). Then, the ratings could be used to derive a multiple regression equation that provides weights for each dimension, component, or task. This idea of weighting is produced more directly in the next method, the dominant profile.

Dominant Profile

The Dominant Profile method is used when (1) different parts of the test measure different skills or types of knowledge, (2) separate scores are computed and reported for each part of the test, and (3) there is a reason *not* to have only a single overall cut score. The outcome of this method is not a single cut score, but a set of decision rules. Thus the task is more than evaluating work samples and making an overall judgment, it requires participants to state their decision rules explicitly (Plake, Hambleton, & Jaeger, 1997).

In alternate assessment, this method is particularly useful when tests are scored on several dimensions, such as performance, progress, generalization, and complexity. Panelists determine rules for the cut score, explicating whether there needs to be a minimum score on each dimension, or the total test, or some combination. It highlights the difference between compensatory and conjunctive scoring and requires panelists to explicitly state whether a high score on one dimension can compensate for a low score on another. The panelist's task is to become familiar with the meaning of each dimension and to specify rules for determining which combinations of scores on these dimensions represent acceptable performance and which do not. The rules can combine information from the dimension in various ways, as in the following example:

To reach proficient, the student must

- Score at least 2 points on each dimension
- Score at least 3 points on the achievement dimension
- Produce a total score of at least 10 points

This method should be used only when the score reports provide scores on each dimension separately. It assumes that conjunctive decision rules are important. For example, when policymakers decide that a high independence score cannot compensate for a low complexity score, a conjunctive model based on score profiles would be appropriate.

Portfolio Pattern

Considered by some to be a form of Reasoned Judgment, the Portfolio Pattern method also shares some similarities with Dominant Profile and Judgmental Policy Capturing. This method actually creates a cell for every possible score point combination and asks the panelists to determine whether that particular pattern of scores represents Basic, Proficient, or Advanced performance (or whatever the performance levels are). After the initial ratings are complete, panelists are shown actual portfolios that match the profiles they indicated as being right on the cut score.

For example, the panelists would be asked to complete a chart such as the abbreviated one shown in Figure 2, where there are three dimensions, each scored 1–4.

Figure 2. Sample Pattern Rating Sheet

Pattern	Basic	Proficient	Advanced
111			
112			
113			
114			
121			
122			
123			
124			
…			

(The first number is for complexity, the second for performance, and the third for independence.)

The panelists would check the appropriate box to indicate which performance level each pattern represents. Results would be collated and shared with the group.

Once agreement had been reached on where the differentiation lies between each of the performance levels, portfolios would be selected to share with the panelists to allow them to see which portfolios would be categorized as Basic, Proficient, or Advanced given their ratings. In the example shown in Figure 3, panelists would be given portfolios with the scores 221, 231, 212, 222, 213, 223, and 214. They would be asked to sort those portfolios into either Basic or Proficient and then examine to see if the profiles they placed in each category matched the placement of the rubric profile. Then, the same process would be followed for the profiles around the cut scores for the Proficient/Advanced distinction.

This approach assumes a conjunctive scoring method. That is, not all students who score a total of 6 points will be treated the same. In this example, some would be classified as Basic and other Proficient. So, it serves to weight one dimension over another. It allows for different patterns of scores without requiring panelists to verbalize their rules. It is intended to be used with portfolios or other body of evidence approaches to alternate assessment. Georgia has used this method recently.

Figure 3. Sample Matrix with Assigned Performance Levels.

Complexity receives a score of One (1)					
	Performance score points received				
Independence score points received		1	2	3	4
	1	B	B	B	B
	2	B	B	B	B
	3	B	B	B	B
	4	B	B	B	B

Complexity receives a score of Two (2)					
	Performance score points received				
Independence score points received		1	2	3	4
	1	B	B	P	P
	2	B	P	P	P
	3	B	P	P	P
	4	P	P	P	P

Complexity receives a score of Three (3)					
	Performance score points received				
Independence score points received		**1**	**2**	**3**	**4**
	1	P	P	P	P
	2	P	P	P	A
	3	P	P	P	A
	4	P	A	A	A

Complexity receives a score of Four (4)					
	Performance score points received				
Independence score points received		**1**	**2**	**3**	**4**
	1	P	A	A	A
	2	A	A	A	A
	3	A	A	A	A
	4	A	A	A	A

Performance Profile Method

The Performance Profile method works best with tests consisting of small numbers of performance questions or tasks (generally, seven or fewer). It has been used on tests for very young test takers and also works well with performance assessments designed for those with significant cognitive disabilities (Morgan, 2004).

The participants review individual test taker score profiles, each profile showing a test taker's scores on the individual performance questions. The profiles are arranged in order of their total scores, from lowest to highest in an Ordered Profile Booklet. The

participants first become familiar with the meanings of the scores in the profile. The participants then examine the ordered profiles and select the first profile that is indicative of borderline performance at a performance level. For example, a participant may decide that a profile of 4, 6, 3, 5 with a total score of 18 is indicative of borderline Proficient performance.

There will probably be several different profiles with the same total score as the selected profile. Participants next examine all of the profiles provided with the same total score as the selected profile. To continue the example, profiles of 9, 2, 1, 6 and 5, 4, 5, 4 have the same total score as the selected profile. If all of the profiles at the selected total score are judged to represent Proficient performance, the selected total score is the participant's cut score for the Proficient performance level.

