Criteria for Standard Setting from the Sponsor’s Perspective

William D. Schafer

Department of Measurement, Statistics, and Evaluation
University of Maryland

Two concerns related to setting performance standards on educational assessments are discussed. First, criteria for a standard-setting process from the point of view of a standard-setting sponsor, called here “institutional criteria,” are developed using a state department of education as an example. Four institutional criteria are proposed: (a) consistency with policy goals, (b) legal defensibility, (c) generation of assets for support, and (d) efficiency. Some steps sponsors may take toward meeting these criteria are introduced by reviewing, and in some cases elaborating on, Hambleton’s (2001) 20 questions from the perspective of the 4 criteria. Second, the concept of vertical moderation of performance standards is reviewed, and considerations that can lead to different choices for moderation of standards are explored. Realistic state data are used to illustrate some of these choices. How moderated standards may be consistent with the 4 institutional criteria is discussed.

PERSPECTIVES ON STANDARD SETTING

Standard setting (or more precisely, setting cut scores that operationalize standards) was criticized as too arbitrary (Glass, 1978) and has even been called “the Achilles’ heel of the high-stakes testing business” (Glass, 2003, p. 102). In contrast, Mehrens and Cizek (2001), in their concluding chapter in a book devoted to standard setting, stated that “the technology of setting performance standards is … well developed” (p. 484) and Cizek (personal communication, July 6, 2004) believes that “nearly all psychometricians now agree that standard setting is not an arbitrary endeavor but a fairly well-developed technology for applying human
judgment in reasonable, reproducible, and defensible ways.” Its prevalence, if nothing else, suggests that no acceptable alternative has been found to satisfy the need for judgments of whether achievement in a specified domain is sufficient for some adjective such as basic, proficient, or advanced.

The popularity of standard-setting procedures and subsequent decision making implies that it is useful to consider criteria for the success of a standard-setting event. There are at least four perspectives from which criteria might be approached. I list them here in order of increasing usefulness to the sponsoring agency.

**Definitional Perspective**

To be called performance standards, there must be operationally defined, mutually exclusive, exhaustive ordered categories and a decision process based on one or more assessments to place students in those categories. This is not very helpful in determining whether the quality of the decision rule is adequate.

**Psychometric Perspective**

Standards result in categorizations of students, and inferences are drawn from those categories about students and about the programs that, at least in part, are responsible for those categorizations. As such, they form a scale that can be evaluated using the well-known criteria of reliability, validity, and utility. Kane (2001) took this perspective in his discussion of approaches to validating inferences made from categorizations using cut scores based on standards. The definitional approach seems subsumed by the psychometric.

**Legal Perspective**

Performance standards are a part of a decision-making process. Assuming the decisions have importance (i.e., stakes, which imply the possibility of harm), the process may be held to criteria that courts have determined are crucial for legal acceptability. Considerations such as equal protection (fairness) and due process (including adequate notice) have been applied through judicial proceedings to high-stakes decisions about students (Phillips, 2001) and may in the future be applied to decisions about schools (Parkes & Stevens, 2003). Along with the criterion of having a legitimate purpose, the legal perspective has been proposed by Cizek (1993) and further explored by Camilli, Cizek, & Lugg (2001). I place these criteria here on the continuum as more useful to the sponsoring organization because they subsume the psychometric criteria.
The Institutional Perspective

The four general criteria proposed in this section seem most salient from the point of view of the organization sponsoring (and likely having the eventual responsibility to implement) the performance standards (and cut scores, which are part of the standards’ operationalization). They also seem to subsume the other three perspectives. I describe briefly the four institutional criteria following. I then explore their implications.

**Consistency with policy goals.** Standards are developed for a purpose, although there may be a variety of goals that different stakeholders expect them to serve. As Kane (2001) pointed out, standards are justified in part by their relationship with legitimate goals. This suggests that the sponsoring organization, or legitimate authority (Haertel, 2002) has intended outcomes (perhaps unstated) that can be used as criteria for making judgments about the eventual success of the standard-setting process. It is against this criterion that the concept of moderation of standards (Lissitz & Huynh, 2003; discussed later) seems to fit best because moderation may or may not be implied by the policy goals of the sponsoring organization.

