Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process[1,2,3,4]

Ronald K. Hambleton
University of Massachusetts at Amherst

Educational assessments and credentialing examinations are often used today to classify examinees into ordered performance categories such as masters and non-masters, or Advanced, Proficient, Basic, and Below Basic. These performance categories are typically defined with respect to a well-defined domain of content and skills. The domain of content and skills for educational assessments may be the product of collaboration among curriculum specialists, teachers, and policymakers; and for credentialing examinations, may come from the findings of a job analysis or role delineation study. Performance categories rather than the test scores themselves will sometimes be a more meaningful way to communicate test results. For example, with National Assessment of Educational Progress (NAEP) score reporting, the significance of a change in the average mathematics score of 2 points between 1992 and 1996 for a group of examinees may not be understood, and the meaning of the two-point difference may be difficult to communicate clearly. *The percentage of examinees performing at the Proficient level increased from 30% to 35% in the interval between the two administrations of the assessment* may be a more meaningful way to report the results, assuming, of course, that the meaning of proficient level performance has been clearly articulated.

The use of performance categories for score reporting is not always a matter of choice; it may be fundamental to the intended uses of the scores. For example, the purpose of an assessment may be: (a) to make pass-fail decisions about examinees, as it is with many high school graduation tests and credentialing exams; or (b) to place examinees into, say, four ordered performance categories for individual or group evaluation, as it is with many state assessment programs. In these instances, performance categories are needed, along with performance standards.

[1] This paper is based on work supported by the National Science Foundation Grant No. 9555480. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

[2] Laboratory of Psychometric and Evaluative Research Report No. 377. School of Education, University of Massachusetts, Amherst, MA.

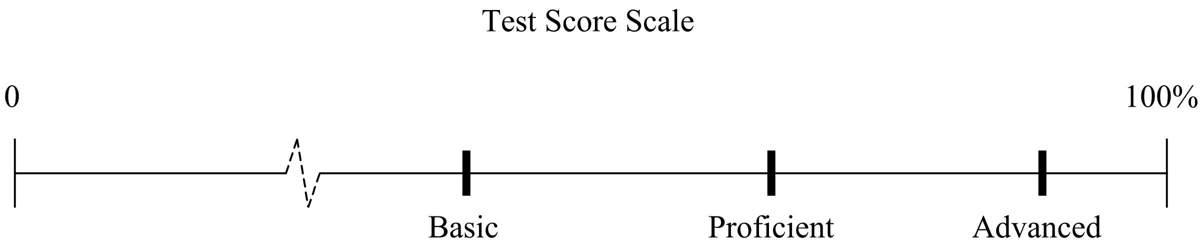
[3] This paper was given as an invited presentation at the Edward F. Reidy, Jr., First Interactive Lecture Series, Dover, New Hampshire, sponsored by the National Center for the Improvement of Educational Assessment.

[4] To appear in G. Cizek (Ed.). (in press). Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Well-defined domains of content and skills and performance categories for test score interpretation are fundamental concepts in educational assessment systems aimed at describing what examinees know and can do. The primary purpose of these assessments is not to determine the rank ordering of examinees, as is the case with norm-referenced tests, but rather to determine the placement of examinees into a set of ordered performance categories. Another important characteristic of these assessment programs, then, is *performance standards*: typically, points on a test score scale that are used in separating examinees into performance categories. Figure 1 shows three performance standards on a typical scale for reporting test scores.

Figure 1

Performance standards on the test score scale



Educational assessments are sometimes used to determine whether examinees have achieved sufficiently high levels of content and skill mastery in a subject area to be eligible to receive high school diplomas. This assessment requires a single performance standard (or alternatively called a standard, achievement level, passing score, minimum proficiency level, threshold level, mastery level, or cutoff score) on the test score scale, or the scale on which achievement is reported (more often scaled scores are used in score reporting and not the actual test scores), to separate examinees into two performance categories, often labeled masters and non-masters, passers and non-passers, or certifiable and not-certifiable.

With the NAEP and many state assessments, examinees are separated, based on their performance on an educational assessment, into multiple performance categories. In the NAEP, examinees are classified into four ordered performance categories called Advanced, Proficient, Basic, and Below Basic. In Massachusetts, the four ordered performance categories used in reporting examinee performance on the state's proficiency tests are called Advanced, Proficient, Needs Improvement, and Failing. The classification of examinees into three to five performance categories in each subject area is common with many state assessment programs. But how are the performance standards set? Are there steps that can be followed to increase the validity of the performance standards that are produced from the process?

In this paper, steps for setting performance standards on educational assessments are presented. In addition, criteria for evaluating a standard-setting study are offered. Both the steps and the criteria should be useful to testing agencies for designing and monitoring the standard-setting process to increase the chances that the process will produce defensible and valid performance standards. The paper is organized into three main sections. The first section addresses background information to provide a foundation for understanding current concepts and practices in standard setting. In the second section, 11 steps for setting performance standards on an educational assessment are offered. Each step is described in detail. In the third section, a set of 20 evaluative criteria for judging a standard-setting study are offered along with a brief discussion of each evaluative criterion.

Background

Three points are important to make at the outset. First, it is important to clearly distinguish between *content standards* and *performance standards*. There is evidence in the assessment literature that many persons, but especially policymakers, fail to correctly distinguish the two. Content standards refer to the curriculum and what examinees are expected to know and to be able to do. Examinees, for example, might be expected to carry out basic mathematics computations, read a passage for comprehension, or carry out a science experiment to identify the densities of various objects. Performance standards, on the other hand, refer to the level of performance that is expected of examinees to demonstrate, say, Basic, Proficient, and Advanced level performance in relation to the content standards. In other words, performance standards communicate how well examinees are expected to perform in relation to the content standards (Linn & Herman, 1997).

For example, we might require examinees to solve 10 of 20 mathematics computations to be judged Basic, and require that examinees solve 14 of 20 problems to be judged Proficient. In reading comprehension assessment, examinees may be expected to answer 60% of the questions to be judged Basic, 80% to be judged Proficient, and 90% to be judged Advanced. Content standards should be thought of as what we expect examinees to have learned; whereas performance standards indicate the levels of expected performance of, say, Basic, Proficient, and Advanced examinees on the educational assessments that measure the content standards. For some researchers (Jaeger & Mills, in press; Kane, 1994) a distinction is made between performance standards and cutoff scores; cutoff scores are defined as the points on the test score scale separating examinees into performance categories, and the performance standards correspond to the performance category descriptions of these cutoff scores. The distinction introduced by Kane (1994) between performance standards and cutoff scores is not adopted in this paper, as it is still uncommon to do so in the standard-setting literature. At the same time, it should be noted that the distinction has been especially helpful in generating validity questions and conducting validity investigations.