A participant may decide, however, that some of the profiles at the same total score represent Proficient performance, but others are below Proficient. For example, a participant may decide that 5, 4, 5, 4 represents Proficient performance, but 9, 2, 1, 6 does not. In that case, the participant has to decide whether it is preferable to treat all of the profiles at that total score as Proficient or to treat all of the profiles at that total score as below Proficient.

If a participant believes that it is preferable to treat all of the profiles at a total score as below Proficient, the participant repeats the process at the next higher total score, and so on until a total score is found at which the participant decides that it is preferable to treat all of the profiles at that total score level as Proficient.

This method allows participants to evaluate different ways of obtaining the same total score and to decide if all of the ways of obtaining the same score are good enough to be acceptable within a performance level or not. The total score of the profile selected by the participant is the participants' cut score. It can be used with a performance assessment that includes only a few tasks. California has used this method to set cut scores on its alternate assessment.

Methods Requiring Judgments of People or their Products

The final type of method discussed here involves asking panelists to judge the students themselves or actual samples of student work. The panelist must either be familiar with selected students' knowledge and abilities or have the opportunity to examine work samples produced by those students. Thus, these methods are used most often to evaluate portfolio assessments. However, they could be used to evaluate any assessment if sufficient evidence is gathered.

Contrasting Groups

This method is based on the idea that the test takers can be divided into two contrasting groups on the basis of judgments of their knowledge and skills: a group that is qualified to belong in a performance level and a group that is not qualified to be

in a performance level (Livingston & Zieky, 1982; Zieky, Perie, & Livingston, forthcoming). For example, you could ask a sample of teachers, who do not know their students' test scores, to identify their students who are performing at the Proficient level in mathematics and those who are performing at the Basic level in mathematics based on the performance level descriptors. That is, they match the knowledge and skills of their students to the knowledge and skills described in one of the descriptors.

Once the teachers have divided the test takers into these two groups, you can obtain the test takers' scores and find the test score that best separates the two groups. For example, you can find the test score at which test takers are equally likely to be in the Basic group and in the Proficient group, which would be the borderline between the two groups. To do that, consider the test takers with a particular test score and ask, "What percent of the test takers are Proficient?" If the test score is low, most of the test takers will be in the Basic performance level. As you go up the score scale, the proportion of the test takers who are Proficient will increase. At the lower score levels, the Basic test takers will outnumber the Proficient test takers. At the higher score levels, the Proficient test takers will outnumber the Basic test takers. One reasonable choice for a cut score would be the score at which 50 percent of the test takers are Proficient because that would represent the borderline of the Proficient performance level.

In most cases it will not be practical to get judgments of all test takers who have taken the test. You will have to choose a cut score on the basis of a sample of the test takers. How should you choose the sample? If you have to choose the sample of test takers before you have given the test, you should try for a representative sample of all the people who will be taking the test. (One way is to choose them at random, for example, by lottery.) But if you can choose them *after* they have taken the test, there is a better way. You can choose the test takers so that their scores are spread evenly throughout the portion of the score range where the cut score might possibly be located. For example, on a 100-question test, you might choose 10 test takers from each five-point score interval (31–35, 36–40, etc.). The important principle to remember is that the sample of test takers you select *at each test score level* must be representative of all the test takers *at that test score level*.

This method can be used with any alternate assessment as it does not involve judgments about the test or the items, but rather judgments about the people. For certain assessment types, like portfolios, judgments can also be made about the student products in much the same fashion as were made about the people. Judging student products minimizes distortion resulting from inaccurate teacher judgment or personal opinion of an individual student.

Up-and-Down Variant of Contrasting Groups

The Up and Down method is a way of focusing the participants' judgments where they will do the most good, in the area of the score scale where the cut score will be (Livingston, 1980; Zieky, Perie, & Livingston, forthcoming).

A problem that often arises in cut score studies is the effort and expense involved in getting the participants' judgments. In a Contrasting Groups study, the cost may depend heavily on the number of judgments of individual test takers. The most useful judgments will be those of test takers whose test scores are fairly near the cut score(s). But until you have collected the judgments, you don't know what part(s) of the score range that will be. This method provides a way to resolve the dilemma. Two conditions are required for this method: (1) the test takers take the test before their skills are judged, and (2) the test takers can be selected for judgment one at a time. There must be some evidence for panelists to review and judge. An example would be a foreign-language speaking test, in which the participants' responses are recorded and available for judgment.

Select a test taker near the point on the score scale where you think the cut score might be. If the first test taker is judged to be clearly below the Proficient level, for example, select the second test taker from a somewhat higher score level. If the second test taker is judged to be clearly above the Proficient level, select the third test taker from a score level that is lower than that of the second test taker, but higher than that of the first test taker. Continue until the borderline of the performance level is found, where it becomes difficult to decide if the test takers are above or below the Proficient level. At that point, we assume that if we had an infinite number of examples of work at this level, about half the test takers will be judged to be Proficient and half will be judged to be below Proficient. The score of that borderline test taker becomes the cut score.

This method can be used with any alternate assessment that produces student work in any form that can be judged. It is particularly well suited for circumstances where the number of samples at any given score point is low.

Body of Work

The Body of Work method is an approach that focuses on categorizing student work rather than the students themselves (Kingston, Kahl, Sweeney, & Bay, 2001). The method is designed for tests with performance questions or tasks that yield observable products of a test taker's work, such as essays or recorded speech or musical performances. The method does not work well for tests that include large numbers of multiple-choice questions, but it will work if there are some multiple-choice questions with the performance questions.