**Legal defensibility.** To the extent that important, high-stakes decisions are made, there is the likelihood that actions will be taken that subject the standard-setting process to an evaluation based on legal principles. Some of these have been mentioned previously. It is natural that the sponsoring organization should desire as much assurance as possible that its use of the resulting standards will be upheld if challenged in court.

**Generation of assets.** The sponsoring organization has a need for means in order to achieve public or political support for the use of the cut scores that operationalize its standards. Assets for acceptance by stakeholder groups may be of several types, including persons involved in the process and outside experts (e.g., psychometric and content-area professionals), materials such as technical manuals, and research into such aspects as consistency with other indicators and consequences for students and for programs. The latter might generate data to evaluate consistency with policy goals where they may exist (I use illustrative state data as an example).

**Efficient use of resources.** The sponsoring organization will want to spend its resources wisely. There may even be a need to account publicly for how prudent it has been.
IMPLICATIONS OF THE INSTITUTIONAL CRITERIA

Are the institutional criteria suggested previously consistent with or do they differ from current thinking about “best practice” in standard setting? What do the institutional criteria imply for the conduct of standard-setting studies? Hambleton (2001) posed 20 questions that appear to cover a broad range of activities that various authors have determined to be useful in standard setting. I use them here, restated as topics and suggestions, as a comprehensive and up-to-date list of recommendations to which the institutional criteria might be applied. As seems appropriate, I then elaborate each as suggested by the institutional criteria to see if the criteria are helpful. I raise two additional issues from the standpoint of the four institutional criteria.

I review Hambleton’s (2001) 20 evaluation criteria from the perspective of the sponsoring organization in the following sections of this article. The model used is that of a state department of education interested in setting standards for school and student accountability. Clearly other organizations, such as professional certification boards and national and international testing associations, are also sponsoring organizations for which the institutional criteria may well be useful. Testing the application of the institutional criteria to these other agencies is beyond the scope of this article but would be a helpful expansion.

CRITERIA FOR STANDARD SETTING

The institutional perspective suggests that two topics that appear outside of Hambleton’s (2001) compendium be introduced before those proposed by Hambleton. These are (a) goals clarification and (b) constituent identification. I add them here because they are best implemented in the beginning phases of a standard-setting project.

Goals Clarification

So that the resulting standards are consistent with the sponsoring agency’s policy goals, it is necessary that those goals be explicitly stated. Although this activity may be accomplished during or even after the panels have completed their work, it seems more helpful if goals clarification were to be done beforehand. Only then can the goals be used to make judgments about the adequacy of the standard-setting process.

Constituent Identification

It is important to note that Hambleton (2001) did not specify an audience for his questions (i.e., those who he intended should be convinced that the answers to the
Generating assets for achieving public support is easier if the public constituencies who need to be convinced are already defined. Knowing what each will want to know and what evidence each will find convincing should be suggestive of the sorts of assets that need to be generated to lead to their support.

**Composition of the Panel**

Hambleton (2001) was concerned about whether the groups who ought to be represented were represented by qualified individuals. Because it goes directly to validity, this is primarily a legal defensibility issue. Haertel (2002) suggested that an array of stakeholder groups should have impact on both the content and achievement standards that are implemented through the use of cut scores. However, it may also be considered a development of assets issue because the composition of the panel may be judged in terms of the groups who would best make the case to the public that the standards should be supported. Besides educators, the producers of graduates, those who further educate or employ graduates would be natural groups to consider. This suggests including employers and university educators in setting standards for high school students (perhaps with an emphasis on those who employ and train those candidates who are minimally successful) and including teachers from higher grades when setting standards for elementary and middle school students. Those who provide resources to the sponsor (e.g., political leaders, who also help shape public opinion) are another group whose support would be useful, along with persons who have credibility in making recommendations about special populations, such as advocates for persons with disabilities, the limited-English proficient, and demographic minorities, especially those who may be adversely affected by the assessment program. Whether including these groups is feasible within the resources of the sponsor and whether including them on panels is the best way to generate their helpful support are open questions, but including them on panels can be a viable way to accomplish the purpose.