Performance standards are usually scores on a test score scale, but not always. They may correspond to verbal descriptions that can be used in classifying examinee test performance into performance categories. For example, the scoring rubrics for a writing assessment might define advanced, proficient, basic and below basic writing, and then examinees are classified by raters by matching examinee written work to the verbal descriptions of four levels of writing proficiency.

Second, all standard-setting methods involve judgment. With the contrasting-groups method, judges must decide which examinees are in each performance category (see Clauser & Clyman, 1994); with the Angoff method, panelists must estimate the performance level of borderline candidates (Angoff, 1971); with the paper selection method, panelists must sort examinee work into performance categories (Plake & Hambleton, 2000); and so on. Judgment is also involved in deciding the composition and number of panelists who will be asked to participate in a standard-setting process, what the process should be, whether or not panelists will be provided feedback on their own ratings and the ratings of other panelists, and even the form of that feedback. All of these judgments, and many more, are an integral part of the standard-setting process. Standard setting is mainly a judgmental process. It is for this reason that factors such as the selection of panelists, the training of panelists, and the processes that panelists are asked to follow during the course of setting standards are central in the overall evaluation of a standard-setting process and the defensibility and validity of the resulting performance standards.

The point is not to disparage performance standards because judgments are involved; in fact, judgments are involved in every aspect of education including the specification of curriculum frameworks and content, the choices of textbooks and other instructional materials, and the selection of optimal teaching methods to match examinee learning styles and aptitudes. The point is that care needs to be taken about who provides the judgments for setting the performance standards and the context in which those judgments are provided and interpreted.

Finally, methods for setting performance standards on educational assessments using the multiple-choice item format are well developed and steps for implementation are generally clear (see Livingston & Zieky, 1982). Most districts and states and credentialing agencies have set defensible performance standards using one of the acceptable methods (e.g., Angoff, 1971; see Berk, 1986, for an excellent review; Ebel, 1972). On the other hand, standard-setting methods for educational assessments that include constructed response items such as writing samples and performance tasks are not as well developed at this time, and certainly none of them have been fully researched and validated (see Hambleton, Jaeger, Plake, & Mills, 2000a, 2000b). Readers are referred to Hambleton, et al. (2000a) for a review of methods and issues for setting standards on educational assessments involving constructed response items.

Typical Steps in Setting Performance Standards

Any defense of performance standards in educational assessment begins with a defense of the full standard-setting process itself. It is important to document that a reasonable, systematic, and thoughtful process was followed in arriving at the final standards (Hambleton & Powell, 1983; Plake, 1997). The defensibility of the resulting standards is considerably increased if the process reflects careful attention to: (a) the selection of panelists; (b) training of panelists; (c) the sequence of activities in the process; (d) validation of the performance standards; and (e) careful documentation of the process. If, on the other hand, panelists are chosen because, for example, they live near the meeting site, they demand to be on the panel, or they happen to be known by the coordinator of the meeting, or a process was implemented that did not allow panelists to carefully consider their judgments or if the panelists had reservations about the process, questions would be raised about the validity of the resulting performance standards. Other common problems that can reduce the validity of the performance standards include: (a) the use of ambiguous descriptions of the performance categories; (b) failure to train panelists fully on the standard-setting method; (c) failure to allow sufficient time for panelists to complete their ratings in a satisfactory manner; and (d) failure to validate and document the process that was implemented to set the performance standards.

A presentation and discussion of 11 steps for setting performance standards follow; the steps are summarized in Table 1. These steps generally apply to standard-setting methods that focus on judgments about the assessment items and associated scoring rubrics, examinee work, or examinee score profiles (e.g., Angoff, Ebel, Nedelsky, paper selection, booklet classification, bookmark, body of work, dominant profile, and more). The steps are not completely applicable to methods such as the contrasting groups method that are focused on judgments about candidates.

Table 1

Steps for Setting Performance Standards on Educational Assessments

1. Choose a panel (large, and representative of the stakeholders).

2. Choose one of the standard-setting methods, and prepare training materials and finalize the meeting agenda.

3. Prepare descriptions of the performance categories (e.g., basic, proficient, and advanced).

4. Train panelists to use the method (including practice in providing ratings).

5. Compile item ratings and/or other rating data from the panelists (e.g., panelists specify expected performance of examinees at the borderlines of the performance categories).

6. Conduct a panel discussion; consider actual performance data (e.g., item difficulty values, item characteristic curves, item discrimination values, distractor analysis) and descriptive statistics of the panelists' ratings. Provide feedback on interpanelist and intrapanelist consistency.

7. Compile item ratings a second time that could be followed by more discussion, feedback, and so on.

8. Compile panelist ratings and obtain the performance standards.

9. Present consequences data to the panel (e.g., passing rate).

10. Revise, if necessary, and finalize the performance standards, and conduct a panelist evaluation of the process itself and their level of confidence in the resulting standards.

11. Compile validity evidence and technical documentation.

Choose a panel (large and representative of the stakeholders). Who are the stakeholders in the decisions that will be made with the educational assessments? These are the persons who should be involved in the standard-setting process. In the case of NAEP, teachers, curriculum specialists, policy makers, and the public (30% of the panels by policy) make up the standard-setting panels. With many state assessments, standard-setting panels consist of state teachers, curriculum specialists, and school administrators. Representatives of the public sometimes are included with high school graduation tests.

In the typical state assessment situation, 15 to 20 persons are often placed on a panel to provide the diversity that is needed (geographical, cultural, gender, age, technical background, educational responsibilities), and to provide stable estimates of the performance standards (Jaeger, 1991; Jaeger & Mills, in press; see also Raymond & Reid, in press). The composition of the panel is arbitrary but it is important for the agency setting performance standards to be able to demonstrate that the issue of composition was considered, and that there was a rationale for the composition of the panel that was formed.

One of the questions that is raised more often today than ever concerns the impact of the particular choice of panelists on the final performance standards. The question is, "Were a second sample of panelists to be drawn with the same characteristics as the first, would the second sample produce a similar set of performance standards?" If it cannot be demonstrated that similar performance standards would result with a second panel, the generalizability of the performance standards is limited, and the validity of the performance standards is significantly reduced.

There are at least two consequences of the current demand in the education field to demonstrate performance standard generalizability over panels. First, extra panelists can be selected—at least twice the number assumed appropriate to set performance standards. Second, in designing a study to address performance standard generalizability over panels, two separate panels of roughly the same size are needed. The ideal is to conduct separate meetings, but it is common because of the cost to hold a single meeting, provide a common orientation, and common training to both groups of panelists, and then split them up for the remainder of the meeting. The result is two sets of independently derived performance standards that could be compared.