A test taker's responses to all of the questions in a test are placed in a Response Booklet. There is a separate booklet for each test taker. The word b*ooklet* is used very loosely. The Response Booklet could be a booklet of responses to essay questions, or it could be a CD or DVD containing audio or video recordings of a test taker's responses. The Response Booklet could be a portfolio of artwork or a set of x-ray studies. What the Response Booklet must be is a collection of observable responses to (mostly)

performance test questions in a format that the participants can conveniently evaluate during the cut score study.

If the test contains multiple-choice questions, include them in the Response Booklet with the correct answers and the answers the test takers selected. Include the test taker's number correct on the set of multiple-choice questions as well. Inform participants of the weight to be given to the multiple-choice questions in comparison to the weight to be given to the performance questions.

The participant's judgment is based on the *Response Booklet* containing a test taker's responses to all of the questions on the test. The participant makes a single judgment about the entire set of responses in the Response Booklet, matching the knowledge and skill exhibited in the responses to the knowledge and skill required to be in a performance level.

There are usually three iterations of the judgments. The first iteration is a training round, the second is a range-finding round, and the third is a pinpointing round. As the names imply, the range-finding round identifies the part of the score scale in which each cut score lies, and the pinpointing round identifies the actual cut score within that part.

The cut score between two performance levels is chosen by finding the point on the score scale that best distinguishes between the Response Booklets placed in each of the performance levels.

This method is well suited for portfolio assessments or any assessment where the student produces a body of evidence. Rhode Island and New Hampshire are two portfolio states that used the Body of Work method to set cut scores.

Generalized Holistic

In their chapter on standard setting on alternate assessments, Cizek and Bunch (2007) focus on one particular approach to setting cut scores on portfolio assessments that they call a Generalized Holistic method. Virginia used this method on their portfolio assessment.

This method is very similar to the Body of Work, although with a slight variation. In this method panelists review student portfolios (or bodies of evidence) that have been ordered from lowest overall score to highest. They classify each portfolio into one of the performance levels based on their holistic judgment of the entire portfolio. There is only one round of classification, discussion, and adjustment. The cut score is calculated by taking the midpoint of between adjacent category means. That is, the score points for each portfolio assigned to a particular level are averaged for each performance level. For example, Basic=20 points, Proficient=30 points, Advanced=40 points. Then, the midpoint between the two levels is used as the cut score. So, in this example, the cut

score for Proficient would be 25 points, and the cut score for Advanced would be 35 points.

The procedure described in this chapter included a couple of variations. One variation was that not all panelists reviewed every portfolio. Because of time constraints and the amount of material in each portfolio, panelists were only able to review eight to nine portfolios, on average. Each portfolio was reviewed by a minimum of two panelists. The second variation was in the determination of the cut score. Using the method of averaging adjacent categories resulted in two cut scores that were only one point across. At this point, the facilitator asked the panelists to discuss their ratings and then vote on the final cut score. The vote moved those two cut scores five points apart.

This method seems appropriate under the conditions mentioned, short time frames and large amounts of materials to review, but the problems documented highlight the importance of applying a rigorous methodology while allowing for flexibility in the application. Participants need ample time to discuss their judgments and come to agreement.

Hybrid: Rubric with Evidence

A final method that has been used more recently could fit under this category or the previous one, as it combines some of the features of a dominant profile approach with the body of work methodology. That is, panelists first focus on a set of rules that each student must achieve to reach each performance level. The first part of this task is exactly as described in the Dominant Profile Method. Then, once consensus has been reached on the desired minimum profile for each performance level, student sample work is pulled that meets that profile and that is similar to that profile. Then, the second part of the standard-setting workshop is similar to the pinpointing round of a Body of Work approach. The panelists would review each sample portfolio and categorize it as either meeting or not meeting the criteria (or as Basic or Proficient, for example). The results of the analysis of the student work would be used to either confirm or modify the profile rules.

Consider the following example, shown in Table 4, where the panel recommended profile is shown in the first column and the profiles of all the portfolios pulled for examination are shown in the other columns.

Table 4. Example of Analysis for Part Two of the Hybrid Approach

Dimension*	Minimum Profile Rules	Profiles of Ten Sample Portfolios for Analysis									
		A	B	C	D	E	F	G	H	I	J
Accuracy	4 pts	4	3	4	4	4	2	3	3	4	4
Complexity	2 pts	3	3	2	3	3	4	3	3	2	2
Independence	3 pts	3	3	3	2	3	3	4	3	3	2
Generalizability	2 pts	2	3	2	2	2	3	2	2	3	3
Total points	12 pts	12	12	11	11	12	12	12	11	12	11

*Assumes a rubric with five dimensions, each scored 0–5.

In this example, the panelists set a minimum criterion for each dimension and a separate minimum criterion for the total score. If the decision rules were followed exactly, only portfolios A, E, and I would be judged sufficient to be categorized as meeting the criteria.

