

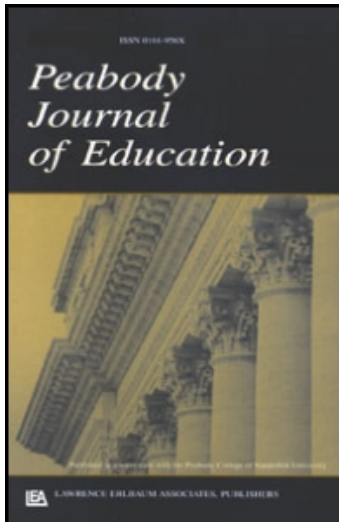
This article was downloaded by: [Peabody College Vanderbilt University]

On: 10 November 2009

Access details: Access Details: [subscription number 776124334]

Publisher Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Peabody Journal of Education

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t775653692>

Writing Performance Level Descriptors and Setting Performance Standards for Assessments of Modified Achievement Standards: The Role of Innovation and Importance of Following Conventional Practice

Karla L. Egan ^a; Steve Ferrara ^a; M. Christina Schneider ^a; Karen E. Barton ^a

^a CTB/McGraw-Hill,

To cite this Article Egan, Karla L., Ferrara, Steve, Schneider, M. Christina and Barton, Karen E. 'Writing Performance Level Descriptors and Setting Performance Standards for Assessments of Modified Achievement Standards: The Role of Innovation and Importance of Following Conventional Practice', Peabody Journal of Education, 84: 4, 552 – 577

To link to this Article: DOI: 10.1080/01619560903241028

URL: <http://dx.doi.org/10.1080/01619560903241028>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Writing Performance Level Descriptors and Setting Performance Standards for Assessments of Modified Achievement Standards: The Role of Innovation and Importance of Following Conventional Practice

Karla L. Egan, Steve Ferrara, M. Christina Schneider, and Karen E. Barton

CTB/McGraw-Hill

Alternate assessments of modified academic achievement standards (AA-MAS) must be designed, developed, implemented, and validated following the same rigorous principles and procedures used for other assessments. However, the uniqueness and unfamiliarity of the target population for these assessments requires innovative thinking, especially in regard to writing relevant performance level descriptors (PLDs) and setting appropriate achievement standards. In this article we discuss considerations for (a) likely designs of AA-MAS; (b) the writing of PLDs that are relevant to the target examinee population for these assessments; and (c) standard setting methods that would be appropriate for these assessment designs, as well as special considerations relevant to those standard setting methods.

Alternate assessments of modified academic achievement standards (AA-MAS) require both conventional and innovative education achievement testing practices. On one hand, designing, developing, implementing, and validating AA-MAS should follow the same technical procedures and meet the same technical standards as other grade-level assessments. This is both a commonsense requirement and a stipulation in the *Standards and Assessment Peer Review Guidance* (U.S. Department of Education [USDE], 2007b). On the other hand, achievement test designers and psychometricians have little familiarity with the intended target examinee population. Their limited experience with students with mild and moderate disabilities—the students we think are most likely to be appropriately eligible for AA-MAS—comes from deliberations about providing test administration accommodations on grade-level assessments. Test designers and psychometricians have had little direct and focused experience with who these students are, what they are studying, how they learn, and the disabilities that create their persistent learning difficulties. Federal guidelines do not define the target population clearly, although they stipulate that AA-MAS

We thank Adele Brandstrom for editing this manuscript. We also thank the reviewers for their comments, which have improved the final product.

Authors Karla L. Egan and Steve Ferrara are randomly ordered.

Correspondence should be sent to Karla Egan, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA, 93940.
E-mail: karla.egan@ctb.com

extend only to students with disabilities. We do know that as a group these students have a history of poor performance on large-scale assessments.

State assessment program practice in defining the target examinees and developing AA-MAS suggests that we have a long way to go. For example, in March 2008 six states submitted evidence on their AA-MAS for Title I peer review. No state met all requirements. Further, four general shortcomings appeared across all of the state submissions (Filbin, 2008). Two of them are directly relevant to this article:

1. Identification of the target examinees for AA-MAS.
2. Clarification of the relationships among AA-MAS, alternate assessments of alternate achievement standards, and the corresponding grade-level assessments.

In this article we describe five steps for designing, developing, implementing, and validating AA-MAS. We describe these steps to reinforce our point that standard procedures for designing, developing, implementing, and validating all educational achievement tests apply equally rigorously to AA-MAS. However, these assessments are unique, as are the students who are intended to be eligible for them. Both are somewhat unfamiliar to our field. For these reasons, we focus on considerations for (a) possible designs for these assessments; (b) the writing of performance level descriptors (PLDs) that are relevant to the target examinee population and content standards for these assessments; and (c) standard setting methods that would be appropriate for these assessment designs.

STEPS FOR DESIGNING, DEVELOPING, IMPLEMENTING, AND VALIDATING AA-MAS

The five steps we describe are consistent with conventional educational achievement test practice with one exception: defining the target examinee. Usually, defining the target examinee population is as simple as defining the grade levels for which achievement tests are intended; however, that is not currently the case with AA-MAS. The five steps are briefly listed and then discussed in separate sections of this article.

1. *Define the target examinee population and the academic content standards they are pursuing.* This actually requires three steps: (a) defining the target examinee population, (b) describing the grade-level academic content standards that the target examinees *are* pursuing, and (c) describing the grade-level academic content standards that the target examinees *should be* pursuing. We address these three steps together because defining the population involves, in part, examining the focus of students' instruction and the ways that instruction is adjusted for students' individual needs.
2. *Design, develop, and implement the test.* Whether the design approach is to modify items from an existing grade-level assessment, develop new items from grade-level test blueprints, or use other design approaches (e.g., assessment portfolios), items and test forms must be closely aligned to grade-level content standards and grade-level test blueprints, provide test administration environments that are guided by policies for granting appropriate accommodations,

and enable valid inferences about what students know and can do in relation to grade-level content standards.

3. *Develop appropriate PLDs.* These descriptors should be specific to the target examinee population. They define levels of performance, including the all-important Proficient level. They serve as a policy statement of a state's goals for student performance in relation to standards and represent aspirations for the academic achievement of a segment of the school children of a state.
4. *Set performance standards.* We discuss different standard setting methodologies and the application of these methodologies by the type of test being administered. Some standard setting methods are more appropriate for multiple-choice tests, some for tests comprised of both multiple-choice and constructed-response items, and some for portfolio assessments. For each method, we address challenges to adapting them for AA-MAS.
5. *Conduct validation studies.* It is important to collect validation data throughout the test development process. We discuss ideas for collecting data external to AA-MAS for validation purposes.

THE TARGET EXAMINEE POPULATION

The most recent version of the *Standards and Assessment Peer Review Guidance* (USDE, 2007b) provides guidance on defining this unique student population. The guidance suggests that these are students with disabilities who may achieve "significant growth" in grade-level academic content but who, despite "appropriate instruction . . . will not achieve grade-level proficiency within the [school year]" (p. 3). For these students, neither grade-level achievement standards and assessments nor assessments based on alternate achievement standards are appropriate. Although these students pursue statewide content standards at their grade level, they are not expected to master the volume, scope, breadth, and perhaps depth of content and skills that their grade-level peers are expected to master during a school year. They have been referred to colloquially as "gap" students because expectations for their academic achievement is below that of their peers in the general population and above that of students with significant cognitive disabilities who participate in alternate assessments based on alternate achievement standards (e.g., Ferrara, Swaffield, & Mueller, 2009).

Defining this population of students is an evolving inferential, conceptual, and empirical enterprise. Theoretically, many or most of the students who should be eligible for AA-MAS would have been diagnosed with moderate or severe learning disabilities or even mild mental retardation, and this diagnosis would be cited in their Individualized Education Programs (IEPs). Further complicating matters, it is well known that standardization and consistency in the identification of learning disabilities is questionable; there may well be students without disabilities for whom an assessment of modified academic achievement standards would be most appropriate. Also, some students who begin their school careers participating in assessments of *alternate* achievement standards may, through effective instruction, develop academic knowledge and skills sufficient for them to "graduate" to AA-MAS. From a conceptual point of view, students who may be eligible for AA-MAS function academically and achieve mastery of grade-level content and skills at levels above students who are appropriately eligible for assessments of alternate achievement standards and below students who are appropriately eligible for other

grade-level assessments. Thus, one approach to defining this population may be to characterize the highest levels of achievement of students with significant cognitive disabilities and the lowest levels of achievement of disabled students (and perhaps nondisabled students in the general population).