A popular and less costly variation on designs to investigate generalizability is to take a single panel and form smaller independent subpanels. The stability of the performance standards across independent subpanels is determined at the end of the standard-setting process and used as one of the evaluative criteria. This second design allows for a check on generalizability of performance standards over panels, but uses subpanels that are smaller than those recommended to actually set performance standards. Final performance standards are obtained by averaging the performance standards set by each subpanel. Sometimes, too, studies are designed to investigate both generalizability of performance standards over panelists and over parallel forms of the assessment.

Choose one of the standard-setting methods, and prepare training materials and finalize the meeting agenda. There are many popular methods for setting performance standards. Hambleton, et al. (2000b) offer a classification scheme of methods that is based on the nature of the task posed to panelists:

(1) Make judgments based upon a review of assessment material and scoring rubrics [the Angoff method and variations fit this category (e.g., Angoff, 1971; Cooper-Loomis & Bourque, in press; Ebel, 1972; Hambleton & Plake, 1995) along with the popular bookmark method (see, Mitzel, Lewis, Patz, & Green, in press)];

(2) Make judgments about examinee work [the paper selection, whole booklet classification, analytical judgment, and whole body of work methods fit here (Jaeger & Mills, in press; Kingston, Kahl, Sweeney, & Luz bay, in press; Plake & Hambleton, in press)];

(3) Make judgments about score profiles [the dominant profile method and the policy-capturing method fit in this category (Jaeger, 1995; Plake, Hambleton, & Jaeger, 1997)]; and,

(4) Make judgments about the candidates [the contrasting groups method would fit here (see for example, Jaeger, 1989; Livingston & Zieky, 1982)].

For additional descriptions and reviews of available standard-setting methods, readers are referred to Berk (1986), Hambleton, et al. (2000a, 2000b), Jaeger (1989), and Livingston and Zieky (1982).

Choice of method might be based on: (a) the mix of items in the assessment (e.g., with multiple-choice tests, the Angoff method has been popular; with performance assessments, paper selection, analytic or bookmark methods might be a suitable choice); (b) time available to set standards (e.g., in the information technology [IT] industry, the choice of method needs to require very little time, perhaps only a few hours); (c) prior experience with the method (e.g., prior experience with a method may reduce the need for field-testing which can be costly and time-consuming); and (d) perceptions and/or evidence about the validity of the method (for example, some researchers today would avoid the Angoff method because of concerns about its validity; other researchers have been critical of the contrasting groups method). Readers are referred to Kane (1994) for a good discussion of some of these criticisms of current methods.

It is especially important to use training materials that have been field tested. For example, a miscalculation of the time required to complete various steps in the process may result in panelists needing to rush their ratings to complete their work on time. This problem arose on the first initiative to set performance standards on NAEP and 60 panelists scattered around the country needed to be recalled for a second meeting (Hambleton & Bourque, 1991).

When multiple facilitators are needed, training is important to insure that they handle their panels in much the same way. Pacing and handling of discussions, answers to common questions (e.g., How should the probability of guessing a correct answer be considered in Angoff ratings? What should I do if I think the scoring rubric is problematic?), and order of presentation of information (e.g., panelist ratings, statistical data on the items/tasks), need to be standardized, to avoid inflating differences in subpanel performance standards.

Prepare descriptions of the performance categories (e.g., basic, proficient, advanced). In recent years, time spent defining the performance level descriptions has increased considerably in recognition of the importance of the descriptions for producing valid performance standards (see, for example, Mills & Jaeger, 1998) and for communicating information about the meaning of the performance standards. These descriptions may apply to the performance categories, or sometimes they are prepared to describe examinees at the performance standards of interest. The descriptions consist of statements of the knowledge, skills, and abilities of examinees who would be in each of the performance categories or at each of the performance standards. This focus on clear descriptions is one of the most important advances in recent years in standard setting.

In setting performance standards on the NAEP, for example, more than two full days are spent on the performance level description process. If panelists are to set defensible performance standards, the belief is that performance levels need to be clearly articulated. Panelists are requested to consider the performance of borderline examinees on the assessment material or they may be required to classify examinee work using the performance level descriptions. When these descriptions are unclear, panelists cannot complete their tasks and the resulting performance standards could be questioned. For example, it is not possible to confidently sort examinee work in the booklet classification method into performance categories if the performance category descriptions are unclear.

A critical step in the process, then, is for the panel (or a prior panel) to develop descriptions of examinees in each performance category. Recently, Mills and Jaeger (1998) produced the first published set of steps for producing test-based descriptions of performance levels, and these steps will be of interest to readers. Other times, more generic descriptions are used (see, for example, Cooper-Loomis & Bourque, in press). There is some evidence to suggest that lower performance standards may result with test-based descriptions (Mills & Jaeger, 1998), but this result needs to be replicated, and then if differences are found, comparative studies of the impact of generic versus test-based descriptions of performance categories on the validity of performance standards need to be carried out.

Appendix A provides an example of descriptions of novice, apprentice, and proficient examinees from the Pennsylvania Grade 8 Mathematics Assessment. Appendix B presents the descriptions for basic, proficient, and advanced examinees used in the setting of grade 4 performance standards in the area of Reading on the NAEP. These descriptions provide an idea of the level of detail that is assumed necessary for panelists to complete their rating tasks.

In one currently popular standard-setting method (the bookmark method; see Mitzel et al., in press), the descriptions of the performance categories are not fully developed until the end of the process. The rationale for this view is that panelists are in the best position following their efforts to set performance standards to develop the descriptions. The impact of this decision to place the development of detailed descriptions of the performance categories at the end of the process on the location of the performance standards themselves would be a topic worthy of investigation. Does this decision affect the resulting performance standards, and if so, are the resulting standards more or less valid than standards set with the descriptions being developed at the beginning of the process?

Nellhaus (2000) reported that 18 states were currently using some version of the bookmark method, and so research on this point seems especially timely.

Train panelists to use the method (including practice in providing ratings). Panelists will need effective training and practice exercises to set defensible and valid performance standards. Effective panelist training would include:

(1) Explaining and modeling the steps to follow in setting standards (e.g., estimating the performance of borderline candidates, or sorting examinee papers into ordered categories);

(2) Showing the scoring keys and/or scoring rubrics and insuring they are understood;

(3) Completing easy-to-use rating forms;

(4) Providing practice in providing ratings;

(5) Explaining any normative data that will be used in the process, and so forth;

(6) Familiarizing panelists with assessment content (e.g., the assessment tasks);

(7) Developing borderline descriptions (if used);

(8) Taking the test under standard or near standard conditions; and,

(9) Reviewing the item pool on which the performance standards will be set.