**Size of the Panel**

Hambleton (2001) noted that the panel should be large enough to represent all appropriate stakeholder groups (while assuring that the subject-matter knowledge of all persons is sufficient to enable them to make the needed ratings). Including persons from each of the groups identified previously is likely to stretch the capacity of many sponsors, and compromises may be necessary. When that happens, a process to make these compromises that will demonstrate that all relevant constituencies were allowed an effective voice in reaching a consensus on participation decisions could be helpful to the sponsor in promoting the eventual outcome.
Size is also a statistical issue. A cut score that results from the standard-setting procedure could be thought to estimate the cut score that would in theory result from repeated replications of the procedure with randomly equivalent samples. The standard error of that process is estimable (see the next section) and decreases as sample size increases.

Estimation of Standard Error Across Panels

Two types of standard errors are commonly reported in standard-setting studies: standard errors of panels and standard errors of measurement. The latter are available from psychometric analyses that are independent of the standard-setting study. Standard errors of panels, which are the focus of Hambleton’s (2001) question, are commonly estimated through subdividing the panel into two or more subpanels. In practice, the two standard errors are often combined and used in deliberations subsequent to a given study to provide a range of values (e.g., the recommendation plus or minus two of the combined standard errors) that may be used in making adjustments.

Funding Adequacy

As Hambleton (2001) pointed out, standard setting can consume significant resources. The greater the importance of the decisions made using the standards, the more justifiable are the expenses needed for the process. Hambleton’s concern about sufficiency of funding is appropriate. However, virtually any sponsor has obligations to others to spend resources wisely. Developing a justification that balances costs against value of the product for the level of funding committed to the standard-setting procedure and ancillary activities could provide useful documentation that the efficiency criterion was met.

Field Testing

Hambleton (2001) noted that although uncommon, field testing the study can and likely will yield revisions to the process, and thus it is formative. The tryout can be evaluated against the four institutional criteria (or the topics mentioned here). Including at least five perspectives in the evaluation of the tryout would be helpful: (a) policy groups (for match with their goals), (b) legal staff (for legal defensibility), (c) methodologists (as respected professionals whose opinions are useful for asset generation), (d) representatives of the public and other stakeholders mentioned previously so that other members of the public will believe that all important perspectives were represented (useful for asset generation), and (e) the tryout participants themselves so that those most closely involved in the judgment process have positive opinions (also useful for asset generation).
Appropriateness of the Method

Not only should the method be described in enough detail that it can be replicated (suggested by Hambleton, 2001), but considerations that led to the choice of method in the first place can also become a material asset in communicating the results to stakeholders.

Preparation of the Participants

Hambleton (2001) described two aspects of preparation: (a) an explanation of the purposes and uses of the assessments and (b) exposure to the assessments and how they are scored. Participants who are well prepared should be more likely to support the standard-setting results, whereas participants who were not aware of the nature and consequences of the assessment program might very well disavow the eventual standards when they realize they were making judgments for a context they did not fully understand.

This situation seems akin to the informed consent process in doing research with human participants; the participants must be aware of all factors that might affect their recommendations. When are the assessments administered? What are the stakes for the student? What are the stakes for the school and/or district? Who makes decisions about consequences, how are they made, and when are they reported? When do these consequences begin? What materials exist or are being developed to help teachers reach their and their students’ instructional goals? Are there remediation opportunities? What is new and what is unchanged from that which currently exists? Knowing answers to questions such as these should increase the likelihood that the participants will make better recommendations (i.e., recommendations that are consistent with policy needs) and will support the standards that are eventually set, thus becoming more valuable as assets for generating public acceptance. To this end, it would be helpful to have the script (or talking points) used to brief the participants incorporated into any field testing.