Some researchers have begun to define the population using empirical approaches. For example, Georgia defined “persistently low performing” students (Fincher, 2008) as those students who achieved the lowest performance level on the statewide reading or mathematics grade-level assessments for three consecutive annual administrations. This definition identified 3 to 4% of fifth graders in one or the other content area, 4% of eighth graders in reading, and 9% of eighth graders in mathematics. A subsequent teacher group, after reviewing modified items that had been pilot tested for potential use in an AA-MAS and other relevant information, identified several characteristics of persistently low-performing students. They suggested that students in this group tend to be passive learners, display metacognitive deficits (e.g., they do not generalize skills and concepts to new situations and cannot change topics easily), and display limited vocabulary and prior knowledge. Although this information is valuable for defining IEP goals, providing instruction, and defining target examinees, it provides only limited guidance for identifying item formats that may be problematic for target examinees, designing AA-MAS that are aligned with grade-level content standards, writing performance level descriptors, and setting modified achievement standards.

In another example of empirical work, a consortium of the Minnesota, Ohio, and Oregon assessment programs and the American Institutes for Research has taken a similar approach to identifying persistently low performing students. They defined persistently low-performing students as in the Georgia study. They then evaluated items with acceptable and unacceptable psychometric characteristics for persistently low-performing students and conducted focus groups with teachers to identify learning and performance characteristics of students who should be eligible for AA-MAS. In addition, they have written standards, instruction, achievement, and assessment (SIAA) profiles (Ferrara & Wright, 2007) of students to define the range of achievement and instructional focus for students who they intend to be eligible for AA-MAS. These SIAA profiles include the instructional goals, learning activities, and academic achievement of the intended range of eligible students. Each SIAA profile addresses social and interpersonal skills; instructional goals, academic support, and academic achievement; and appropriate participation in assessments, accommodations, and modifications. The general strategy defines and describes the learning goals and achievement of students at the top and bottom of the range of students who should be eligible for AA-MAS, and the range of students in between.

From a practical or teaching practice point of view, the content of students’ instructional programs may be the best indicator of the need to participate in AA-MAS. Students for whom AA-MAS are intended do not simply perform poorly in school and on grade-level assessments. AA-MAS target students with disabilities who may or may not be striving in school—but who are hindered by persistent learning difficulties when they do strive. Modifications to learning expectations and test items are intended to provide these students access and opportunity to achieve grade-level academic content. At least for those students whose instruction is individually and appropriately tailored through special education services, the focus in instruction should be on grade-level academic content standards where expectations for mastering that material is adjusted as needed and on participation in AA-MAS.

What Gets Modified?

Speculations and confusion about this question abounded (e.g., Filbin, 2008, p. 8) until peer review guidance was updated to address AA-MAS in December 2007 (see USDE, 2007b). The answer is quite clear: The academic content for students in this population is defined by grade-level content standards. Those grade-level content standards *cannot* be modified. The updated guidelines are quite explicit about this point: “Challenging academic standards . . . [must] be the same academic standards that the State applies to all public schools and public school students in the State [and] include the same knowledge, skills, and levels of achievement expected of all students” (USDE, 2007b, p. 2).

Expectations for how much of the grade-level knowledge, skills, and understandings students are likely to master in a school year *can* be modified because it is appropriate and realistic to do so. It follows logically that the achievement standards established by the cut scores on these assessments and the “descriptions of the competencies associated with [each] achievement level” (USDE, 2007b, p. 3) can and should be different from the grade-level achievement standards for the grade-level assessments.

It is interesting to think about defining achievement standards for an entire subgroup of special education students. The bedrock philosophy and principle for delivery of services for students with disabilities in special education is individualization. However, the development of modified achievement standards and assessments for a group of students with disabilities does not have to be viewed as a philosophical shift. Rather, it is public illumination of what academic content this distinguishable group of students should be taught and should be learning, and how much of that content knowledge, skills, and understandings they should be expected to learn each school year.

Likewise, it follows that the content and design of AA-MAS *can and should be modified*. We address this topic next.

POSSIBLE DESIGNS FOR AA-MAS

Designing an assessment system for the target examinees for AA-MAS continues to be a challenge for states. The Federal nonregulatory guidance (USDE, 2007a) provides approaches that states might consider to meet the needs of students in this group. The nonregulatory guidance requires states to assure that the assessments are built following commonly accepted test development processes, provide accessible tests, and ensure reliable and interpretable results. The guidance also states that the content standards used for AA-MAS are to be the same grade-level content standards taught and tested in grade-level assessments.

States can operationalize this guidance in different ways when they define achievement standards and test designs. States may choose to create modified academic achievement standards that are disconnected from grade-level achievement standards or may link to them conceptually via a standard setting process. The guidance requires states to assure that modified academic achievement standards are appropriate and meaningful for target examinees. States must include at least three performance levels (e.g., Basic, Proficient, and Advanced) and the number of levels does not have to be the same number of levels as for the corresponding grade-level assessment. In addition, states cannot simply set lower cut scores on grade-level assessments. The guidance does not stipulate test designs. States may include multiple-choice items, constructed-response items,

portfolio designs, and other approaches. States may also choose to develop AA-MAS by modifying grade-level assessments or by developing entirely new assessments. States are required to ensure that blueprints for AA-MAS are comparable to blueprints for the grade-level assessments (Rigney, 2008; USDE, 2007b, pp. 25–26)

Current Practices

It appears that, so far, states are not choosing to create completely new assessments (Lazarus, Thurlow, Christensen, & Cormier, 2007). Some states are modifying existing grade-level test item pools or intact test forms. In line with the examples provided in the nonregulatory guidance, current approaches to modifying grade-level assessments include eliminating an incorrect answer choice in multiple-choice items, providing shorter reading passages or segmenting them, using plain English principles to simplify items, and/or requiring fewer test items.

Given the current economic climate and the erosion of the tax base in many states, the modification of a state's current grade-level assessment may prove a popular option as states seek affordable ways to develop AA-MAS. For these states, AA-MAS with item types that may be machine scanned (e.g., multiple-choice items) as opposed to open-ended items and portfolios that require human scorers may also be preferred. Although some states may simply opt out of creating AA-MAS, the political realities may force other states to develop AA-MAS.

A report by the National Center on Educational Outcomes (Lazarus et al., 2007) about the six states that reported in 2007 that they had developed or were developing AA-MAS, found the following:

- One state used a portfolio approach.
- The other five states used a traditional testing approach and modified their grade-level assessments:
 - All five states used multiple-choice items, two also included writing prompts, and one included constructed-response items.
 - Some states included items only at the lowest two depths of knowledge levels; most states removed an incorrect answer choice, used fewer test items and simplified test language, and/or used fewer and shorter reading passages (Lazarus et al., 2007).

Other states have focused on improving the accessibility of their grade-level assessments (e.g., using plain English criteria) rather than creating new tests or modifying existing items. For example, Colorado has done extensive work to increase the accessibility and alignment with universal design principles of their existing grade-level assessments and ensuring that their accommodation guidelines and implementation of accommodations in the field are maximized and appropriate (see HB 05-1246 Study Committee, 2005). This falls directly in line with the nonregulatory guidance (USDE, 2007a). Other modification approaches are not consistent with the nonregulatory guidance. For example, selecting reading passages that are not on grade level is considered testing off grade and not acceptable. We note that the determination of what is “grade level” is an oft varied and complex task. Minimizing the breadth of standards covered on an AA-MAS is considered misalignment and not acceptable.

Technical Quality

Whether states choose to retool existing assessments (a seemingly economical solution) or create new assessments, AA-MAS must meet conventional standards for rigor and technical quality. They must be aligned to grade-level content standards, the accompanying modified achievement standards must be well articulated, and both must enable reliable and valid inferences about the academic achievement of grade-level content standards for eligible students.

Current practices in test alignment and standard setting provide guidance on the length of AA-MAS. For example, one might consider no fewer than six items per content standard (Webb, 2006) and as few as 10 items or as many as 20 per achievement level. Test score reliability is a consideration for test length as well. States should consider the amount of testing time that students will need to respond to a full test form, provide sufficient breaks particularly if the test is long, and consider the impact on score reliability of shortening or lengthening tests.

States also will need to consider whether item designs are appropriately accessible for target examinees. For example, multiple-choice items that require multiple steps to achieve a solution may be more appropriately presented as multipart items on AA-MAS (unless of course the item is intended to measure multistep problem solving). These considerations may require different proportions of item types on AA-MAS compared to the proportions on the regular grade-level assessments, and rightly so. As we begin to understand the population, we may find that a distinguishing factor between students taking AA-MAS and students taking regular grade-level assessments is that AA-MAS students possess content knowledge but are less proficient at problem solving, thinking critically, and reasoning than students taking the regular grade-level assessments, as suggested by the Georgia study (Fincher, 2008).