It is not uncommon for training to take at least one day. When the assessment and scoring are complex, considerably more time may be needed. With the policy-capturing method (Jaeger, 1995), applying the method was easy, but two to three days of time were needed to explain the complex assessments on which the standards were going to be set. Each standard-setting study is unique, but appropriate training is required, regardless of the time required.

In addition, panelists need to be informed about factors that may impact on examinee performance and should be considered in the standard-setting process. Such considerations would include, for example: (a) the role of time limits for the assessment; (b) the artificiality of educational assessments (panelists need to remember that when an examinee chooses to write a story, the examinee will often select the topic, have unlimited time to complete the work, and will often prepare several drafts [characteristics which are often not present in the typical writing assessment]); (c) distractors in multiple-choice items that may be nearly correct (and, therefore, increase the difficulty of the item for examinees); and (d) the role of guessing behavior on performance of examinees on multiple-choice items, and so on.

Also, administering the assessment to panelists is often an effective way to demonstrate to them the knowledge and skills that examinees must possess to obtain a high score. It is assumed that panelists are likely to set more realistic performance standards if they have experienced the assessment themselves. The assessments always appear more difficult to panelists when they are completed without the aid of the scoring keys and scoring rubrics!

Finally, it is important that panelists understand their relationship to the board or agency to which the performance standards will be forwarded. Often the performance standards they prepare are only advisory to the board or agency; the board or agency reserves the right to revise the performance standards as they feel appropriate. Panelists must have this information. When boards or agencies revise the performance standards, questions can be raised about the basis for the changes (e.g., Pelligrino, Jones, & Mitchell, 1999; Reckase, 2000). Often the basis for changes is related to information the board or agency may have about the consequences of applying the performance standards, or the board or agency may be interested in building in consistency and coherence among multiple sets of performance standards. For example, with NAEP, performance standards are set in many subject areas and at grades 4, 8, and 12. Unless there is some consistency and coherence to the results across grades in each subject, and even across subjects, meaningful interpretations of the complex array of data become very difficult. In another example, an agency such as a state department of education may want to lower performance standards to reduce the numbers of children requiring special services to a level that can be accommodated with existing funds.

Compile item ratings and/or other rating data from the panelists (e.g., panelists specify expected performance of examinees at the borderlines of performance categories). This step is straightforward if the training has been effective. A summary of the panelists' ratings can be prepared. For example, suppose panelists are asked to judge the minimum expected performance of proficient examinees on a task with a 5-point scoring rubric (e.g., 0 to 4). The median or typical rating and the range of ratings of the panelists could be calculated. Later (step 6), this information can be provided to the panelists and used to initiate discussion about the performance standard for proficient examinees. Reckase (in press) provides more details on the topic of intrapanelist and interpanelist feedback.

Perhaps, in time, panelists themselves can enter their ratings into a computer to speed up data processing. From our experience, data entry of panelists’ ratings is typically a bottleneck and often meetings are scheduled so that lunchtime, breaks, and the evening of the first day can be used for data entry, but this is not always possible. Some standard-setting teams ask panelists to record their ratings on machine-scanable sheets, which can speed up the data entry process considerably, but often constraints must be placed on the panelists’ ratings to match the options on the scan sheet. Scanners are not completely reliable either, and we have seen major delays in standard-setting meetings due to their failures.

Conduct a panel discussion: consider actual performance data (e.g., item difficulty values, item characteristic curves, item discrimination values, distractor analysis) and descriptive statistics of the panelists' ratings. Provide feedback on interpanelist and intrapanelist consistency. With several of the standard-setting methods, panelists are asked to work through the method and set preliminary standards, and then to participate in a discussion of these initial standards and actual examinee performance data on the assessment. The purposes of the discussion and feedback are to provide opportunity for panelists to reconsider their initial ratings and to identify errors or any misconceptions or misunderstandings that may be present. The precise form of the feedback depends on the method, but with several methods, the feedback might include average performance and examinee score distributions on the items or tasks of the assessment, and descriptive statistics of the ratings of the panelists.

More elaborate forms of feedback are also possible. For example, it is possible to determine the extent to which panelists are internally consistent in their ratings (van der Linden, 1982). Panelists who set higher performance standards on difficult tasks than easier tasks would be identified as being "inconsistent" in their ratings. They would be given the opportunity to revise their ratings or explain the basis for their ratings. Sometimes the so-called inconsistencies in the ratings can be defended, but regardless, panelists would rarely be required to revise their ratings if they were comfortable with them. For a full review of factors affecting ratings, readers are referred to Plake, Melican, and Mills (1991).

The impact of the feedback and discussion may be more psychological than psychometric. Often, the main impact is consensus among the panelists. The variability of the panelists' choices of performance standards is decreased, but the performance standards themselves often remain about the same. But the performance standards do not always remain the same and so the iterative process seems worthwhile (see, for example, Plake & Hambleton, in press). Also, we have observed in some of our own studies (Plake & Hambleton, 2000) that panelists feel more confident about the resulting performance standards if there has been discussion and feedback.

Compile item ratings a second time that could be followed by more discussion, feedback, and so forth. This iterative process is common but not essential. Typically, a two-stage rating process is used: Panelists provide their first ratings (independent of other panelists or performance data of any kind), discussion follows, and then panelists complete a second set of ratings. Following the discussion phase of the process, panelists are instructed to provide a second set of ratings. It is not necessary that panelists change any of their initial ratings, but they are given the opportunity to do so. Sometimes this iterative process is continued for another round or two. For example, in some of the NAEP standard-setting work that has been done (see Hambleton & Bourque, 1991; Reckase, 2000), panelists went through five iterations of ratings and discussions.

Not all standard-setting researchers are committed to the use of discussion and feedback in the process. For example, with performance assessments, some researchers such as Jaeger and Mills (in press) argued that better (i.e., more stable) performance standards will result if panelists spend the available time rating more examinee responses rather than participating in discussions and review of statistical data. They take this position because of the knowledge that the typical influence of discussion and review anyway is to reach consensus. Performance standards rarely change between iterations. The competing argument is that it is important for panelists to discuss their ratings and receive feedback. Sometimes discussion and feedback will alter the performance standards, and even small changes in the standards, up or down, can be of practical consequence. Furthermore, standard errors are almost certainly lower, and discussion and feedback may increase panelist confidence and acceptance of the resulting performance standards. This is an area where more research would be helpful.

Panelists like this step very much (or at least they report that they do on postevaluations), appreciate the opportunity to discuss their ratings with their colleagues, find the feedback valuable, and sometimes performance standards do shift significantly up or down, especially when the feedback is a surprise to panelists (see Hambleton & Plake, 1995; Plake & Hambleton, 2000).