Kentucky used this method to set cut scores on their portfolio assessment. One benefit of this method is it helps panelists clearly distinguish between compensatory and conjunctive methods of scoring. Using the rules in the example, only three of the ten portfolios would have met the criteria. However, if the whole score cut score was used (without regard to how it was obtained) then six portfolios would have met the criteria. Examining the two portfolios that achieved the same number of total points but through a different combination of scores will help panelists refine the decision rules and clarify what is important for the assessment. Furthermore, examining Portfolio C which met the minimum criteria for each dimension but did not meet the criteria for total points will also help the panelists see the results of their rules. Even if the decision is made to use one cut score on the total score scale, this process can help clarify the meaning of compensatory scoring for panelists by showing them samples of all possible ways of obtaining that total score.

How to Choose a Method

One of the most popular questions in any overview of standard setting is “How do I choose a method?” The answer depends both on the type of assessment and the values of the policymakers. For instance, as described in the methods section above, some methods lend themselves more to certain assessment formats. For example, body of work is used most commonly with portfolio assessments. However, values come into play too, in terms of whether the state policymakers want to report scores in terms of total score, by content standard, or by dimension, and whether they want to weight one standard or dimension more highly than another. Likewise, policymakers will need to decide whether they want their cut scores to be compensatory or conjunctive. That is, should the cut score be set on the total score, or is it important to have a minimum criterion for each dimension of the test?

Overall, the recommendation is to focus on the actual student work product whenever possible. If the test does not lend itself to producing student work, consider videotaping a sample of the students engaged in the assessment tasks. Focusing on the students or their work is considered the preferential approach by many (cf., Zieky, Perie, & Livingston, forthcoming). This approach is embodied in the contrasting groups approach, body of work, the up-down method, the rubric with evidence method, and judgmental policy capturing.

The next most preferable approach is to consider the rubric, such as with the dominant profile or the performance profile approaches. Although the judgments are not made on student sample work, some samples are needed to provide a context for the various profiles.

Finally, if it is not possible to examine student work or rubrics, consider one of the item-level methods. The two types of extended Angoff are preferable in the sense that they ask the panelist to judge the minimal acceptable rubric level for each task, thus considering the interaction between performance and content. The yes-no method might be the best method available for a skills checklist approach, although again, the training should include examples of a student succeeding and failing on each of the tasks.

The methods that require ordering items by difficulty should only be considered if there is evidence that the difficulty ratings of the items are consistent across disability types. If some students follow different learning progressions than others, then the ordering will not make sense for all students. This inconsistency will cause difficulties for teachers working with these different students. If, however, the items order by difficulty similarly for most students, it may be a valid method for tests comprised primarily of items scored right/wrong or including a smaller proportion of items score through a rubric. These methods also work better when the scores are translated onto a scale rather than reported in the raw score method. However, using an IRT-based approach to scale scores requires a minimum number of students; working with a small, diverse population may preclude this type of scaling in some states. One additional caution is the lack of research available on methods such as Bookmark and ID Matching. See, for example, a recent literature review that highlighted the dearth of published research on the Bookmark method (Karantonis & Sireci, 2006).

Finally, it is worth considering combing methods to best meet the various components of some alternate assessments. The hybrid approach described at the end of the previous section is an example of combining two standard methodologies.

Methods Used by States

According to the 2005 survey conducted by NCEO, only 26 of the 50 states used a formal standard setting process to determine cut scores in 2005. Because of the requirements set by the peer review process, we expect the number of states using a formal standard setting process to be 50 by the end of the 2007–08 school year. Table 5 shows the number of states using each method of setting cut scores. In 2005, the most common method selected was the Body of Work method. Note that many states used more than one method to set cut scores. For example, panelists would used the Judgmental Policy Capturing Method to determine the weighting scheme for each component and then pull portfolios that matched that weighting scheme and determine the final cut score using the Body of Work method.

Table 5. Standard-setting Methods that States Apply to Alternate Assessments*

Method	Body of Work	Reasoned Judgment	Bookmarking/ Item Mapping	Contrasting Groups	Judgmental Policy Capturing
2003	2	11	6	3	1
2005	15	10	9	5	5

*States could select more than one method.

In 2007, we expect to see a wider variation in methods applied. The Dominant Profile approach is being used more widely, often in combination with a Body of Work methodology. The extended Angoff method is being used with some performance task approaches, and the Angoff yes/no method has been used to set cut scores on a skills checklist.

VALIDATING ALTERNATE ACHIEVEMENT STANDARDS

Once the achievement standards have been set, they need to be validated. It is important to understand that it is not the achievement standard (or cut score) itself that is valid or invalid, but the interpretation and use of the achievement standards. Just as there are no absolute criteria against which specific cut scores can be evaluated, there are no perfect criteria for evaluating standard setting studies (Kane, 1994, 2001). Even though there are no absolute criteria you still must provide evidence that the cut scores are reasonable and appropriate.

In examining the validity of the use of the achievement standards, it is important to ask a series of questions about the basic components of those standards. For example, questions may include:

- Was the standard setting procedure internally valid?
- Do the cut scores divide students reasonably in terms of achievement?
- Do the effects of the achievement standards match what was intended?
- Were there any unintended consequences for using the performance levels?

To answer these questions requires states to undertake both short-term and long-term studies.

Short-term Approaches

Some of the short-term studies are really more procedural approaches involving appropriate documentation, evaluation, and review. For example to demonstrate that the cut scores were set in a valid manner, the policymakers should ensure

- A documented method was used
- Appropriate training was provided
- All modifications were justified
- All the typical steps were applied
- The panel composition was appropriate

In addition, to examine the reasonableness of the cut score, policymakers should document the variation across panelist ratings from one round to the next. We would expect to see the variance decrease from one round to the next as panelists converge on a cut score. Documenting the relevant discussions among panelists will also provide

a rich source of data regarding why the cut score was placed where it was, any dissention among the panelists, and further evidence of the rationale for adopting the recommended cut score.