Until new AA-MAS are actually administered to the target examinees, it is difficult to make assumptions about how items and test forms will perform psychometrically. For example, in the design approach involving elimination of a multiple-choice distractor, it is important to review data on the grade-level assessment based on students presumed to fall into the target group and find which distractors are most troubling or least useful for that group, as opposed to looking at distractor data only for the full examinee population (Barton, 2006). States should ensure that the final version of the assessment follows standard steps and procedures for test development, adapted for AA-MAS, as discussed previously.

WRITING PLDS FOR AA-MAS

The PLDs¹ articulate the academic knowledge and skills that students in a particular achievement level are expected to be able to demonstrate (Lewis & Green, 1997; Perie, 2008). As such, PLDs inform students, their families, educators, and the general public about the types of knowledge, skills, and abilities expected of, say, Proficient students in a particular content area. This straightforward purpose of PLDs obscures the controversy surrounding their development. Some researchers assert that PLDs must be developed at the beginning of the test-development cycle (Loomis, 2001; Perie, 2008), whereas others argue that PLDs should only be written following the

¹Also known as achievement level descriptions in the standard setting literature and peer review guidance. We use the terms achievement level and performance level interchangeably.

completion of the standard setting (Lewis & Green, 1997). Peer Review guidance for NCLB has settled this point in K-12 testing, requiring that the initial PLDs are developed prior to standard setting. We contend that both processes for developing PLDs have merit and can be used to the benefit of the standard setting and the testing program (Mercado & Egan, 2005).

By developing PLDs at the front end of the test development cycle, items can be written that target and align with the descriptors (e.g., Bejar, Braun, & Tannenbaum, 2007). These initial PLDs should be developed to provide the state's expectations for examinee performance. For standard setting, the initial PLDs can be used to guide standard setting panelists. The cut scores from a standard setting are an operationalization of the state's PLDs. By refining and validating those initial PLDs after the cut scores have been established, we can be certain that the final PLDs reflect actual student knowledge and performance.

In the next section, we discuss the development of the PLDs for AA-MAS; specifically, we discuss four aspects that states should consider when developing PLDs in the context of AA-MAS: (a) the relationship to the grade-level assessment, (b) policy-based PLDs, (c) test-based PLDs, and (d) final PLDs.

Relationship Between PLDs for AA-MAS and Grade-Level Assessments

In developing the PLDs for AA-MAS, the first step must be to consider the relationship between the PLDs for grade-level assessments and the ones to be developed for AA-MAS. The USDE (2007a) nonregulatory guidance allows for a different definition of Proficient on the AA-MAS even though the same grade-level content standards must be used for both the grade-level assessments and AA-MAS. According to the *Standards and Assessment Peer Review Guidance* (2007b), "modified academic achievement standards must—be challenging for eligible students, but may be less difficult than the grade-level academic achievement standards" (p. 17). Given that AA-MAS are meant to be a less difficult version of a state's grade-level assessments, it is only logical that the AA-MAS proficiency definition will reflect that decrease in difficulty. In other words, students taking the AA-MAS will most likely be expected to master less grade-level content than students taking the regular grade-level assessments. As states move forward with both assessment types, they will want to consider how to communicate the differences in expectations for Proficient on AA-MAS and grade-level assessments.

In part, these differences in expectation may be communicated by using different performance level labels and even a different number of levels. The nonregulatory guidance clearly states that states need not use the same number of performance levels for AA-MAS as for their grade-level assessments (USDE, 2007a, p. 23). As mentioned earlier, this guidance requires at least three performance levels. Elliott, Kettler, and Roach (2008) argued that changing the number of performance levels between grade-level assessments and AA-MAS may prove confusing for stakeholders, and they questioned how the state would decide which performance level to drop.

This is a fair question. If a state currently uses four performance levels (e.g., Below Basic, Basic, Proficient, and Advanced), the highest performance level (e.g., Advanced) seems likely to be dropped. If this were to happen, it is easy to imagine that stakeholders may become upset that eligible students do not have the opportunity to achieve the Advanced level on AA-MAS. On the other hand, it may prove more beneficial to the end users of AA-MAS if the underlying achievement spectrum is partitioned differently than the achievement spectrum underlying grade-level

assessments. It is anticipated that students intended to be eligible for AA-MAS are able to master grade-level content but at a slower pace than other students (Elliott et al., 2008). Thus, we might expect that fewer students will attain the Proficient level on AA-MAS. If this is the case, it may make sense to partition the Basic achievement level into two categories to provide opportunity to students who have not yet achieved the Proficient level to show improvement.

There are also psychometric considerations regarding the number of performance levels. Shorter tests, like AA-MAS, provide less information about student performance, making it desirable to segment the student achievement continuum into fewer achievement levels; moreover, in planning for item-based standard setting methods (e.g., Bookmark, Item Descriptor [ID] Matching), sufficient numbers of items should be present along the continuum of student achievement to enable accurate description of content expectations within each achievement level and to support the breadth and depth of descriptors for each achievement level.

Ercikan and Julian (2002) provided guidance on the number of performance levels by showing the degree of classification accuracy that state assessment programs can expect to achieve on an assessment based on the number of achievement levels and test score reliability. As they pointed out, the greater the number of achievement levels planned for a test score scale, the lower the accuracy will be of correctly identifying the true achievement level of examinees. The same relationship also exists between score reliability and classification accuracy.

Should the same or different performance level names be used for AA-MAS as for grade-level assessments? Beyond saying that three performance levels must be established, the USDE's (2007a, 2007b) guidance does not specify that the same labels should be used for AA-MAS performance levels as for grade-level assessments. Even though AA-MAS assess the same grade-level content, they assess less rigorous content. If a state uses the same labels for the performance levels, stakeholders could confuse the two sets of performance descriptors. The category Proficient will mean something entirely different on AA-MAS than it does on grade-level assessments. It seems appropriate to use different performance level names for AA-MAS to decrease possible confusion between the two tests.

A Process for Developing PLDs

There will probably never be one best process for developing and validating PLDs, and states have taken a variety of approaches to conceptualizing and defining Proficient and other levels of performance (see, e.g., Ferrara et al., 2009). In this section, we describe a three-step process for developing and refining PLDs. Throughout this process, the intent is to clarify, refine, and validate the PLDs. The intended meaning and interpretation of PLDs must remain intact throughout the process of setting cut scores and refining the PLDs.

*Policy-Based PLDs Prior to Standard Setting.*² After determining the appropriate number of and names for achievement levels, the state's next step is to conceptualize the level of

²We refer to policy-based PLDs that are based on and refer explicitly to grade-level content standards. These policy-based PLDs are distinct from policy definitions that apply generically to all grades and content areas, like those developed for the National Assessment of Educational Progress. (See <http://www.nagb.org/policies/PoliciesPDFs/Technical%20Methodology/developing-student-performance.pdf>)

proficiency they expect from students in each achievement level. These policy-based PLDs can be developed directly from the content standards, and outline the expectations that the state holds for what students should know and be able to do based upon the grade-level content standards. In recent years, and because of the influence of the peer review process, policy descriptors have evolved from generic performance statements (e.g., National Assessment of Educational Progress policy definitions) to content-based descriptions.

These more recent PLDs typically reflect the entire range of performance within an achievement level and do not target a particular area of the achievement level such as borderline performance. They assert academic knowledge and skills that the state expects of the Proficient student and students at other performance levels. They also set the tone for the standard setting (as well as the tone of the testing program). For AA-MAS, the tone of the policy-based PLDs should reflect the tone of the PLDs for the grade-level assessment. For example, one assessment should not assert world class standards, whereas the other calls for standards that represent minimal competency.

Developing Policy-Based PLDs. When developing PLDs for AA-MAS prior to setting standards, there are three areas that the state should consider: (a) the grade-level assessments' PLDs, (b) the content standards, and (c) the articulation of PLDs across grade levels. Further, development of any PLDs requires the input of knowledgeable stakeholders, such as assessment specialists, content specialists, and specialists in teaching students with disabilities. The development of PLDs may be accomplished through formal workshops or through informal meetings between the state and stakeholders. The remainder of the discussion in this section assumes a formal workshop is held; however, these steps could easily be applied to an informal meeting.

The grade-level assessment PLDs should serve as a starting point for the development of policy-based PLDs for AA-MAS. The content knowledge and skills targeted in the grade-level PLDs can be discussed in relationship to the students taking AA-MAS. During this exercise, workshop panelists can discuss the relationship between the achievement levels for grade-level assessments and AA-MAS. Panelists should pay explicit attention to the rigor as well as the content expectations asserted in the PLDs for grade-level assessment so that they are able to explicate if and how the rigor and content expectations have changed (or are the same) in the PLDs for AA-MAS.