Compile panelist ratings and obtain the performance standards. At this stage, panelists' ratings are compiled to arrive at the performance standards. Often, this is an average of the performance standards set by each panelist. Median ratings may be preferable with small samples or nonsymmetric distributions of panelist ratings. It is very common to report the variability of the performance standards across subpanels or panels. This error may be used in deciding on the viability of the resulting performance standards; a large error will lead to suspicion about the standards, a small error builds confidence in the standards. This error, too, may be used in adjusting a performance standard.

Present consequences data to the panel (e.g., passing rate). One step that is sometimes inserted into the process involves the presentation of consequential data to panelists. For example, a panel might be given the following information on the percentage of examinees that would be classified into each category if the current performance standards were applied:

Category	Percent of Examinees
Advanced	7.0%
Proficient	33.2%
Basic	42.5%
Below Basic	17.3%

If these findings were not consistent with the panelists' experiences and sense of reasonableness, they may want an opportunity to revise their performance standards. For example, panelists may feel that a performance standard that resulted in, for example, 80% of the examinees being classified as below basic is simply not reasonable or consistent with other available data about the examinees, and they may want, therefore, to change the standard for basic examinees. And, in so doing, the number of basic examinees would be increased, and the number of below basic examinees would be decreased.

There remains considerable debate about the merits of providing normative information, the timing of the presentation, and even the format in which it is presented. Many policymakers believe that panelists should set performance standards without knowledge of the consequences of applying those standards; it is the policymaking board’s prerogative to review the consequences and take appropriate actions. As for timing, one view is that if the normative data are provided too early, this may unduly influence the panelists because they have not had the chance to settle on their own views. However, if the data are provided too late in the process, panelists may be reluctant to consider it because there are often fairly confident in the process they went through and the performance standards that they have set.

Panelists often report that after working for 2 or 3 days through a process of reviewing test items and examinee work, and striving to minimize their own inconsistencies and differences with other panelists, they find it very hard to revise their performance standards when confronted with normative data. They are often willing to stick with the consequence of applying the performance standards to examinee data because they have confidence in the process and they feel that to change their performance standards would be "playing with the numbers." More research on the use of normative data in the standard-setting process is very much in order because many persons have strong views on both sides.

Revise, if necessary, and finalize the performance standards, and conduct a panelist evaluation of the process itself and their level of confidence in the resulting standards. Again, panelists are given the opportunity to revise their ratings to increase or decrease their performance standards. In addition, a panelist evaluation of the process should be conducted. One sample evaluation form for the booklet classification method appears in Appendix C. This is a modified version of an evaluation form from Hambleton, et al. (2000a), and it can be used as a basis for generating an evaluation form for other standard-setting initiatives. Information about the panelists' level of satisfaction with the performance descriptors, training, standard-setting process, and final standards is an important piece of the evidence for establishing the validity of the performance standards.

Compile validity evidence and technical documentation. It is important not only to be systematic and thoughtful in designing and carrying out a performance standard-setting study, but it is also necessary to compile validity evidence and document the work that was done and by whom (Kane, 1994). A description of the process from the specification of panelist characteristics to the selection of panelists, to training, implementation of the method, and final results, all need to be fully documented. In addition, the evaluative results from the panelists are important. Next, evidence of internal validity is needed including evidence that addresses the extent of intrapanelist and interpanelist consistency, the variability of performance standards across subpanels, and agreement between panelists' judgments and any relevant empirical evidence that might be available. Evidence that there is agreement between the performance categories and what they say examinees know and can do compared to the actual performance of examinees in these categories is also important. Finally, evidence of external validity, though often difficult to collect, can be very compelling information for the acceptance of the performance standards. Evidence of the generalizability of the performance standards across parallel panels, generalizability of the performance standards over parallel forms of the assessment, and agreement between the findings from the assessment and other testing evidence can be very useful for supporting the validity of the performance standards.

Technical documentation is valuable in defending the performance standards that have been set. Good examples of documentation are provided in the reports by Hambleton and Bourque (1991) and ACT, Inc. (1997) in setting performance standards on the NAEP and Mills et al. (2000) in setting a performance standard on the Uniform CPA (Certified Public Accountant) Exam. Often the group setting the performance standards is advisory to the agency or board that ultimately must set the standards. Technical documentation of the process is valuable information for the agency or board who must ultimately set the performance standards.

Criteria for Evaluating a Performance Standard-Setting Study

A number of researchers have offered guidelines or steps to follow in setting and/or reporting performance standards (Cizek, 1996a, 1996b; Hambleton & Powell, 1983; Livingston & Zieky, 1982; Plake, 1997). Following are a set of 20 questions that can guide the setting of performance standards via a judgmental process, or can be used to evaluate a standard-setting study:

(1) Was consideration given to the groups who should be represented on the standard-setting panel and the proportion of the panel that each group should represent? (There is no correct answer to this question. The important point in any study is that the question has been given serious attention in developing the standard-setting process.)

(2) Was the panel large enough and representative enough of the appropriate constituencies to be judged as suitable for setting performance standards on the educational assessment? (One of the most important points for defending a set of performance standards is to demonstrate that the panel is substantial in size and representative of the various stakeholder groups. Capability to make the required ratings is another important point to document. This last point is sometimes applicable when members of the public are being placed on standard-setting panels though it may be relevant with other groups as well.)

(3) Were two panels used to check the generalizability of the performance standards across panels? Were subpanels within a panel formed to check the consistency of performance standards over independent groups? (Setting standards with two panels can be cumbersome and prohibitively expensive for many agencies. At a minimum, however, a single panel can be divided into smaller, randomly equivalent subpanels who work independently of each other to arrive at the performance standards. These randomly equivalent subpanels provide a basis for estimating the standard error associated with each performance standard.)

(4) Were sufficient resources allocated to carry out the study properly? (Standard-setting studies can be costly [panelists' time, accommodations, and travel; staff time, accommodations, and travel; planning and revising the process; preparing training materials and other materials needed to implement the process; training facilitators; field-testing; data analysis; preparation of a final technical report; and so on.])

(5) Was the performance standard-setting method field tested in preparation for its use in the standard-setting study, and revised accordingly? (This is a very important addition to the process. In a recent standard-setting study by Plake and Hambleton [in press], three field tests were conducted of two new methods and each field test provided new and useful results that were incorporated into subsequent field tests. Field tests are especially important for new methods. We found, for example, that one application of a booklet classification method was flawed because panelists had difficulty reading the photocopies of examinee work and did not have sufficient workspace to do the job of booklet classification efficiently. Some problems can be anticipated, others will only be detected from a carefully conducted field test. Determining the times to complete particularly tasks in the process is one important purpose of a field test. Evaluating the training materials is another.)