Understanding features of the test development process would also be helpful to the participants in their deliberations as well as useful in helping them to act as assets after the procedure is completed. Among these are the process used to develop the content standards, how the test items are developed and field tested (including reviews, both substantive and statistical, of items for quality, sensitivity, and bias), and how the tests are constructed, administered, scored, scaled, and reported.

Collection of Panel Qualifications and Demographics

Hambleton (2001) noted that information about the participants’ qualifications and relevant demographics, as well as their motivation for participation, is needed for
documentation. Other relevant information that can enhance public support is how the participants were chosen and solicited.

**Exposure of the Participants to the Assessment**

Hambleton recommended that the participants actually complete the assessment under conditions that simulate the actual testing and most standard-setting procedures now include this feature, although it can be a challenge for some assessment administration methods (e.g., computer adaptive testing) or test lengths (e.g., tests that are given over long periods of time or in more than one session). Yet participants need an understanding of the actual testing context and tasks that students face.

**Training of Participants on the Standard-Setting Method**

The adequacy with which participants are trained to perform their tasks is crucial. Hambleton (2001) also noted that participants themselves and an independent evaluator are two sources of data that can be used in current as well as in future standard-setting studies to evaluate the quality of that training. Checking participants’ understanding throughout the process and having personnel available to correct misperceptions can lead to improved judgments and can leave participants with a more positive impression, enhancing their value as advocates for the standards.

**Development of Performance Category Descriptions**

Hambleton (2001) noted that the use of agreed-upon performance category descriptions by participants in the standard-setting process is a recently introduced and positive feature. Although it is reasonable that a consensus about performance category descriptions reduces interjudge variability, several options exist. There are at least four important (interrelated) dimensions in developing performance category descriptions: target, specificity, source, and timing.

*Target.* Performance category descriptions are intended to specify what students in each of the categories are expected to know and be able to do. However, each category is actually a range of degrees of achievement. As Lewis and Green (1997) pointed out, some standard-setting procedures target the average student(s) within the range and others, sometimes called borderline descriptions, target the lowest-performing student(s). Other locations are possible, though less likely. Lewis and Green recommended that performance-level descriptions target the minimum levels of performance in each category, in part so that the description represents knowledge and abilities common to all students in that range.
**Specificity.** Category descriptions may be content-specific or general. General descriptions (e.g., “ready to study the material at the next level”) do not describe what examinees know and can do and therefore require elaboration. They are often used as preliminary category (or achievement-level) descriptions at the beginning of a standard-setting study. Content-specific descriptions, such as those that are almost universally developed by the end of the study, specify the understandings and skills that examinees in each category should possess. Although sometimes developed by the participants, these have on occasion been developed by an outside group, making the participants responsible only for judgments about which of the category descriptions is represented by a given student performance.

**Source.** If content-specific performance category descriptions are developed prior to (or perhaps during) the study, then there may be some group other than the standard-recommending panel that is involved. In that event, issues related to the size and composition of the description-recommending panel parallel those of the standard-recommending panel. The stability of their judgments is also a (researchable) question that naturally arises. On the other hand, if content-specific category descriptions are developed during the study, then the standard-setting panel becomes (or is likely to be) the description-recommending panel.

**Timing.** Preliminary category descriptions may be developed prior to or during the standard-setting recommending study. If the latter, they may be developed at the end or while the study is in progress. Developing them during the study can be attractive. Indeed, continually refining the category descriptions (or borderline descriptions) may be an effective way for facilitators to focus the discussions that arise in panels at each of the rounds (Loomis & Bourque, 2001).