A second set of analyses could examine whether or not the impact data make sense. Do they fit expectations? Are they aligned with the goals for the program? For instance, if the assessment was written to brand new content standards that were newly introduced to schools and teachers, policymakers might expect to see a low percentage of students reaching Proficient the first year. Do the impact data support this expectation? Gathering data on expectation from various stakeholder groups and comparing them to the actual impact data will also provide evidence on the reasonableness of the cut scores. Examining impact data across grades also provides additional validity evidence.

Finally, gathering information from the panelists themselves through evaluation forms can provide additional evidence of the procedural validity of the cut scores. The evaluation form should ask questions regarding the clarity of the procedures, adequacy of training, completeness of the materials, and their confidence in the results.

Long-term Approaches

Validation efforts do not end with the aforementioned approaches, it must continue for the longer term. Some of the methods described require following students and/or teachers longitudinally to determine the effects of the alternate achievement standards. Others need to be done a year or two after the standard setting to examine the location of the cut scores and the resulting decision accuracy for the students.

One example of a longitudinal study is one that follows students over time. One simple question that can be answered simply by tracking their progress on the alternate assessment over time is “Do the classifications make sense from one grade to the next?” For example, we might be concerned if the results of the alternate assessment placed the same student in the Proficient category at grade 3, Basic in grade 4, Proficient in grade 5, and Basic in grade 6. Unless that particular student is scoring right at the cut score, it is unlikely that his/her performance is varying to that degree each year. Another question to ask might be “is improvement captured by the performance levels?” If a student is truly making great strides but remaining in one performance level over the years, this may argue for additional performance levels or to add a growth dimension to the assessment program. On the aggregate, tracking trends over time on the percentage of students in each performance level would also provide useful information. Are the same percentages of students scoring in each level each year? Is there movement in one direction or another? How much? In some cases, great care has been taken to articulate the standards across grades so that similar percentages of students are being classified into each performance level. However, five years later, there are large inconsistencies in the percentage of students scoring at proficient or above across grades. Watch for this. If this pattern is emerging, additional studies may

need to be done to determine if changes in curriculum and instruction might warrant adjustments to the cut scores.

Another type of longitudinal study that can be done is to survey teachers and other stakeholders (e.g., parents) over time. To help determine the validity of the use of the standards have and whether they are affecting classrooms in a positive manner, ask questions such as:

- Are the performance levels capturing the true achievement of your student(s)?
- Do you see possibility for improvement for the students (i.e., do you think it's possible for your student(s) to reach proficient with sufficient instruction)?
- Has understanding these alternate achievement standards changed your (or your teacher's) approach to instructing your student(s)? How?
- What suggestions do you have for modifying PLDs and/or cut scores?
- Are you satisfied with results?

Some of these questions touch on the issue of how alternate achievement standards might effect students' opportunity to learn. This idea can be explored further by looking at changes in enrollment in teacher professional development courses over time. Periodic but systematic classroom observations could also look for pedagogical changes or changes in the types of curriculum students are exposed to. Any of these areas can also be explored through teacher focus groups.

It is also important to look for any unintended negative consequences. Are certain content standards being de-emphasized in the classroom because they are not assessed? Are certain students being neglected because they are so far away from the proficient cut score that the teachers are choosing to work more with the students closer to that mark? Is teacher retention decreasing? This last question can be answered by looking at school or district records. The first two will require either anonymous surveys or classroom observation (or both).

In addition to examining trends over time, snapshot studies at various points also will inform the validity of the alternate achievement standards. Consider, for instance, the scenario given above where drift in performance has resulted in a loss of articulation of results across grades. If three years after the cut scores are set, 60% of kids are scoring Proficient or above at grade 3, 70% in grade 4, 65% in grade 5, 35% in grade 6, 45% in grade 7 and 25% in grade 8, what does this mean? Are the cut scores at grades 6 and 8 too strict? Are they too lenient in grade 4? Or are there other changes happening in the classroom that may explain this pattern of scores?

One option is to conduct a contrasting groups study with a sample of teachers and students. Ask 100 teachers (per grade and subject) in your state to read the performance level descriptors carefully and then place each student in your class (who takes the alternate assessment) into one of those three categories based on which

descriptors best matches that student's knowledge and skills. This part of the study should be conducted BEFORE the student takes the assessment, but as close to the assessment date as possible. Then, the students' scores and resulting achievement levels on the assessment can be added to the teacher classifications for a complete data set. There are a couple of analyses that can be conducted with these data. First, the cut scores can be recalculated using the contrasting groups methodology (see Zieky, Perie, and Livingston, forthcoming, for instructions). Then, these cut scores can be compared to the ones currently in use. Another analysis could be a misclassification analysis that compares the teacher's classification of the student to the assessment's classification of the student. For either analysis, ask how different are the cut scores and classifications? Do they appear to go in a certain direction by grade (teacher always classifies the student in a higher/lower category)? Note that there is some research that suggests that this type of contrasting groups study typically will result in a lower cut score than other studies where teachers are judging the work of students they do not know. However, if you find different degrees of misclassification across grades, that may provide further evidence that the cut scores need to be revisited.