(6) Was the standard-setting method appropriate for the particular educational assessment and was it described in detail? (Older methods such as the those developed by Angoff,

1971, Ebel, 1972, and Nedelsky, 1954, have been applied successfully to multiple-choice tests. The selection of a method is more difficult when the assessment consists of performance tasks only or a mixture of multiple-choice items and performance tasks. For example, the examinee booklet classification method seems problematic when the majority of the assessment is multiple-choice and there are only a few performance tasks. Instead of focusing on the quality of examinee work, panelists are counting the number of correct multiple-choice questions and using this number in classifying examinee test booklets. Research is needed to sort out the advantages and disadvantages of various new methods when applied to assessments where the proportion of performance materials varies from very low to 100%. As for details, terms like Angoff, Extended Angoff, booklet classification, and so on, have little meaning and need to be defined. Full details are needed about the method to enable others to evaluate the process. A good rule of thumb is to provide sufficient details so that someone else could replicate the study.)

(7) Were panelists explained the purposes of the educational assessment and the uses of the test scores at the beginning of the standard-setting meeting? Were panelists exposed to the assessment itself and how it was scored? (A briefing on the uses of the assessment scores and the assessment itself and scoring is fundamental for panelists to set appropriate performance standards. Very different standards may result depending on the purpose of the assessment. For example, were the purpose of an assessment principally diagnostic, panelists might be expected to set fairly high standards to maximize the number of examinees who might receive assistance. A very different set of performance standards would result if the same test were being used to award high school diplomas.)

(8) Were the qualifications and other relevant demographic data about the panelists collected? (This information is needed to fully inform reviewers about the suitability and composition of the panel setting the performance standards. All information pertinent for the evaluation of the panel or panels should be compiled. Even the panelists' motivation for participation may be relevant information.)

(9) Were panelists administered the educational assessment, or at least a portion of it? (Experience has shown that panelists benefit from taking at least part of the assessment under testlike conditions. They become aware of the pressure to perform well on a test, time limits, difficulties in using the test booklets, nuances in the test questions, and so on. All of this learning probably makes for more realistic performance standards.)

(10) Were panelists suitably trained on the method to set performance standards? For example, did the panelists complete a practice exercise? (One of the major changes in standard-setting practices in the last 10 years has been the commitment to fully train panelists in the method they are applying. The panelists' evaluation of the process is often helpful in documenting the extent to which the training was helpful. The presence of a formative evaluator may be useful also in cataloging strengths and weaknesses in the training process and the overall implementation of the study.)

(11) Were descriptions of the performance categories clear to the extent that they were used effectively by panelists in the standard-setting process? (This is another of the major changes in standard-setting practices over the last 10 years. Years ago, this activity may not have even been included in the process. Today the predominant view seems to be that arriving at consensus and clarity about the performance categories is an essential first step in developing meaningful performance standards. It is one of the ways that the variability of standards across panel members can be reduced. Reporting of results, too, is enhanced, if the performance category descriptions are clear.)

(12) If an iterative process was used for discussing and reconciling rating differences, was the feedback to panelists clear, understandable, and useful? Were the facilitators able to bring out appropriate discussion among the panelists without biasing the process? (The importance of these questions seems obvious to the validity of the overall process. Often, a postquestionnaire like the one in Appendix C provides the essential information. Low standard errors associated with the performance standards is another indicator of the effectiveness of the feedback. The role of the facilitator is often taken for granted, but the facilitator can have immense control over the final performance standards. The role of the facilitator in the standard-setting process deserves to be more thoroughly researched. Researchers such as van der Linden [1996] wrote about the need to demonstrate that the results, that is, the performance standards, are robust to minor changes in the process. The role of the facilitator is one of the factors that deserves more investigation.)

(13) Was the process itself conducted efficiently? Were the rating forms easy to use? Were documents such as examinee booklets, tasks, items, and so on, simply coded? If copies of examinee work were being used, were they easily readable? Were the facilitators qualified? (The process needs to flow smoothly from one activity to the next. Delays need to be minimized out of respect for the panelists' time and the desire to finish the process within the time allocated. Often a careful review of the process will turn up inefficiencies. For example, often examinee and booklet numbers contain more digits than are needed for a standard-setting study. Simplifying these codes can reduce errors among the panelists in recoding their data and save some valuable time. A good example of a problem reported by a number of researchers is the difficulty of producing copies of examinee test booklets for panelists. Often the examinee writing is light and in pencil and does not copy well. This creates problems for the panelists to read the examinee work.)

(14) Were panelists given the opportunity to "ground" their ratings with performance data, and how was the data used? (For example, were panelists given performance data of groups of candidates at the item level [for example, item difficulty values], or the full assessment level [for example, a score distribution]? The goal is for the data to be helpful, but not dictate the resulting standards. Often a very high correlation between panelists' ratings and candidate score information is taken as evidence that the empirical data are driving the standard-setting process.)

(15) Were panelists provided consequential data (or impact data) to use in their deliberations, and how did they use the information? Were the panelists instructed on how to use the information? (The intent of consequential data is to provide panelists with information that they can use to judge the reasonableness of the standards, and to make modifications in the performance standards, if they feel it is appropriate to do so.)

(16) Was the approach for arriving at final performance standards clearly described and appropriate? (The approach for arriving at performance standards from the data provided by panelists may involve some complex operations—see the procedures described by Cooper-Loomis and Bourque [in press] and Plake and Hambleton [in press]. Fitting statistical models, transforming panelist and examinee data to new scales, combining standards over sections of an assessment, and making adjustments for standard errors and/or measurement errors are all common steps in arriving at performance standards. Regardless of their complexity, they need to be clearly explained, and understandable to panelists who must ultimately decide on the acceptability of the performance standards. Ultimately, too, the approach used in arriving at the standards must be explained to boards and agencies.)

(17) Was an evaluation of the process carried out by the panelists? (This is another of the important ways to defend a set of performance standards. Did the panelists find the process credible? Did they have confidence in the training, the performance category descriptions, and the method? Again, Appendix C provides an example of an evaluation form that could be adapted for use in standard-setting studies.)

(18) Was evidence compiled to support the validity of the performance standards? (One of the main advances in recent years has been the attention in standard-setting studies to compile procedural, internal, and external validity evidence to support the validity of the resulting performance standards. These points have been developed in detail in Kane [in press].)