Utility of Feedback to Participants
Between Rounds of Judgments

For iterative standard-setting processes—which most are—this topic refers to how feedback was presented to the participants and how well they were able to make use of it. Hambleton (2001) suggested that this topic could be evaluated in two ways: postquestionnaires and low recommended cut-score standard errors. This topic seems more associated with the generation of assets and legal defensibility criteria.

**Efficiency**

There are two aspects to efficiency. First, as Hambleton (2001) noted, the study should proceed smoothly. The materials should be clear, participant time should be used effectively, and transitions should flow easily. These are aspects that are most
closely associated with the generation of assets criterion and for which a field test can be especially helpful.

Another aspect to efficiency is effective use of the sponsor’s resources. For example, not all questions and assumptions need be addressed in each standard-setting implementation. Insights developed from research in similar contexts are becoming more and more available as the field matures and these can be used to justify commonly accepted methods. For example, it is sometimes recommended that multiple approaches be used in standard setting to show that the results generalize across methods. However, including multiple approaches in each study can be and likely is beyond the resources of most sponsors (which is another way of saying that the funds could be used more effectively in other ways). Instead, literature that evaluates consistency between procedures could be noted in a technical manual. The efficiency criterion suggests broadening this topic to include cost effectiveness.

Grounding With Performance Data

Hambleton (2001) recommended that participants have data on how well examinees performed at the item (e.g., \( p \) values) and test (e.g., deciles) levels. Hambleton suggested that the goal should be to provide helpful but not directive information that makes judges aware of typical levels of responding by examinee groups. This topic is closely related to the next.

Use of Impact Data

Participants are usually given data that show the consequences of their interim recommended cut points on examinees and/or examinee groups. Hambleton (2001) recommended that the participants be instructed on how to use the data so that they could decide whether they need to revise their recommendations on the basis of reasonableness.

There may be a tension within panels, particularly if the participants have been chosen broadly. Some members may tend to try to ground their recommendations with consequences and others with content-related understandings and skills. Facilitators are commonly asked to remind participants that they are being asked to base their recommendations on comparisons of the content standards with examinee descriptions. However, impact information, once given, is difficult to ignore, and informed negotiations between participants with different orientations could actually be a healthy process toward reaching publicly supportable performance standards and assets to help generate that support.

How Final Standards Were Achieved

One or more groups normally review final standards following the participants’ recommendations. Hambleton (2001) noted that whatever steps and processes are
used, they should be clear to the participants and explainable to important boards and agencies. Awareness on the part of the judges of the possibility that their recommendations will be amended seems important to maintaining their good will following the study and therefore their value as assets.

Participant Evaluation

Hambleton (2001) pointed out that positive participant evaluations about their training, the method, and the category descriptions can be used in defending the performance standards, a direct instance of the development of assets criterion. Ensuring that they have ample opportunities to raise concerns about any of these aspects during the study can make positive evaluations more likely. As long as the participants do not learn anything new following the study that might have affected their recommendations, they are likely to continue to support the resulting standards and thus continue to be useful in gaining public acceptance.

Validity Evidence for the Performance Standards

Kane (2001) defined two characteristics on which the use of cut scores is based. He called them the policy assumption (that the standards are appropriate for the decision, which can be associated with the consistency with policy goals criterion) and the descriptive assumption (that each cut score can be used to separate examinees into those who do and those who do not meet the performance standard, which is associated with the legal defensibility and generation of assets criteria). Hambleton (2001) suggested that validity evidence in support of these assumptions should be developed along Kane’s three evidentiary sources: procedural (quality of the design and implementation of the standard-setting procedure), internal (reliability of classifications), and external (evaluating classification decisions based on other relevant evidence about student performance). As Kane noted, consequential evidence (e.g., competencies or lack thereof on the part of successful candidates, presence or absence of unfair adverse impact) is needed as well. Although these recommendations are consistent with the psychometric perspective, they are no less important from the institutional perspective. This topic is clearly broad and seems to apply to all the institutional criteria except efficiency.