Finally, consider using external measures to validate the alternate achievement standards. That is, compare other types of evaluations that rate the students with the alternate achievement standards. For instance, compare the grades students receive in ELA and math to the performance levels student are placed in for ELA and math. If the teacher evaluations of these students are incongruous with the achievement level classification, the validity of the alternate achievement standards may be called into question.

DOCUMENTING THE PROCESS OF DEVELOPING ALTERNATE ACHIEVEMENT STANDARDS

An often overlooked but critical component of standard setting is documenting the process. Two professional *Standards* (AERA, APA, & NCME) directly address the importance of documenting the rationale, procedures, and results:

- Document the PLDs, selection of panelists, training provided, ratings, and variance measures (Standard 1.7)
- Document the rationale and procedures for the methodology used (Standard 4.19)

Recent work has focused on the technical documentation of alternate assessments (Marion & Pellegrino, 2006). In the recommendation by Marion and Pellegrino (ibid, p. 56), standard setting should be one chapter within the "nuts and bolts" of alternate assessment technical documentation. They divide the standard setting chapter into two sections: methodology and results. Although their recommendations are probably sufficient for a complete technical documentation manual, it is useful to have a stand-alone standard setting report that includes full detail of all parts of the procedures, results, and follow-up work.

There are eight important components that should be documented after the standard setting process:

1. Performance level descriptors
2. Panelists
3. Rationale
4. Training
5. Procedures
6. Ratings and variance
7. Any adjustments and adoption of cut scores
8. Validity evaluation

Each of these components will be discussed in the following sections.

Document the PLDs

It is important not only to include the performance level descriptors (PLDs) in the technical documentation but also to document their development. The documentation should answer questions about who developed them, when they were developed, and what procedures were followed. If they were developed by committee, what instructions were given to the committee? Any reviews and adjustments done by policymakers should be documented as well.

The documentation should demonstrate that the alternate assessment PLDs

1. Represent highest standard possible for this population
2. Capture essential skills
3. Align with state content standards
4. Progress logically across levels (e.g., is Proficient appropriately higher than Basic)
5. Progress logically across grade levels (e.g., is Grade 5 proficient sufficiently more advanced than grade 3 proficient)
6. Represent knowledge and skills that are evaluated by the assessment (e.g., don't discuss independence in the PLD if your assessment doesn't measure independence)

If these PLDs were developed for assessments required under NCLB, the PLDs also must be approved formally by the state policymakers. That approval process should also be documented.

Document the Panelists' Qualifications

Most technical documentation includes a section that describes the panelists involved in the standard-setting workshop. This documentation typically includes a table listing the number of panelists and summarizing characteristics such gender, race/ethnicity, and years of experience. Other information that should be documented includes the panelist's experience, e.g., content expert, special education teacher, higher education

representative, or parent. It also is important to discuss the representation of the state in terms of geographic and demographic diversity.

Equally as important as documenting the characteristics of panelists at the standard-setting process is documenting the process by which they were selected. The panelists should represent the pool of people qualified to be panelists. Provide evidence that the pool of potential panelists was determined in a manner that best represents the constituency of the state. Were all teachers invited to participate? If not, how were they selected? Did principals nominate teachers? Did the state department of education include only people they knew? Then, the documentation should describe the process by which potential panelists were contacted, the information given to them, how the final panel came to be, and how well it represents all potential panelists in the state.

Document the Rationale

This section needs to defend the selection of methodology. It should answer the question: Why did you choose the standard-setting methodology you did? To answer this question, first consider the characteristics of the alternate assessment. Does it include performance items? A checklist? A full body of evidence? Were rubrics used to score part or all of the assessment?

Next, discuss the types of judgments made and how they fit into the model for the alternate assessment. For instance, was each item considered individually or were holistic judgments made? Was the rubric an important part of the process? If the theory behind the assessment includes a discussion of the importance of the various dimensions of a rubric, were these dimensions considered separately in the standard-setting process? Discuss the importance of compensatory or conjunctive standards and how these values were communicated through the selection of methodology.

Document any logistical considerations such as time constraints, panelist accessibility, availability of student data or sample work. Any of these considerations could have implications for the methodology chosen. The important piece of this section is to explain how the methodology fits the model of the assessment while acknowledge its constraints.

Document the Training

The panelist training is arguably the most important part of the workshop. Therefore, it should be recorded thoroughly in the documentation. Each step of the training should be discussed from familiarizing the panelists with the content standards, assessment and PLDs to training on the standard setting methodology and providing and opportunity for practice. This section should also include a summary of the evaluation form given to panelists at the end of the practice session. It should provide evidence that the panelists felt sufficiently trained and were prepared to continue with the process.

Document the Procedures

The number one rule of documenting the procedures is that it should be of sufficient detail to allow someone else to replicate the study. Include the exact instructions that were given to panelists. Explain how each round differed in terms of the judgmental task, feedback given, and discussion allowed. Describe the types of data given to panelists at each round.

If possible, describe the focus of the group discussions. Record any major issues or concerns of panelists as well as any driving factors in their judgments. For example, were there any particular items, student work, or parts of the rubric that caused a particular challenge for the panelists? If so, how did they resolve their concerns? Document any problems that arose during the workshop and how they were resolved. Use evaluations effectively and record the results.