(19) Was the full standard-setting process documented (from the early discussions of the composition of the panel to the compilation of validity evidence to support the performance standards)? (All of the questions prior to this one need to be answered and presented in a technical report for reviewers to read. Attachments might include copies of the agenda, training materials, rating forms, evaluation forms, etc).

(20) Were effective steps taken to communicate the performance standards? (In some cases, the performance category descriptions may be sufficient for effective communication. Often, exemplar items that can describe the performance of candidates either in the performance categories or at the borderlines of performance categories are helpful [ACT, 1997]. This is a fairly new area of concern and research is presently being conducted by agencies such as the National Assessment Governing Board, who have the responsibility of communicating NAEP results to the public in meaningful ways.)

These 20 questions provide a framework for judging the quality of a standard-setting study. The same questions might be used in the planning stages of a standard-setting study to eliminate the possibility that important issues are skipped over. For more detailed criteria, readers are referred to the paper by Plake (1997).

Conclusions

Many researchers, policymakers, and educators are still not comfortable with several of the current performance standard-setting methods (see, for example, Pelligrino, Jones, & Mitchell, 1999; Shepard, Glaser, Linn, & Bohrnstedt, 1993). Criticisms center on both the logic of the methods and the ways in which several of the methods are being implemented. Clearly, there is a need for new ideas and more research. New methods, improved implementation of existing methods, and increased efforts to validate any performance standards are needed.

At the same time, performance standards are being set on many educational assessments with methods that appear to be defensible and valid (see, for example, Hambleton et al., 2000c). The steps described in the second section of this paper should be helpful to agencies planning or evaluating standard-setting studies. The steps are based on the best standard-setting practices found in the educational measurement field. The 20 questions described in the previous section might be asked at the planning stage of a standard-setting study, during the course of a standard-setting initiative, or at the end when conducting an evaluation of the full standard-setting process.

Perhaps the most controversial problem in educational assessment today concerns setting standards on the test score scale (or other score scale that is being used) to separate examinees into performance categories. It is now widely recognized by workers in the educational testing field that there are no true performance standards waiting to be discovered. Rather, setting performance standards is ultimately a judgmental process that is best done by an appropriately constituted panel who understand their tasks well, and are prepared to spend the necessary time to complete the work. In addition, full documentation of the process must be compiled, along with a validity study commensurate in size to the importance of the educational assessment. Following the steps described in this paper, implementing them well, and answering the 20 questions successfully will not keep an agency out of court. At the same time, these activities will increase considerably the likelihood of producing defensible and valid performance standards so that the educational assessments can achieve their intended goals.

Author's Note

The author wishes to acknowledge the helpful comments of Mary Lyn Bourque, Gregory Camilli, and Gregory Cizek on an earlier draft of the paper.

References

ACT, Inc. (1997). Setting achievement levels on the 1996 NAEP in science: Final report volume III achievement level-setting study. Iowa, City, IA: Author.

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

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 137-172.

Cizek, G. J. (1996a). Standard-setting guidelines. Educational Measurement: Issues and Practice, 15, 12-21.

Cizek, G. J. (1996b). Setting passing scores. Educational Measurement: Issues and Practice, 15, 20-31.

Clauser, B. E., & Clyman, S. G. (1994). A contrasting-groups approach to standard setting for performance assessments of clinical skills. Academic Medicine, 69(10), S42-S44.

Cooper-Loomis, S., & Bourque, M. L. (in press). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. Cizek (Ed.)., Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Ebel, R. L. (1972). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.

Hambleton, R. K., & Bourque, M. L. (1991). The levels of mathematics achievement: Initial performance standards for the 1990 NAEP Mathematics Assessment (Tech. Rep., Vol. 3). Washington, DC: National Assessment Governing Board.

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, J., & Zwick, R. (2000). A response to *Setting Reasonable and Useful Performance Standards* in the National Academy of Sciences' *Grading the Nation's Report Card*. Educational Measurement: Issues and Practice, 19, 5-13.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000a). Handbook for setting standards on performance assessments. Washington, DC: Council of Chief State School Officers.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (2000b). Setting performance standards on complex educational assessments. Applied Psychological Measurement, 24(4), 355-366.

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

Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard-setting. Evaluation & the Health Professions, 6, 3-24.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 485-514). New York: Macmillan.

Jaeger, R. M. (1991). Selection of judges for standard setting. Educational Measurement: Issues and Practice, 10, 3-6, 10.

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

Jaeger, R. M. , & Mills, C. N. (in press). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. Cizek (Ed.)., Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Kane, M. (1994). Validating the performance standards associated with passing scores. Review of Educational Research, 64, 425-462.

Kane, M. (in press). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.)., Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

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

Linn, R. L., & Herman, J. L. (1997). A policymaker's guide to standards-led assessment. Denver, CO: The Education Commission of the States.

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

Mills, C. N., Hambleton, R. K., Biskin, B., Kobrin, J., Evans, J., & Pfeffer, M. (2000). A comparison of the standard-setting methods for the Uniform CPA Examination. Jersey City, NJ: American Institute for Certified Public Accountants.

Mills, C. N., & Jaeger, R. J. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), Handbook for the development of performance standards (pp. 73-85). Washington, DC: US Department of Education and the Council of Chief State School Officers.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (in press). The bookmark procedure: Cognitive perspectives on standard setting. In G. Cizek (Ed.)., Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Nellhaus, J. (2000). States with NAEP-like performance standards. Washington, DC: National Assessment Governing Board.

Pelligrino, J. W., Jones, L. R., & Mitchell, K. J. (Eds.). (1999). Grading the nation′s report card. Washington, DC: National Academy Press.

Plake, B. S. (1997). Criteria for evaluating the quality of a judgmental standard setting procedure: What information should be reported? Unpublished manuscript.

Plake, B. S., & Hambleton, R. K. (2000). A standard setting method designed for complex performance assessments: Categorical assignments of student work. Educational Assessment, 6(3), 197-215.

Plake, B. S., & Hambleton, R. K. (in press). The analytic judgment method for setting standards on complex performance assessments. In G. Cizek (Ed.)., Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Plake, B. S., Hambleton, R. K., & Jaeger, R. M. (1997). A new standard-setting method for performance assessments: The dominant profile judgment method and some field-test results. Educational and Psychological Measurement, 57, 400-411.

Plake, B. S., Melican, G. J., & Mills, C. N. (1991). Factors influencing intrajudge consistency during standard-setting. Educational Measurement: Issues and Practice, 10, 15-16, 22-26.

Raymond, M. R., & Reid, J. B. (in press). Who made thee a judge?: Selecting and training participants for standard setting. In G. Cizek (Ed.)., Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Reckase, M. D. (2000). The evolution of the NAEP achievement level-setting process: A summary of the research and developmental efforts conducted by ACT. Iowa City, IA: ACT, Inc.