Documentation of the Process

As Hambleton (2001) noted, all the topics thus far can be evaluated individually, and a technical report should be prepared to document success. The report, itself can become a material asset for the sponsor. This task could be the responsibility of the sponsor (especially because several decisions that need to be documented are
commonly determined even before a request for proposals [RFP] is issued, such as the general standard-setting approach), or the sponsor may make the report the sole responsibility of a contractor, with or without specifications about what is to be included. If the sponsor retains responsibility, a straightforward way to accomplish that could be to assign the development of the technical report to in-house personnel from the beginning, with sections to be completed by the contractor(s) as noted in the RFP. It would be helpful for the assessment community to develop recommendations about what material to include and how to format it. Possible starting points for formatting would be to use the evaluation criteria or the topics explored in this article as sections for the document.

Effective Communication of Standards

The performance of examinees who fall in each of the performance categories is usually communicated using content-specific category descriptions. Hambleton (2001) suggested that although these category descriptions may be sufficient, exemplar items, either of the categories or at the borderlines, could enhance communication of what examinees in each category know and can do. Material and people used to communicate the standards become assets for public acceptance, and data showing their reasonableness and consistency with other information are research assets.

Most assessments are actually composites of highly but not perfectly correlated content strands and the use of a cut score defined on a single scale therefore represents a compensatory decision rule. Yet performance categories are typically presented as if the rule were conjunctive (i.e., students in this category can do all of these things). It would be helpful for educators, candidates, and the public (as well as the participants) to understand the compensatory nature of the use of a single cut score between categories.

MODERATION OF STANDARDS

The institutional criteria seem reasonable when compared with Hambleton’s (2001) questions from a theoretical perspective, but do they help in evaluating a standard-setting orientation? I use them to review the method known as vertical moderation of standards.

Lissitz and Huyhn (2003) proposed vertically moderated standards (VMS) for statewide tests. After reviewing the current status of vertical scaling from the perspectives of validity and practicality, Lissitz and Huyhn concluded that vertical scales are too problematic to recommend for many current situations in which they have been considered. Using the development of cut scores in South Carolina as an
example, their process for VMS as an alternative to vertical scaling seems to have these elements:

1. Determine a policy definition for proficiency (or whatever category the cut score is intended to operationalize as the minimum). This is done separately for different content areas. The policy definition should include forward-looking elements such as that the student should be able to maintain the same level of proficiency in the next year.
2. The material presented to the participants should include the curriculum in at least 2 years (the target grade and the subsequent or previous grade), the actual test items and blueprints, and scatter plots displaying the relationships between the 2 years that will help maintain (or at least understand) normative (in)consistency between the standards.
3. In the event that not all grades are available for standard setting (e.g., because development of assessments in some grades lags development in others), Lissitz and Huynh (2003) allow setting cut scores by interpolation or extrapolation of proficiency-level achievement rates (impacts) to intermediate assessments.
4. Within margin-of-error bands resulting from confidence intervals generated using standard errors of the participant-recommended cut scores, adjustments may be made based on policy considerations when the consequences justify them.
5. Annual studies should be conducted to validate the standards.

A Hypothetical Example of VMS

The promise of VMS is to be able to make inferences about growth in student achievement without vertical scaling in place. How does this concept fare when judged according to the criteria for standard setting developed from the institutional perspective? That question is discussed using examples of realistic outcomes of standard-setting procedures in four simulated but realistic states: East Pacific, North Mexico, South Canada, and West Atlantic.

Figures 1 through 8 show the hypothetical percentages of students above the minimum (i.e., proficient) cut for making No Child Left Behind (NCLB) Act (2001) decisions in reading and math for the four states. Also shown are hypothetical percentages of students above the cuts for the most recent National Assessment of Educational Progress (NAEP) results for the basic and proficient levels (upper and lower dotted lines, respectively). Finally, hypothetical ranges of two standard errors (measurement and standard-setting combined) above and below the panels’ recommended cuts are included (the dashed lines below and above the solid line, respectively). Typical of actual states, data for all grade levels were assumed to be unavailable, so for all these series, linear interpolation and extrapolation were used...
FIGURE 1  Reading results for East Pacific.