Document the Ratings and Variance

The panel-recommended cut scores should be recorded and reported at the end of each round. The central tendency (median, mean, or trimmed mean) should be recorded along with a measure of the variance. At the least, provide the range of the judgments, meaning the minimum and maximum. Record the full distribution of panelist recommendations for the final round. Provide a standard error of the variance in panelist judgments for each round to document any convergence that occurred across rounds.

The easiest way to convey this information is to develop summary tables showing results across rounds, including means (or medians), minimums, maximums, and standard errors of judgment.

There should be 1–2 final summary tables, depending on whether or not the results were adjusted after the conclusion of the workshop. These tables should include the recommended and/or adopted cut scores along with the resulting impact data. It also may be important to provide comparative information, such as impact data for other subjects or grades not included in the study.

Document the Process for Adopting the Cut Scores

After the standard setting workshop is complete, the PLDs and cut scores must be adopted. Document any subsequent reviews and other procedures conducted after the workshop concluded. Discuss any adjustments made to results such as smoothing across subjects or grades. Provide tables of the final results including the relevant impact data. Record how the final decision was made on which cut scores to adopt. Include answers questions such as to

1. Who adopted them?

2. What data did they consider?
3. What was the rationale for any changes made?

Other Considerations

Ideally, we should also document the study coordinator's and facilitators' qualifications. The facilitators should have experience running workshops, understand the alternate assessment, have a strong familiarity with the methodology and have a facility with the data used throughout the workshop. If multiple facilitators are working in several rooms simultaneously (to cover multiple subjects, for instance), they should be working from a script to ensure consistency across facilitators. If that is the case, the script should be included in the documentation. If a full script is not used, the actual instructions given to panelists regarding the judgmental task should be recorded and said the same way in all sessions. These instructions should be recorded. At a minimum, document the training given to the facilitators prior to the workshop.

In addition, if student sample work was used in the workshop, include a section in the documentation that describes how that work was selected. If impact data are included, record where the data came from and how they were analyzed for use in the workshop. Include blank rating forms, evaluation forms, and the agenda in appendices.

Document the Validity of the Interpretation of the Alternate Achievement Standards

Finally, the documentation should conclude with a section on validity. All evaluation forms should be summarized and presented to demonstrate the internal validity of the procedure. This section should also include a discussion of procedural validity, describing how the procedures followed in the standard-setting workshop contribute to the validity of the interpretation of the alternate achievement standards.

Again, a good validity section will describe how the rationale for the entire alternate assessment program was carried into the standard setting procedures. Document how the choice of methodologies, evidence presented, and decisions made match the purpose the assessment is intended to serve. Finally, document any validity studies that were completed or are planned for a future date.

OTHER ISSUES

Other issues that are emerging with alternate assessments include setting modified achievement standards for the two percent population and setting achievement standards on growth scales.

Modified Achievement Standards

In April 2007, the federal government released additional Title I regulations providing an exception for allowing an additional 2% scores for students with disabilities to count toward proficiency determination for Annual Yearly Progress (AYP). These proficient

scores must be derived from assessments based on modified achievement standards (U.S. Department of Education, 2007). This new allowance raises the question of how to set modified achievement standards that are aligned to grade level content standards in terms of depth, breadth and complexity. An important a priori consideration in the design of such an assessment is a comprehensive description of the population and the trajectory of learning in academic content domains. In most cases, performance data do exist for this population in terms of assessment results. Student descriptions will be essential in the writing of performance level descriptors, and the performance level descriptors will be essential to setting cut scores.

Little else will change in developing modified achievement standards. The same steps described in this paper for alternate achievement standards should be followed for modified achievement standards. The same methodologies are available to use in developing cut scores. The tests themselves are likely to look quite different, meaning a different methodology may be used to set the modified cut scores than was used to set the alternate cut scores. It may, in fact, make sense to use the same methodology used to set grade-level achievement standards for the general assessment. Regardless, the same considerations in choosing a method still apply. The modified achievement standards will also need to be validated and documented in the same way as the alternate achievement standards.

Growth

Ultimately, in looking at growth standards, we are asking ourselves “if a student has been taught well, i.e., received quality instruction, how much progress would we expect to see by the end of the school year?” Then, proficient growth would be defined as that amount of progress. The challenge lies in determine how much progress can be reasonably expected of a student in one school year. Determining appropriate methodologies for answering these questions is still being researched for the general population. There are additional challenges in answering this question for the students with significant cognitive disabilities. We know less about learning progressions or whether there is an interaction between the type or degree of cognitive disability and the learning progression.

One possibility for exploring the idea of growth with this population is to collect evidence of student work at different points during the school year. Experts, both in content and in this population, could then examine that evidence and evaluate the degree of growth it shows. The evidence could be examined for students with different disability types or different severities of cognitive disabilities. This exercise would allow us to qualitatively describe the types of growth we see for each content area, grade span, and disability type or degree. The next step would be to quantify it. How this evidence is quantified will depend heavily on the type of assessment, the scoring mechanism, and scaling techniques. Once we can represent the qualitative statements of growth onto a quantitative scale, determining an appropriate cut score becomes a matter of judgment. Panelists would go back to the evidence of student work collected

throughout the year, and the descriptions and determine a minimum level of growth for each performance level. Then, that level of growth could be mapped onto the scale to determine the cut scores.