Deconstruction of the content standards is another important component to the development of policy-based PLDs. During this process, workshop panelists consider each content standard and how students in each achievement level may perform on the content standard. The content standard may be parsed as panelists discuss how students at different performance levels are expected to perform on portions of the standard. For example, as panels deconstruct the range of inferences that can be made from informational and literary texts and how various texts influence the complexity of the inference, they can begin to identify inference types they expect to be present in particular performance levels on AA-MAS. Once the content standards have been deconstructed, panelists should examine the articulation of expected academic knowledge and skills across the achievement levels. Panelists should check that the academic knowledge and skills are coherent across the performance levels, and that the level of expectations increases as the achievement levels increase.

As a final step in the development of policy-based PLDs, a meta-panel should examine the PLDs for the articulation of expected academic knowledge and skills across the grade levels. The meta-panel would consist of representatives from the grade-level panels. This panel should ensure that the expectations of the academic knowledge and skills increase meaningfully from grade to grade.

Test-Based PLDs. The policy-based PLDs summarize the state's expectations of student performance in each achievement level. They articulate what students *should* do to be in a particular achievement level. The test-based PLDs, on the other hand, summarize student performance and how students did perform in each achievement level. Ideally the two would be one and the same; however, research has shown that policy-based PLDs are not always an accurate reflection of student performance once tests are administered (Burstein et al., 1996; Mercado & Egan, 2005). By themselves, policy-based PLDs may provide a somewhat inaccurate picture of what students in each achievement level know and can do and may confuse parents and teachers regarding the content knowledge and skills students actually hold.

Although policy-based PLDs reflect the range of student performance within an achievement level, test-based PLDs focus on a specific part of the range, such as the threshold or middle of the achievement level. Test-based PLDs should reflect the tone of the policy-based PLDs. Because test-based PLDs describe the knowledge, skills, and abilities found in the test items, they are sometimes considered a snapshot of the skills held by students in each achievement level.

Developing Test-Based PLDs. The development of test-based PLDs begins with standard setting, when panelists deconstruct the policy-based PLDs to focus on a target student. The target student is the student for whom cut scores are being set and provides a starting point for standard setting panelists. The discussion of the target student for each performance level helps panelists come to a common understanding of the expectations that they will operationalize through the standard setting. When panelists first discuss target students, they use three sources of information: their own knowledge, the content standards, and the state's policy-based PLDs. In the case of AA-MAS, it may also be helpful to provide panelists with summaries of the differences between the expectations for students on the grade-level assessments and the AA-MAS as well as the differences in the student population taking the two tests.

Throughout the standard setting, panelists gain greater insight about the target student through discussion with their peers, analysis of the items, and exposure to impact data, which may influence their original description of the target student. As such, the target student is a dynamic product that panelists discuss throughout the standard setting. At the end of the standard setting, it is important for panelists to have time to finalize the target student description based on everything gleaned during the standard setting.

The format of the final session will depend on the standard setting procedure used. This is a routine step in the Bookmark Standard Setting Procedure: panelists write descriptors of the items that precede the cut score. As an aggregate, these items describe the student who just enters the achievement level, that is, the skills of the target student. With other standard setting methods, it may be necessary to have the items mapped so that panelists can describe them.

States that use an item response theory (IRT) model to scale their assessments can collect information regarding how the knowledge, skills, and abilities targeted by items that map to

TABLE 1
 Partial Item Map With Completed Knowledge and Skill Descriptors

	Order of Difficulty	Item No.	Knowledge and Skill Descriptors
AL3	38	75	Draw conclusions from informational text
	37	35	Draw conclusions from informational text
	36	67	Identify supporting detail in informational text (persuasive letter)
	35	78	Draw conclusions from informational text
	34	62	Draw conclusions from informational text
	33	15	Draw conclusions from informational text
	32	89	Locate specific information in informational text (table)
AL2	31	11	Draw conclusions from narrative with graphics
	30	86	Identify main idea in informational text
	29	84	Locate specific information in informational text
	28	64	Determine the author’s purpose
	27	76	Identify supporting detail in informational text.
	26	52	Identify supporting detail in informational text (narrative)
	25	31	Identify main idea in informational text (narrative text)

an achievement level score range. Mapping the items this way relates item demands to the descriptions of what students should know and be able to do, regardless of the standard setting method used. After the operational standard setting is completed, an analysis for each item may be conducted in which a short phrase termed an item descriptor is developed. These item descriptors can be compiled into a PLD that reflects the target student. Because AA-MAS may be relatively short, supplementing final PLDs with illustrative items could be particularly helpful. Table 1 shows an example of an item map that has been partially completed with item knowledge and skill descriptors.

Final PLDs. The final PLDs should reflect both the policy- and test-based PLDs. Both types of PLDs provide different types of information regarding student performance in each achievement level. In this final phase, the two types of PLDs can be combined so that the final PLDs reflect what students in each achievement level are expected to (should) know as well as examples of actual academic knowledge and skills currently held by the students (i.e., what students can do). During the final phase, it may be necessary to update the policy-based PLDs to reflect knowledge gained from the development of the test-based PLDs.

Developing Final PLDs. The development of the final PLDs provides the state with the opportunity to synthesize the ideas and information found in the policy- and test-based PLDs. This process may reveal areas of disconnect between the two types of PLDs where academic knowledge and skills that students *should* have (via the policy-based PLDs) are not supported by the test-based PLDs. To finalize the PLDs, the state will want to have stakeholder input either through formal or informal meetings.

The first step in finalizing PLDs is to compare the skills in the two PLDs, side by side, for concurrence. If a skill was expected at the Proficient level in the policy-based PLDs but was

Guidance for the interpretation of these descriptors: The Performance Level Descriptors in this document give details about the knowledge and skills that students at the *cut score* for that particular achievement level are expected to know and be able to do. The particular standards and items that are tested may change from year to year to ensure comprehensive coverage of the state's standards as a whole. The skills described below indicate summary statements of skills that we expect students to consistently demonstrate, but these should not be interpreted as the only skills measured on these assessments or the only knowledge and skills students at the cut score possess.

FIGURE 1 Sample interpretive guidance.

easily mastered by Basic students according to the test-based PLDs, then this conflict will need to be examined. It may be that the skill simply needs to be re-aligned to the Basic category in the policy-based PLDs, or it may be that the skill needs to be further deconstructed because it was assessed only by items that captured a simpler component or depths of knowledge level of the skill.

Field testing PLDs may be an appropriate final step prior to release of the final version of the PLDs. First, if a state sets standards on one test form, a field test period would allow the state to validate the PLDs with data from a second test form to provide more stable information. Second, panels of teachers could review the PLDs and their interpretive guidance for clarity and understandability. Figure 1 is an example of suggested interpretive guidance. Teachers could provide feedback regarding how helpful (or not helpful) the PLDs are in providing useful feedback about student abilities.

PLDs Over Time. States must be very clear in how they define final PLDs, especially the skills from the test-based PLDs. The skills described in a test-based PLD often reflect a single form of the test. Such PLDs are a single snapshot in time of the skills that students in each achievement level *can* do. If a state does not plan to revisit the PLDs, then it is important that stakeholders are informed that the PLDs reflect a subset of skills held by students in each achievement level. In this case, PLDs are seen as static definitions that do not change as new information becomes available.

In other cases, PLDs are seen as fluid definitions that can and should be updated as new test forms are administered to students. Each new form will provide more information regarding the skills that students can do. Schneider, Egan, Kim, and Brandstrom (2008) found that only 19% to 30% of the original PLDs remained stable over time as new items were introduced in subsequent test forms. This suggests that original PLDs are limited in their ability to reflect actual student performance. This research echoes advice from Crane and Winter (2006), who recommend that PLDs be reviewed periodically to ensure that the descriptors still align with the competencies of the test. We also recommend including a discussion during test-based performance descriptor writing that reflects not only the test form(s) included in standard setting but also the content standards and items that are expected to appear in subsequent forms.

All of this points to the need for policy-based and test-based PLDs to work hand in hand. Policy-based PLDs are often more extensive than test-based PLDs and may be able to withstand

the differences in test forms over time. It is important that there is alignment between the policy-based and test-based PLDs.

SETTING MEANINGFUL AND APPROPRIATE MODIFIED ACHIEVEMENT STANDARDS

Setting cut scores and establishing modified achievement standards is the culmination of the design and implementation process of a new and unique assessment program and is the final step prior to reporting of assessment results. The standard setting design should be based on consideration of the test design, the test length, and the way the test is scaled and student performance reported. In addition, particularities of AA-MAS deserve special consideration when designing a standard setting. These special considerations must be addressed regardless of the standard setting method to be implemented.