FIGURE 2  Math results for East Pacific.

FIGURE 3  Reading results for North Mexico
FIGURE 4  Math results for North Mexico.

FIGURE 5  Reading results for South Canada.

FIGURE 6  Math results for South Canada.
to generate results for all grade levels, 3 through 10. It should be noted that a policy using interpolation and extrapolation would be unfair to Lissitz and Huynh (2003) because the process they recommend involves thoughtful judgments about annual relations and these quantitative methods are to be substituted only when necessary (see point 3 previously). The figures are used here for illustration only, and fictitious names are used. Nevertheless, they are intended to be realistic and to portray fairly the challenges typical of states across the country (it would be straightforward to construct such a figure for a real state, given the appropriate data). In these diverse contexts, I turn to a discussion of each of the four institutional criteria.

**Consistency with policy goals.** The need to include forward-looking elements in the characterizations of proficiency levels is not necessarily a part of a
state’s expressed policy goals. Nevertheless, it is almost surely consistent with the state’s general intent. The assumption that standards should be moderated through interpolation and extrapolation seems a bit more problematic because there may be other considerations than smoothness of pass rates. There was remarkable consistency across grade levels in three of the states (North Mexico, South Canada, and West Atlantic), but the rates were less smooth in East Pacific. Perhaps the apparent lack of moderation in East Pacific represents some beliefs of the state, such as about the difficulties of the content standards across the grades, the comparative rigor of the performance standards, the progression of learning over time, or the relative success of current instruction at different grade levels.

Whether cuts for content areas should be set together may also be an issue for a state to consider from a policy perspective. In the figures, the percentages above cut were about the same for reading and math for two states (North Mexico and South Canada), but reading cuts seem more lenient than math cuts in two others (East Pacific and West Atlantic). Moderation across content areas may be just as important as moderation across grade levels.

A reasonable conclusion is that moderation may be a policy goal, but it competes with others. Although lack of moderation, either across grades or across contents, can lead to unbalanced consequences and asymmetric remedies, those outcomes may be consistent with policy goals. Moderation does not appear to be an end in itself.

**Legal defensibility.** Of course, the proposed moderation process has not been tested against the legal criteria of legitimacy, due process, and fairness. Can one anticipate some issues that may arise in comparison with another approach, such as setting standards using one of the currently accepted procedures? Legitimacy seems not to be a problem because the authority and purpose of the agency is the same whether or not vertical moderation is used. Similarly, fairness should not be impacted differently. However, due process may be of concern. Standards could be set based almost exclusively on normative data. Inferences about the knowledge and skills of examinees would not then have the support of professional and other stakeholder evaluations of the curriculum, its alignment with the assessments, and the achievement-level descriptions of the performance categories, including remediation recommendations. If moderation is used to set cut scores, it is likely that this area is in significant need of being addressed, such as in the program of research that Lissitz and Huynh (2003) recommend as a step in implementing their suggestions.

**Generation of assets.** Earlier, assets were broadly classified into three types: persons, materials, and research studies. If a state were to develop some of its cut scores using vertical moderation, at least some of these assets would not exist. In this analysis, I describe some of the assets that each of the four states might
have had available at each of the grade level and content combinations to use for public acceptance.

One assumes the contractors in each state documented in technical reports at the end of the study the characteristics and beliefs of the persons who were involved. Whether those beliefs were maintained after the study is an open question. Additionally, one assumes each state incorporated an outside technical advisory group that reviewed the procedures and provided evaluative commentary. These reports provide material assets for the state.