CONCLUSION

The bottom line is that the process of developing alternate achievement standards has much in common with the process of developing general achievement standards. The necessary components stay the same, the steps are identical, and the detail of documentation required should be equivalent. The primary differences are in the type of assessment and the low numbers of students in the population. However, these differences do not imply that a lower level of rigor is acceptable in the implementation of standard setting procedures for alternate achievement standards compared to general achievement standards.

The key components of the standard-setting process include first writing clear performance level descriptors linked to the grade-level content standards and describing the supports and contexts appropriate for students with significant cognitive disabilities. A second component is choosing an appropriate standard-setting methodology by matching the features of the alternate assessment to the judgmental task of the method. A third component is validating the alternate achievement standards, both in terms of the process of establishing them and the interpretation and use of them. The final component involves the documentation—writing complete descriptions and justifications for each step of the process, following the same rigor required for the general assessment.

Ultimately the alternate achievement standards should be grounded in student work but provide educators with incentives for continuing to teach these students grade-level curriculum. Developing strong alternate achievement standards can help move the education process out of status quo and provide additional opportunities for students with significant cognitive disabilities.

REFERENCES

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

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.) *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Arnold, N. (2003). *Washington Alternate Assessment System technical report on standard setting for the 2002 portfolio* (Synthesis Report 52). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 16, 2006 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis52.html.

Browder, D. M., Spooner, F., Ahlgrim-Delzell, L., Flowers, C., Karvonen, M., & Algozzine, R. (2003). A content analysis of the curricular philosophies reflected in states' alternate assessment performance indicators. *Research and Practice for Persons with Severe Disabilities*, *28*, 165–181.

Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Erlbaum.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Ferrara, S., Perie, M., & Johnson, E. (2007). Matching the judgmental task with standard setting panelist expertise: The Item-Descriptor (ID) Matching procedure. *Journal of Applied Testing Technology*.

Gong, B., & Marion, S. (2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities* (Synthesis Report 60). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved May 16, 2007, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis60.html.

Hambleton, R. K. & Pitoniak, M. J. (2006). Setting performance standards. In R.L. Brennan (Ed.). *Educational Measurement*. Westport, CT: Praeger.

Hambleton, R. K. & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*, 41–56.

Impara, J. C. & Plake, B. S. (Eds.). (1995). Standard Setting for Complex Performance Tasks [Special issue]. *Applied Measurement in Education*, *8* (1).

Impara, J. C. & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*, 353–366.

Jaeger, R. M. (1995) Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education, 8*, 15-40.

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

Karantonis, A. & Sireci, S. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25* (1), p. 4-12.

Kingston, N.M., Kahl, S.R., Sweeney, K.P., & Bay, L. (2001). Setting performance standards using the body of work method. In G.J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kleinert, H. L., & Kearns, J. F. (2001). *Alternate assessment: Measuring outcomes and supports for students with disabilities*. Baltimore, Maryland: Brookes Publishing.

Livingston, S. A. (1980). Choosing minimum passing scores by stochastic approximation techniques. *Educational and Psychological Measurement, v. 40,* no. 4, pp. 859-873.

Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Marion, S. F. & Pellegrino. J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25* (4), p. 47–57.

Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I*, (pp. 73–85). Washington, DC: Council of Chief State School Officers.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The Bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Morgan, D.L. (June 2004). *The performance profile method (PPM): A unique standard setting method as applied to a unique population*. Presented at the annual meeting of the Council of Chief State School Officers in Boston, Massachusetts.

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

Olson, B., Mead, R., & Payne, D. (2002). *A report of a standard setting method for alternate assessments for students with significant disabilities* (Synthesis Report 47). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 16, 2006 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis47.html.

Perie, M. (2006). *Convening an Articulation Panel after a Standard Setting Meeting: A How-To Guide.* National Center for the Improvement of Educational Assessment. Dover, NH: NCIEA.

Perie, M. (2007). *A guide to understanding and developing performance level descriptors*. Dover, NH: National Center for the Improvement of Educational Assessment.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard setting method For performance assessments: The Dominant Profile Judgment method and some field-test results. *Educational and Psychological Measurement, 57*, 400–411.

Quenemoen, R. Rigney, S., & Thurlow, M. (2002). *Use of alternate assessment results in report and accountability systems* (Synthesis Report 43). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 16, 2006 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis43.html.

Quenemoen, R., Thompson, S. & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring criteria*(Synthesis Report 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved May 16, 2007, from the World Wide Web: http://education.umn.edu/NCEO/OnlinePubs/Synthesis50.html.

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 16, 2006 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html.

Thurlow, M.L., & Ysseldyke, J. R.. (2001). Standard-setting challenges for special populations. In G.J. Cizek (Ed.) *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

U.S. Department of Education, Office of Elementary and Secondary Education. (April 9, 2007). *Title I--Improving the Academic Achievement of the Disadvantaged; Individuals With Disabilities Education Act (IDEA); Final Regulations*. Washington, DC: U.S. Department of Education.

U.S. Department of Education, Office of Elementary and Secondary Education. (2004). *Standards and assessments peer review guidance: Information and examples for*

meeting requirements of the No Child Left Behind Act of 2001. Washington, DC: U.S. Department of Education.

Wiener, D. (2002) Massachusetts: *One state's approach to setting performance levels on the alternate assessment* (Synthesis Report 48). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved August 16, 2006 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis48.html.

Zieky, M., Perie, M., & Livingston, S. (forthcoming). *Cutscores: A manual for setting performance standards on educational and occupational tests*. Princeton, NJ: Educational Testing Service.