In this section, we discuss two special considerations to take into account when planning standard setting for AA-MAS. Then we address general considerations that are relevant to all standard setting methods. Finally, we examine standard setting methods appropriate for AA-MAS comprised of multiple-choice and constructed-response items and methods appropriate for portfolio assessment designs.

Special Considerations for Setting Performance Standards for AA-MAS

The AA-MAS are based on the same content standards as a state's grade-level assessments. Expectations for what students should know and be able to do, however, are modified for students who are eligible for AA-MAS. In turn, the complexity of the PLDs and the difficulty of the test items are also adjusted to be consistent with the modified achievement standards. This presents a unique situation. The same content standard structure underlies two assessment programs that are administered to two different groups of students—supporting the idea that both assessments target the same achievement construct. A link exists between the two assessments that should be addressed as part of standard setting.

Standard setting panelists should understand the relationship between AA-MAS and grade-level assessments. This can be addressed by the state having a clear definition of the target population for AA-MAS and explaining the relationship between the two student populations for whom the tests are designed. This also can be addressed through a study of the policy-based PLDs, which should be unique for each of the assessments. As a first step for any standard setting method, panelists should actively discuss the policy-based PLDs. In the case of AA-MAS, the panelists should study the policy-based PLDs for the modified achievement standards and compare them to the policy-based and test-based PLDs for the grade-level assessment. In most states, the policy-based and/or test-based PLDs for the grade-level assessment have been published and disseminated for several years. This comparison activity will aid panelists in understanding the difference in expectations between the two assessments. It should be noted that this is an additional activity on the meeting agenda that, at least up to now, would not be seen at a standard setting for a grade-level assessment.

Some states may attempt to link the AA-MAS to the grade-level assessments using statistical linking methods so that they can locate the grade-level assessments' cut scores on the test scale for the AA-MAS. For example, a state may use equipercentile equating to establish cut scores on the AA-MAS that produce similar impact data for both assessments. Alternatively, a state could link the scales from the two assessments in a random sample of students from the general examinee population that takes both the AA-MAS and the grade-level assessments. With grade-level assessments' cut scores projected onto the scale of AA-MAS, standard setting panelists could review and adjust the cut scores through a review process instead of a full standard setting. This approach enables panelists to review and consider the specific attributes of AA-MAS and the target examinee population while showing them where the cut scores for the grade-level assessments would be located.

If a state attempts to link AA-MAS and grade-level assessments statistically, it is even more important that the standard setting panelists understand the similarities and differences between the two assessments. Almost all of the standard setting methods we consider can be used for a cut score review as well as for a full standard setting. Of the methods we consider, however, the item mapping methods such as the Bookmark Standard Setting Procedure and ID Matching best lend themselves to cut score reviews.

General Considerations for Setting Cut Scores for AA-MAS

Some aspects of setting performance standards for AA-MAS are universal to all assessments, regardless of a test's format.

Recruiting Standard Setting Panelists. Two questions regarding panelist recruitment must be answered: Who? and How many? The answer to these questions is, A diverse group that is as large as possible. For educational achievement assessments, Hambleton and Pitoniak (2006) recommended 15 to 30 panelists per standard setting committee to obtain adequate representativeness and stability of cut score recommendations.

The importance of diversity of a standard setting panel has long been recognized (Hambleton & Pitoniak, 2006). It is also important that all panelists are knowledgeable about the students who are assessed and the content standards on which they are assessed (Plake, 2008). Obviously, educators who teach the students who are appropriately eligible for AA-MAS are most knowledgeable. It is, however, important to recruit panelists from other stakeholder groups because a diversity of perspectives enhances the group decision-making process (Surowiecki, 2004). The group of panelists should represent the views of all stakeholders, including teachers, parents, students, administrators, and policymakers. Panelists for the AA-MAS standard setting may include teachers of students with learning disabilities, teachers of English-language learners, teachers from regular education classrooms, interested members of the public with some requisite background knowledge of test content and the target examinees, administrators, and parents. The views of education policymakers who are responsible for establishing expectations for student achievement—for example, state board members—should be represented as well. Elliott et al. (2008) cited the recent Individuals with Disabilities Education Act (commonly known as IDEA) that expects a diverse group of stakeholders be involved in the standard setting process.

It is important to underscore that panelists should have knowledge of the content standards and student population for whom the cut scores are being established. As Plake (2008) pointed out, the validity of a standard setting may be undermined if the panelists do not have the requisite knowledge base prior to entering the standard setting. Although we agree with Plake that knowledge of the content and student are important, we believe that the diversity of the panel is of equal import. A diverse group of standard setting participants ensures that many different points of view will be brought to the table and reflected in the cut scores. For example, parents of students taking the AA-MAS may not have content knowledge, but they have advocated for that student through IEP meetings, parent-teacher conferences, and so on. They bring a unique perspective of the student experience to the table that will differ from teachers. Notice that we do not argue that content or student knowledge should be entirely disregarded for the sake of diversity; however, we do believe that a more informed decision will be made only by bringing the various stakeholder groups to the table. This being said, all panelists must be trained to develop an adequate understanding of the content standards, target examinees, and standard setting procedure that will be implemented.

Standard setting panels should be diverse, representing the population of the state. Rationales that guide recruitment should be noted explicitly in the standard setting technical report, and analyses of panel diversity should be documented. Some states have great difficulty in recruiting minority teachers. In these cases, a state may consider basing their recruiting efforts on the demographics of the students that educators teach instead of the demographic characteristics of the educators themselves.

Vertical Articulation. The AA-MAS can be administered in multiple grades and content areas, making the articulation of the cut scores and/or impact data an important consideration for standard setting. Procedures for vertical articulation of cut scores are described elsewhere (e.g., Cizek & Bunch, 2007; Ferrara et al., 2007; Hambleton & Pitoniak, 2006; Lewis & Haug, 2005). Any standard setting design should consider the issue of vertical articulation and how it will be addressed. For AA-MAS, the considerations are unique. For example, many of the earliest states have developed AA-MAS for a subset of contiguous grades rather than Grades 3 through 8 and high school, and one content area rather than both reading and mathematics. Such states should consider articulating achievement standards across the subset of grades when they set standards. They will also need to consider how to articulate the achievement standards in the future if they add grades and content areas. Current procedures for articulating standards (e.g., writing articulated PLDs, articulating impact data by adjusting cut scores; see Ferrara et al., 2007) can be adapted.

Standard Setting for AA-MAS Composed of Multiple-Choice and/or Constructed-Response Items

As mentioned previously, of the six states that have built or are currently building AA-MAS, five used only multiple-choice items or multiple-choice items with constructed-response items or writing prompts. Assessments with multiple-choice items and those with both multiple-choice and constructed-response items lend themselves to conventional standard setting methods (e.g.,

Bookmark, modified Angoff, contrasting groups). This is not surprising. The predominant approach to developing AA-MAS is to adapt items and test blueprints from existing grade-level assessments. We provide brief overviews of the methods we cover but leave out important details (e.g., when panelists share insights, when they make independent judgments) that are covered extensively elsewhere (e.g., Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). We propose approaches to using conventional standard setting methods to meet the unique challenges of AA-MAS.

Bookmark Standard Setting Procedure and Other Item Mapping Methods. The Bookmark Standard Setting Procedure (Lewis, Mitzel, & Green, 1996) is one of the most popular methods currently used in state assessment programs (Karantonis & Sireci, 2006). Other item mapping methods that have been used in other assessment programs include the Item-Mapping method (Wang, 2003) and ID Matching (Cizek & Bunch, 2007; Ferrara, Perie, & Johnson, 2008). These three methods are related psychometrically because they required panelists to make judgments about items mapped to IRT score scales. Their cognitive-judgmental tasks are quite different (e.g., Ferrara et al., 2008). Although our discussion about setting cut scores for AA-MAS focuses on the Bookmark Procedure, the discussion is applicable to the Item-Mapping and ID Matching methods.

The Bookmark Standard Setting Procedure is typically implemented in three rounds of judgments using two key components: the Ordered Item Booklet (OIB) and the item map. The OIB is composed of multiple-choice and constructed-response items, ordered by item difficulty. The item map provides details on the items in the OIB (e.g., item difficulty, scoring key, and the content standard that each item targets). Standard setting panelists record their responses or reactions to two questions on the item map or in the OIB: (a) What does this item measure? That is, what does a student need to know and be able to do to respond successfully to this item or score point? and (b) Why is this item more difficult than the preceding items? In Rounds 1, 2, and 3, a panelist places a bookmark on a page in the OIB to indicate a recommendation for a cut score. The panelist's cognitive-judgmental task is stated as "Place your bookmark on the page where two-thirds (or one-half) of those students who are just barely in a performance level would respond successfully."³ Typically, panelists view impact data (e.g., the percentage of students at or above the Proficient level) to inform their final bookmark placements. Final cut scores are based on the median of the page numbers on which panelists place their bookmarks in the final round of judgments. Full descriptions of materials, training, and procedures for the Bookmark Procedure appear in a number of sources (Cizek & Bunch, 2007; Lewis, Green, Mitzel, Baum, & Patz, 1998).