Reckase, M. D. (in press). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. Cizek (Ed.)., Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). Setting performance standards for achievement tests. Stanford, CA: National Academy of Education.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard-setting. Journal of Educational Measurement, 19(4), 295-308.

van der Linden, W. J. (1996). A conceptual analysis of standard-setting in large-scale assessments. In Proceedings of the Joint NCES-NAGB Conference on Standard-Setting for Large-Scale Assessment (pp. 97-118). Washington, DC: U.S. Government Printing Office.

Revised Version
Thursday, December 14, 2000

Appendix A

Pennsylvania Grade 8 Mathematics Assessment Descriptions of Novice, Apprentice, and Proficient Student Performance

Novice. Novice students demonstrate minimal understanding of rudimentary concepts and skills. They occasionally make obvious connections among ideas, providing minimal evidence or support for inferences and solutions. These students have difficulty applying basic knowledge and skills. Novice students communicate in an ineffective manner.

Apprentice. Apprentice students demonstrate partial understanding of basic concepts and skills. They make simple or basic connections among ideas, providing limited supporting evidence for inferences and solutions. These students apply concepts and skills to routine problem-solving situations. Apprentice students' communications are limited.

Proficient. Students performing at the proficient level demonstrate general understanding of concepts and skills. They can extend their understanding by making meaningful, multiple connections among important ideas or concepts, and provide supporting evidence for inferences and justification of solutions. These students apply concepts and skills to solve problems using appropriate strategies. Proficient students communicate effectively.

Appendix B

NAEP Grade 4 Reading Descriptions of Basic, Proficient, and Advanced Level Student Performance

Basic. Fourth-grade students performing at the basic level should demonstrate an understanding of the overall meaning of what they read. When reading text appropriate for fourth graders, they should be able to make relatively obvious connections between the text and their own experiences.

Proficient. Fourth-grade students performing at the proficient level should be able to demonstrate an overall understanding of the text, providing inferential as well as literal information. When reading text appropriate to fourth grade, they should be able to extend the ideas in the text by making inferences, drawing conclusions, and making connections to their own experiences. The connection between the text and what the student infers should be clear.

Advanced. Fourth-grade students performing at the advanced level should be able to generalize about topics in the reading selection and demonstrate an awareness of how authors compose and use literary devices. With reading text appropriate to fourth grade, students should be able to judge texts critically and, in general, give thorough answers that indicate careful thought.

Appendix C

An edited version of a sample panelist evaluation form from the <u>Handbook for Setting Standards on Performance Assessments</u> by Hambleton, Jaeger, Plake, and Mills (2000a)

Grade 8 Science Assessment
Standard-Setting Study

<u>Evaluation Form</u>

The purpose of this Evaluation Form is to obtain your opinions about the standard-setting study. Your opinions will provide a basis for evaluating the training and the standard-setting methods.

Please do not put your name on this Evaluation Form. We want your opinions to remain anonymous. Thank you for taking time to complete this Evaluation Form.

1. We would like your opinions concerning the level of success of various components of the standard-setting study. Place a A%@ in the column that reflects your opinion about the level of success of these various components of the standard-setting study:

Component	Not Successful	Partially Successful	Successful	Very Successful
a. Introduction to the Science Assessment				
b. Introduction to the Science Test Booklet and Scoring				
c. Review of the Four Performance Categories				
d. Initial Training Activities				
e. Practice Exercise				
f. Group Discussions				

2. In applying the Standard-Setting Method, it was necessary to use definitions of four levels of student performance: Below Basic, Basic, Proficient, Advanced.

   Please rate the definitions provided during the training for these performance levels in terms of adequacy for standard setting. Please CIRCLE one rating for each performance level.

			Adequacy of the Definition		
Performance Level	Totally Inadequate				Totally Adequate
Below Basic	1	2	3	4	5
Basic	1	2	3	4	5
Proficient	1	2	3	4	5
Advanced	1	2	3	4	5

3. How adequate was the training provided on the science test booklet and scoring to prepare you to classify the student test booklets? (Circle one)

   a. Totally Adequate
   b. Adequate
   c. Somewhat Adequate
   d. Totally Inadequate

4. How would you judge the amount of time spent on training on the science test booklet and scoring in preparing you to classify the student test booklets? (Circle one)

   a. About right
   b. Too little time
   c. Too much time

5. Indicate the importance of the following factors in your classifications of student performance.

	Factor	Not Important	Somewhat Important	Important	Very Important
a.	The descriptions of Below Basic, Basic, Proficient, Advanced	__________	__________	__________	__________
b.	Your perceptions of the difficulty of the Science Assessment material	__________	__________	__________	__________
c.	Your perceptions of the quality of the student responses	__________	__________	__________	__________
d.	Your own classroom experience	__________	__________	__________	__________
e.	Your initial classification of student performance on each booklet section	__________	__________	__________	__________
f.	Panel discussions	__________	__________	__________	__________
g.	The initial classifications of other panelists	__________	__________	__________	__________

6. How would you judge the time allotted to do the first classifications of the student performance on each booklet section? (Circle one)

a. About right
b. Too little time
c. Too much time

7. How would you judge the time allotted to discuss the first set of panelists' classifications? (Circle one)

   a. About right
   b. Too little time
   c. Too much time

8. What confidence do you have in the classification of students at the ADVANCED level? (Circle one)

   a. Very High
   b. High
   c. Medium
   d. Low

9. What confidence do you have in the classification of students at the PROFICIENT level? (Circle one)

   a. Very High
   b. High
   c. Medium
   d. Low

10. What confidence do you have in the classification of students at the BASIC level? (Circle one)

   a. Very High
   b. High
   c. Medium
   d. Low

11. What confidence do you have in the classification of students at the BELOW BASIC level? (Circle one)

   a. Very High
   b. High
   c. Medium
   d. Low

12. How confident are you that the Standard-Setting Method will produce a suitable set of standards for the performance levels: Basic, Proficient, Advanced? (Circle one)

a. Very Confident
b. Confident
c. Somewhat Confident
d. Not Confident at all

13. How would you judge the suitability of the facilities for our study? (Circle one)

a. Highly Suitable
b. Somewhat Suitable
c. Not Suitable at all

Please answer the following questions about your classification of student performance.

14. What strategy did you use to assign students to performance categories?

15. Were there any specific problems or exercises that were especially influential in your assignment of students to performance categories? If so, which ones?

16. How did you consider the multiple-choice questions in making your classification decisions about student performance?

17. Please provide us with your suggestions for ways to improve the standard-setting method and this workshop:

Thank you very much for completing the Evaluation Form.