Technical reports often provide details about the development and variability of the recommended standards. These data can be compared with the adopted standards to gauge whether the cut scores actually used are consistent with the recommendations of the judges. As seen in the figures, each state in both contents set cut scores that were within two combined standard errors of the committee recommendations, with one exception (South Canada in math at the two higher grades, which are extrapolated values in the simulation). Documentation of consistency between the participants’ recommendations and the adopted standards is a valuable material asset.

In addition to the technical report of the standard-setting study, technical manuals for the tests, and the decision rules used at the state level are needed. For example, a state’s rules for determining adequate yearly progress are part of its policy implementation and are thus needed for an evaluation of the state’s accountability program, whether by an internal process or by interested members of the public. Although they are independent of the standard-setting procedure implementation, documenting these points can provide additional material assets that would be helpful for a state.

The figures also include an example of a research outcome. In each case, simulated results of state NAEP for the basic and proficient performance levels (and above) were plotted (with interpolation and extrapolation, used here only for purposes of illustration). This allows a comparison between the state’s achievement levels and those of a credible external assessment. Although these comparisons may suffer from different degrees of alignment between the state NAEP frameworks and a state’s content standards, they do provide evidence that the public might use in order to evaluate the rigor of a state’s performance standards.

These comparisons reveal some sharp differences among the hypothetical states shown in Figures 1 through 8. In South Canada, the adopted cuts for both contents appear more lenient than NAEP basic. In North Mexico and East Pacific, they appear between NAEP basic and NAEP proficient. The cuts for reading in West Atlantic are also between NAEP basic and proficient, but for math they are approximately as severe as NAEP proficient. Although the implications of these different degrees of rigor for educational impacts remain to be seen, members of the public would likely place their own values on such comparisons, and thus,
there are implications of these comparisons for public acceptance of the cut scores that operationalize the performance standards. These comparisons would be possible whether or not the cut points were determined using moderation.

**Efficient use of resources.** While gathering data to support the cost-effectiveness of a standard-setting study would be helpful to the sponsoring organization and is possible (e.g., by comparison with other organizations that have conducted similar studies), the efficiency introduced via the competitive bidding process required in most states when contracting with an external vendor for testing services may be sufficient for most audiences. It is an open question whether the cost saving provided by developing VMS rather than conducting studies at all grade levels is sufficient to compensate for the loss of person and material assets that result from the studies.

**CONCLUSIONS**

Guided by earlier work on goals for implementation of such activities, I proposed four criteria for standard setting from the perspective of the institution that will use the standards. The four criteria are (a) consistency with policy goals, (b) legal defensibility, (c) generation of assets (including people, materials, and research evidence), and (d) efficient use of resources. I reviewed the comprehensiveness and utility of these four goals using Hambleton’s (2001) extensive list of questions that should be addressed in any standard-setting implementation. It was found that Hambleton’s suggestions could be supported in terms of the goals. Further, the goals led to an expanded list of recommendations for standard-setting studies in several areas and to two new issues that were not raised by Hambleton.

Although I developed and discussed them from the perspective of a state department of education, whether the four criteria may apply equivalently in other contexts is an open question. It is possible that one or more might not apply well or that they might apply but in different ways. For example, elements needed for legal defensibility might be different for a private as opposed to a public sponsor.

There appear to be implications of all four of the institutional criteria for the advisability of using vertical moderation to establish cut scores for performance categories. Vertical moderation is certainly more efficient, but whether that is sufficient to offset the loss of assets is debatable. The consistency between VMSs and policy goals is an open question that should be addressed explicitly in each application because the goals themselves are likely to be idiosyncratic. Whether the justification of VMS is sufficient to withstand legal challenge is another unanswered question. Finally, I suggested horizontal moderation (across content areas at the
same grade) as a reasonable concept for policymakers to consider in relation to their policy goals.

ACKNOWLEDGMENT

This project was partially funded by the Maryland State Department of Education (MSDE) under a contract to the Maryland Assessment Research Center for Education Success (MARCES). The opinions expressed do not necessarily reflect those of MSDE or of MARCES.

REFERENCES