Challenges in Using the Bookmark Procedure for AA-MAS. Two challenges should be considered before implementing the Bookmark Standard Setting Procedure for AA-MAS. These challenges must be addressed in a Bookmark standard setting for any assessment; however, they require special consideration for AA-MAS. The first challenge is the likely shorter lengths of

³It is common to state the cognitive-judgmental task in another way: Place the bookmark on the first page where you judge that a student who has the knowledge and skills to demonstrate mastery of the items before the bookmark would be classified as Proficient [or whatever the achievement level may be].

these assessments; the second has to do with the response probability used to order the items. We discuss ways to address these challenges.

Challenge 1: Test length. In the Bookmark Standard Setting Procedure, the primary focus for decision making is the items that comprise the test. Over the course of two or three rounds of discussion and decisions, panelists build an understanding of what the target student should know and be able to do in each of the performance levels. To do this successfully, panelists must view an OIB with enough pages (each multiple-choice item and constructed-response score level appears on a separate page) so that they can build adequate understanding of the knowledge and skill requirements of the test. In other words, the panelists use this information to build a story of what the target student is able to do. The score scale underlying the OIB should not have large gaps between item locations. If a test is short, the items may not cover the entire scale range, which leaves gaps in the story participants build about what the target student is able to do.

This is an important consideration for AA-MAS, where states may modify and shorten versions of their grade-level assessments. Our experience has shown us that there should be at least 10 and as many as 20 items or score points per cut score. This guideline ensures that adequate numbers of items appear in the OIB to cover the contents of the test blueprint, minimize gaps between item locations (or difficulty), and provide adequate separation on the test score scale between cut scores. Requirements for AA-MAS require at least three achievement levels (i.e., two cut scores). This means that OIBs should be between 30 to 60 pages long for AA-MAS. The upper end of this range represents a goal, not a requirement. We recognize that it is unlikely that AA-MAS will include 60 points. We anticipate that most states will create AA-MAS that are toward the low end of our suggested range.

It is sometimes possible to fill in large gaps in the location of items on the underlying score scale by augmenting the OIB with items from the broader item pool. When augmenting, it is important to consider covering the entire range of the test scale, the test content blueprints, and the gaps in item locations. If a broader item pool is available (perhaps from a field test administration), one can begin by identifying items to fill location gaps. At the same time, it is important to avoid making the OIB too long and burdensome for panelists to review. This can be done efficiently by selecting multiple-choice items or one-point constructed-response items. If the item pool offers sufficient choice of items, then it is also important to consider the test content blueprint. To the degree possible, the OIB should proportionally match the test blueprint. Often, though, it is the case that the available item pool offers a limited number of items and it may be fortunate to find any items to enhance the scale coverage. Many assessments do not have item pools from which to augment OIBs. For short tests where the scale is not adequately covered, states may want to consider an Angoff or Body of Work standard setting method.

Challenge 2: The response probability criterion. The response probability (RP) criterion is, perhaps, the most discussed aspect of the Bookmark Standard Setting Procedure. In item mapping methods such as Bookmark, items are ordered in an OIB using an RP criterion. The selected criterion indicates the probability that an examinee whose proficiency on the score scale is equal to an item's location will respond correctly to a multiple-choice item or at a given score level on a constructed-response item. The RP criterion also indicates students' mastery of content by specifying the likelihood with which students at a particular scale score will answer items

correctly. Using a response probability of 0.67 (RP67) as an example, multiple-choice items and score levels for constructed-response items are located on the underlying IRT score scale and placed in an OIB so that examinees with proficiencies equal to an item's RP67 location have a two thirds likelihood of answering the item correctly (often, with guessing factored out). Thus, when a cut score is placed in an OIB, we can say with some certainty (i.e., two thirds probability) that students who meet or exceed the cut score have *mastered* the content which the standard setting participants expected them to master. The students at or above the cut score will have at least a two thirds likelihood of answering the item correctly. The RP criterion influences cut scores, impact data, and the ordering of the items in the OIB.

For AA-MAS, the RP criterion should be selected in light of state policy and methodological implications. From a policy perspective, what RP criterion is appropriate for students eligible for AA-MAS? Should items be ordered using RP67, thus defining content mastery at two thirds? Or is a 50-50 (RP50) likelihood a more appropriate definition of mastery? Is it important to select the same RP criterion that was used for the standard setting for the grade-level assessments?

The importance of the methodological implications of the RP value cannot be overstated. The distribution and order of items along the test score scale differs depending on the RP value used. In some cases, using RP67 locates very few items at the low end of the scale, whereas RP50 may distribute the items more widely across the scale score range. In this case, RP67 might force panelists to choose higher cut scores than they intend, whereas RP50 might provide more flexibility in lower ranges of the score scale. In some cases the preferred RP value from a policy perspective may be at odds with the RP value that is best from a methodological perspective. In these cases, states must decide which RP criterion is the most appropriate choice for standard setting. In the end, the best decision from a standard setting methodology point of view may be to select the ordering information and criterion that ensures that adequate numbers of items are available in the OIB for the full range of examinee proficiency.

Modified Angoff Standard Setting Method. Even though the modified Angoff cognitive-judgmental task has been famously deemed a complex, "fatally flawed" cognitive task (Shepard, Glaser, Linn, & Bohrnstedt, 1993), it, along with the Bookmark Procedure, account for the majority of standard settings for state assessment programs over the last 15 years. Use of the modified Angoff method for educational achievement tests can be traced to the 1970s. The modified Angoff method typically is implemented in two or three rounds of judgments and discussion,⁴ where panelists discuss item response demands and target examinees and make judgments about every item in a test book. The key components in a modified Angoff standard setting are the test items, typically presented in the test books that examinees use, and conceptualizations of target students. In each round, the panelists' cognitive-judgmental task is to estimate the percentage of target examinees (e.g., students who are just barely Proficient) they expect to respond successfully to a multiple-choice item or at a score level on a constructed-response item. Typically, panelists view impact data (e.g., percentages of students at or above the Proficient level) to inform their final judgments. To calculate cut scores for each round, the average expected percentage for each item across all panelists is calculated, the averages summed and translated into a total test cut score.

⁴In the original (unmodified) Angoff method, standard setting panelists judged simply whether a target examinee would respond correctly or incorrectly to a multiple-choice item.

Full descriptions of materials, training, and procedures for the modified Angoff method appear in several texts (Cizek & Bunch, 2007; Zieky, Perie, & Livingston, 2008).

Challenges in Using the Modified Angoff Method for AA-MAS. One main challenge should be considered before implementing the modified Angoff method to set cut scores for AA-MAS. As with the Bookmark Procedure, this challenge must be addressed in using the modified Angoff method for any achievement test. However, it requires special consideration for AA-MAS.

Challenge 1: Probability judgments. People are notoriously inaccurate in making probability judgments. The amount of empirical research and the range of contexts in which this observation has been documented are impressive. People can be trained to estimate probabilities “moderately well” (Nickerson, 2004, p. 433), but they are susceptible to judgmental biases and are prone to making errors when judging the probability of an occurrence (Nickerson, 2004, chap. 11; Plous, 1993, p. 144). Although probability type judgments are embedded in the Bookmark judgmental task through the response probability criterion, probability judgments in the form of percentages *are* the cognitive-judgmental task for the modified Angoff method. We acknowledge that the modified Angoff method continues to be used to produce appropriate and acceptable standards. In addition, standard setting panels have been shown to produce cut scores from the modified Angoff method which are not that different from cut scores produced using the Bookmark Procedure (cf. Karantonis & Sireci, 2006, p. 8), indicating that standard setting panels produce appropriate and acceptable cut scores using the modified Angoff method despite its complexity.

Probability judgments, however, may be more difficult for AA-MAS, at least for now. The field is struggling to define the target examinee population, so it may be even more difficult for standard setting panelists to estimate percentages in the modified Angoff method, especially if the target examinee population is not clearly defined and understood. Training can help improve probability judgments (Nickerson, 2004, p. 433). Training panelists to understand clearly the target examinee population is a special requirement for setting standards for AA-MAS using the modified Angoff method.

Challenge 2: The number of judgments required. The judgmental task associated with the modified Angoff method asks panelists to consider target examinees and estimate the percentage that would answer a multiple-choice item correctly or achieve a score level for a constructed-response item. Although this is not a challenge unique to assessments of modified achievement standard, it does warrant consideration. Using a 30-item AA-MAS and only two cut scores to produce Advanced, Proficient, and Basic levels (the minimum requirement) to illustrate, panelists would have to estimate the probability of successful responses for 30 items for the Basic/Proficient cut score and 30 for the Proficient/Advanced cut score. Panelists would have to make 60 probability judgments—judgments that they probably find challenging—for each of two or three rounds of standard setting.

Standard Setting for AA-MAS Using Collections of Evidence. At least one state so far has designed an AA-MAS using portfolios containing student work as evidence of achievement.

Assessments based on collections of evidence are most compatible with standard setting methods that base judgments on student work rather than on test items. Here we address two of the most widely used of those methods.

Body of Work and Performance Profile Sorting Methods. Some states may base their designs for AA-MAS, not on conventional grade-level assessments, but on the design approaches for their alternate assessments based on alternate achievement standards. Eleven of 31 states with information available on their Web sites use portfolio designs for their alternate assessments (Ferrara et al., 2009, Table 1 and surrounding text). Portfolio assessments contain collections of student academic work, video and audio recordings of them as they perform academic tasks, and other evidence of their current performance in relation to extended grade-level content standards. Portfolios, or collections of student work, have been used to guide instruction effectively, but with limited psychometric rigor, for assessing students with significant cognitive disabilities and for assessing mathematics and writing achievement of all students in selected grades in Vermont (Koretz, Stecher, Klein, & McCaffrey, 1994). Portfolio assessments lend themselves to standard setting methods that focus on student work or ratings of the quality of the work, rather than on test items.

Two of the most widely known of these methods are the Body of Work and Performance Profile sorting methods (see Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006). In the Body of Work method, standard setting panelists examine student work, typically contained in portfolios, that represents the range of possible total portfolio scores. They sort the portfolios in two different rounds into groups that correspond to the achievement levels for which standards are being set. Cut scores are calculated using logistic regression (simpler calculations are also possible) to identify portfolio scores that most clearly distinguish the scores of panelists' groupings of student work into performance levels. Because preparing and organizing student work is such a large logistical task, and because sorting large numbers of portfolios can be arduous for panelists, some standard setters may prefer the Performance Profile sorting method (see Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006).

As in Body of Work, Performance Profile panelists sort evidence of student performance into achievement levels. In this case, score profiles are sorted, rather than student work. Specifically, panelists review visual or numerical displays of score combinations (i.e., score profiles) of student work (e.g., accuracy scores on a 4- or 6-point scale for several pieces of student work),⁵ sorted from lowest to highest, and draw lines to determine cut scores for achievement levels. Because some total test scores can be achieved by several score profiles, panelists examine score profiles for each total test score. Cut scores are calculated as described for the Body of Work method. Although Performance Profile sorting reduces the logistical burden of the Body of Work method, the number of score profiles that panelists must consider and sort can be large. The subtlety and complexity of judgments required to distinguish profiles is challenging, for example, whether a score profile of (6, 6, 7) should be sorted into the same achievement level as a score profile of (4, 5, 10).

⁵Scoring scales for some portfolio assessments can be as high as 10 per collection of evidence, as when evidence is scored once for accuracy on a 1- to 5-point scale and once for support on a 1- to 5-point scale.

Challenges in using the Body of Work method for AA-MAS. We already have alluded to the significant logistical challenge posed by the Body of Work method. In addition, it has been common to discover that almost no student work can be located for the lower score levels for alternate assessments, because so few students receive low scores. For some alternate assessments, it has been necessary to simulate performance at the lower score levels, which may degrade the quality of student portfolios. We do not know yet whether this finding or the opposite—too little evidence at the top end of the scale—will occur for AA-MAS.

Challenges in using the Performance Profile sorting method for AA-MAS. In Performance Profile sorting, the logistical challenge is reduced because only small numbers of exemplar student work collections are necessary. However, the cognitive-judgmental burden and complexity increase significantly for standard setting panelists. For example, a simple portfolio that includes three pieces of evidence, each scored on a 1- to 4-point scale, produces 4^3 (or 64) score profiles for panelists to examine and sort into achievement levels. As we illustrated earlier, some profiles cannot be sorted easily. For example, because panelists know that requirements of the three pieces of evidence may not be equivalent, they may struggle with deciding whether to sort a score profile of (4, 3, 4) into the same achievement level as the score profile (4, 3, 2) or with the score profile (4, 4, 3). Making this discrimination may be particularly difficult for AA-MAS because the target examinees are capable of challenging grade-level work and display uneven performance in content areas because of their disabilities.

Standard setting designers can devise methods for reducing the number of profiles to be examined. For example, there is little question that profiles of (1, 1, 1) and any combination of two 1s and a 2 are going to be sorted into the lowest achievement level, so reviewing them may not be necessary. States may also provide panelists direction on a strategy that emulates the state's perception of target study by achievement level. For example, the state may direct the panelists to give greater "weight" to certain standards as reflective of the blueprints or, as in the case of alternate assessments, to higher levels of independence. In any case, managing the complexity of judging the performances of students with disabilities may remain a challenge for standard setting panelists.

VALIDATION STUDIES

Evidence to support the validity of the interpretation of cut scores, performance level descriptions, and performance standards is conceptually complex and requires significant sustained effort. Collecting validation evidence begins with design, development, and implementation of a test and carries through to the implementation of the cut scores themselves. In the discussion of standard setting methods for AA-MAS, we addressed Kane's (2001) concern that standard setting methods should be consistent with the design and type of test. We suggested two designs for AA-MAS and corresponding standard setting methods. The rationale for the method selected should be clearly stated in the standard setting design.

In this section, we touch on external checks that a state can reasonably undertake in collecting evidence of validity. We discuss evidence based on relations to other measures, as required in peer review guidance for AA-MAS (USDE, 2007b) and the Standards for Educational and

Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Evidence of relations with other measures would compare the results from the application of cut scores on AA-MAS to other relevant results. Students' classification into achievement levels can be compared to performance on other tests, in classroom assessments, and other meaningful measures. It is important to articulate the intended interpretations of performance on AA-MAS and how performance in the achievement levels should compare to levels of performance on the grade-level assessments.

Contrasting Groups Survey

A second standard setting method sometimes is used as evidence of external validation of performance standards. However, it is common to find that different standard setting methods produce different results. Assuming both methods are well implemented, this leaves a state in a predicament, having two sets of recommended cut scores from which to produce one set of cut scores. If the purpose of the second method is to validate, it is not helpful or useful to obtain a different set of results.

Instead of implementing a second standard setting workshop, the Contrasting Groups method may be implemented as a survey. Using the survey, teachers classify each of their students into an achievement level using the PLDs for an assessment. In using the Contrasting Groups method for validation rather than setting cut scores, the purpose of the survey would be to see how well teacher judgments of student performance align student performance on the assessment—specifically, the performance levels in which their scores place them. New cut scores would not be derived from the survey results. Instead, classification consistency analyses would be conducted to compare teacher classifications of student performance and classification of students based on the assessment.

The survey could be implemented either through a paper-and-pencil mailed survey or a Web-based approach. The challenge of implementing the Contrasting Groups approach as a survey is ensuring that all participants consider the PLDs before classifying students into an achievement level. If the survey is implemented through the Web, teachers simply can click on a PLD for each student. The amount of time spent on each screen could be captured and analyzed to ascertain if participants actually spent time analyzing the PLDs. With both approaches, participants can be asked how rigorously they considered the PLDs before classifying students. Unfortunately, there is no way to guarantee that the participants study the PLDs before classifying students. To encourage thoughtful responding, states can administer the survey at other test-related workshops, such as Content and Bias Review, where a diverse group of panelists is available. A small portion of time could be devoted to the survey. The survey probably cannot be administered during a standard setting workshop. In many cases, state departments or boards of education adjust cut scores in light of policy considerations following a standard setting workshop so that final cut scores may not be available until well after the standard setting workshop. It is desirable that an independent group provide the validation information.

If the results from the survey and the test-based classifications do not correspond closely, the state will need to examine the reasons for this. It may uncover a disconnect between the PLDs

and the test performance. It may also reveal that teachers tend to over- or underestimate student performance in relation to the performance levels and PLDs.

Comparison of Results from AA-MAS and Grade-Level Assessments

Another approach to collecting validation evidence could involve two groups of students taking both AA-MAS and grade-level assessments. One group would be a diverse sample of examinees who are eligible for the AA-MAS (the target examinee group), the other a diverse sample of students who are eligible for the grade-level assessments (the comparison group). We would expect the average scores of the two groups to be considerably different on both assessments. Further, we know from other studies (e.g., Fincher, 2008) that large numbers of the target examinee group will perform in the lowest achievement level of the grade-level assessment. Likewise, we expect large numbers of the comparison group students to perform at the highest achievement level of the AA-MAS. If the PLDs for the AA-MAS are developed in relation to the PLDs for the grade-level assessments, we would expect student performance to provide support. For example, if Proficient on AA-MAS is intended to be similar to Basic on the grade-level assessments, we would expect that students classified as Basic on the grade-level assessment would, in general, be classified as Proficient or Advanced on the AA-MAS.

Of course, grade-level assessments are not appropriate for students who are appropriately eligible AA-MAS. Other approaches that examine response demands of items that define each cut score (e.g., Ferrara et al., 2007) rather than require administration of items could provide useful validation evidence. In an item response demands study, the academic knowledge and skill demands of items that define Proficient performance on the AA-MAS could be compared to the knowledge and skill demands of items that define a corresponding level on the grade-level assessments.

CONCLUSION

As we indicated at the beginning of this article, requirements for developing and validating educational achievement tests apply to AA-MAS with the same rigor and importance that they apply to grade-level assessments. However, AA-MAS are unique because the target examinee population is unique and relatively unfamiliar to designers of large-scale educational achievement tests. In addition, AA-MAS are new enough that information is still emerging about the most effective and affordable approaches to assessing the target examinees.

Some points are clear. It is crucial to define and describe the target examinee population. This has become a routine process for grade-level assessments. The field has made progress in defining the target examinee population for alternate assessments for students with significant cognitive disabilities. However, progress in defining the target examinee population for AA-MAS is only beginning. It is standard practice to develop PLDs prior to standard setting and, in fact, early in the test design and development process. This practice also applies to AA-MAS, although it will continue to be particularly challenging to define the PLDs until we are able to define the target examinee population clearly. Finally, it is important to consider the challenges we

describe to implementing conventional standard setting methods for AA-MAS. For AA-MAS, both conventional and innovative practices are required.

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.
- Barton, K. (2006, April). Approaches to developing modified alternate assessments. Paper presented in H. Huynh (Chair), *Alternate and modified assessments for accountability and AYP requirements: Policy, technology, and implementation considerations*. Symposium conducted at the annual meeting of the National Council on Measurement in Education, San Francisco.
- Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 31–63). Maple Grove, MN: JAM Press.
- Burstein, L., Koretz, D., Linn, R., Sugrue, B., Novak, J., Baker, E. L., et al. (1996). Describing performance standards: Validity of the 1992 National Assessment of Educational Progress PLDs as characterizations of mathematics performance. *Educational Assessment*, 3(1), 9–51.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
- Crane, E. W., & Winter, P. C. (2006). *Setting coherent performance standards*. Washington, DC: Chief Council of State School Officers.
- Elliott, S. N., Kettler, R. J., & Roach, A. T. (2008). Alternate assessments of modified achievement standards: More accessible and less difficult tests to advance assessment practices? *Journal of Disability Policy Studies*, 19, 140–152.
- Ercikan, K., & Julian, M. (2002). Classification accuracy of assigning student performance to performance levels: Guidelines for assessment development. *Applied Measurement in Education*, 15(3), 269–294.
- Ferrara, S., Perie, M., & Johnson, E. (2008). Matching the judgmental task with standard setting panelist expertise: The Item-Descriptor (ID) Matching procedure. *Journal of Applied Testing Technology*, 9(1). Retrieved February 11, 2009, from http://www.testpublishers.org/Documents/JATT2008_Ferrara%20et%20al.%20IDM.pdf
- Ferrara, S., Phillips, G., Williams, P., Leinwand, S., Mahoney, S., & Ahadi, S. (2007). Vertically articulated performance standards: An exploratory study of inferences about achievement and growth. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 31–63). Maple Grove, MN: JAM Press.
- Ferrara, S., Swaffield, S., & Mueller, L. (2009). Conceptualizing and setting performance standards for alternate assessments. In W. D. Schafer & R. W. Lissitz (Eds.), *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 93–111). Baltimore: Paul Brooks Publishing.
- Ferrara, S., & Wright, L. (2007, July 26). *Alternate assessments based on modified achievement standards: Following standard design and development principles enables valid interpretations and decisions*. Invited presentation in a meeting of the Special Education Partnership, Washington, DC.
- Filbin, J. (2008). *Lessons from the initial peer review of Alternate Assessments Based on Modified Achievement Standards*. Washington, DC: USDE.
- Fincher, M. (2008, June). Georgia's journey towards an AA-MAS. In *1%, 2%: skim or whole assessments*. Paper presented at the National Conference on Student Assessment, Orlando, FL. Retrieved November 6, 2008, from <http://www.ccsso.org/content/PDFs/14%5FALL.pdf>
- Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.
- HB 05-1246 Study Committee. (1995, December 31). *Assessing "students in the gap" in Colorado: Report from the HB 05-1246 Study Committee*. Minneapolis, MN: National Center on Educational Outcomes. Retrieved February 11, 2009, from the National Center on Educational Outcomes Web site: <http://cehd.umn.edu/NCEO/Teleconferences/tele11/ColoradoStudy.pdf>
- Kane, M. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Erlbaum.

- Karantonis, A., & Sireci, S.G. (2006). The Bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice*, 25(1), 4–12.
- Koretz, D., Stecher, B. M., Klein, S. P., & McCaffrey, D. F. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement: Issues and Practice*, 13(3), 5–16.
- Lazarus, S. S., Thurlow, M. L., Christensen, L. L., & Cormier, D. (2007). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2007* (Synthesis Rep. No. 67). Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Lewis, D. M., & Green, R. (1997, June). *The validity of PLDs*. Paper presented at the National Conference on Large Scale Assessment, Colorado Springs, CO.
- Lewis, D. M., Green, D. R., Mitzel, H. C., Baum, K., & Patz, R. J. (1998). *The Bookmark Standard Setting Procedure: Methodology and recent implementations*. Paper presented at the annual meeting of the Council of State School Officers, Phoenix, AZ.
- Lewis, D. M., & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education*, 18(1), 11–34.
- Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). *Standard setting: A bookmark approach*. Paper presented at the National Conference on Large-Scale Assessment, Phoenix, AZ.
- Loomis, S. C. (2001, June). *Judging evidence of the validity of the National Assessment of Educational Progress Achievement Levels*. Paper presented at the National Conference on Student Assessment, Houston, TX.
- Mercado, R. L., & Egan, K. L. (2005). *Performance level descriptors*. Paper presented at the National Council on Measurement in Education, Montréal, Quebec, Canada.
- National Assessment Governing Board. (1995). *Developing student performance levels for the National Assessment of Educational Progress: Policy statement*. Washington, DC: Author. Retrieved from <http://www.nagb.org/policies/PoliciesPDFs/Technical%20Methodology/developing-student-performance.pdf>
- Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Erlbaum.
- Perie, M. (2008). A guide to understanding and developing PLDs. *Educational Measurement: Issues and Practice*, 27(4), 15–29.
- Plake, B. S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice*, 27(1), 3–9.
- Plous, S. (1993). *The psychology of judgment and decision making*. New York: McGraw-Hill.
- Rigney, S. (2008). *Federal policy & statewide assessments for students with disabilities*. Minneapolis, MN: National Center on Educational Outcomes. Retrieved February 11, 2009, from the National Center on Educational Outcomes Web site <http://www.cehd.umn.edu/nceo/Presentations/OSEP2008/OSEPprojDir708.ppt>
- Schneider, M. C., Egan, K. L., Kim, D., & Brandstrom, A. (2008, March). *Stability of achievement level descriptors across time and equating methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.
- Shepard, L. A., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). *Setting performance standards for student achievement. Report of the NAE Panel on the Evaluation of the NAEP Trial State Assessment: An Evaluation of the 1992 Achievement Levels*. Washington, DC: National Academy of Education.
- Surowiecki, J. (2004). *The wisdom of crowds*. Boston: Little Brown.
- U.S. Department of Education. (2007a). *Modified academic achievement standards: Non-regulatory guidance*. Washington, DC: Author.
- U.S. Department of Education. (2007b). *Standards and assessments peer review guidance: Information and examples for meeting the requirements of the No Child Left Behind Act of 2001. Revised December 21, 2007*. Washington, DC: Author.
- Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item mapping method. *Journal of Educational Measurement*, 40(3), 231–253.
- Webb, N., Alt, M., Ely, R., Cormier, M., & Vesperman, B. (2006). The Webb alignment tool: Development, refinement, and dissemination. In *Aligning assessments to guide the learning of all students* (pp. 1–30). Washington, DC: Council of Chief State School Officers. Retrieved February 11, 2009, from <http://www.ccsso.org/publications/details.cfm?PublicationID=293>
- Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.