Setting Performance Standards on Complex Educational Assessments

Ronald K. Hambleton, University of Massachusetts at Amherst
Richard M. Jaeger, University of North Carolina at Greensboro
Barbara S. Plake, University of Nebraska at Lincoln
Craig Mills, American Institute for Certified Public Accountants

Performance assessments have become popular in education and credentialing, and performance standards are common for interpreting and reporting scores. However, because of the unique characteristics of these assessments compared to multiple-choice tests (such as polytomous scoring), new and valid standard-setting methods are needed. Well-known standard-setting methods are no longer applicable. A number of promising methods for setting performance standards are described and their strengths and weaknesses are discussed. Suggestions for additional research are offered. *Index terms: achievement levels, passing scores, performance assessment, performance standards, standards.*

Since the early 1900s, human cognitive abilities have been assessed by pencil-and-paper (P&P) tests. These tests are used to measure achievement, aptitude, and knowledge that is assumed to be prerequisite to effective performance in academic and occupational settings. P&P testing is nearly universal in schools, when seeking employment, and when seeking professional licensure and certification.

The validity of P&P testing has recently been challenged. Performance assessments (PAs), which attempt to measure examinee skills and abilities to perform various tasks directly, are being used or are under development. Rather than presenting examinees with printed multiple-choice (MC) questions that require selection of a correct option from among those presented, PAs require examinees to construct responses to a wide range of problems. Although the trend toward PA use is increasing, much of the available methodology for assessing the psychometric quality of measurement instruments was developed for P&P tests. The applicability of these methods to PA measures is limited, questionable, or untested. For example, currently used methods for establishing performance standards on tests either require P&P MC test items (Nedelsky, 1954) or were intended for use with P&P dichotomously scored achievement test items (Angoff, 1971; Ebel, 1972).

Currently, over 40 states in the United States have adopted some form of PA for the assessment of students. Many credentialing agencies have also shifted to the use of PAs. Education and credentialing fields are now using assessments that consist of constructed-response and selected-response item formats (see Parshall, Davey, & Pashley, 2000).

Performance Standards

Whether the field is education or credentialing, performance standards for test score interpretations need to be established. In education, almost any type of criterion-referenced test score

Applied Psychological Measurement, Vol. 24 No. 4, December 2000, 355–366

interpretation requires performance standards to be set. Along with the educational reform movement in the U.S.A., it became common to set multiple performance standards for criterion-referenced test score interpretations. It is common to have three performance standards for classifying examinees; at times, however, two, four, and even five performance standards are set (see Nellhaus, 2000). The performance categories created by multiple performance standards are often designated "Failing" or "Below Basic," "Basic," "Proficient," and "Advanced." In the credentialing field, one performance standard is always required for sorting examinees into two performance categories, "certifiable" or "not certifiable." Sometimes, additional performance standards are set for diagnostic purposes.

PAs are often associated with complex and polytomous (more than two score points per task) scoring rubrics (i.e., criteria used for assigning scores to examinee responses to each task), multidimensionality in the response data (tasks requiring multiple skills for successful completion), interdependencies in the scoring rubrics (e.g., being unable to complete a task because one part of it was missed), and low score generalizability at the task or exercise level (performing well on one group of tasks does not mean a high performance on another). These features create special problems for standard-setting methods (SSMs). For example, several of the popular SSMs (see Berk, 1986; Jaeger, 1989; Livingston & Zieky, 1982; Plake, 1998) are not even applicable to PAs that are polytomously scored. The challenge is to adapt old SSMs or develop new methods to meet the current characteristics of PAs.

New Methods for Setting Performance Standards

Many new methods have been proposed for setting standards on complex PAs (e.g., Cizek, in press). Reckase (2000) described 10 different judgmental methods that have been studied by American College Testing (ACT) for setting performance standards on the National Assessment of Educational Progress (NAEP). Classification of SSMs is complex. Hambleton, Jaeger, Plake, & Mills (in press) offered six dimensions, two of which are identified below.

I. Focus of panelists' judgments
   A. Tasks (stimulus problems presented to examinees).
      1. Individual items or tasks,
      2. Blocks of items or sections of assessments, and
      3. An entire assessment or test.
   B. Examinees (classification of examinees based on evidence outside the test or assessment for which performance standards are to be set).
   C. Work products of examinees (classification of responses of examinees to stimulus material presented in the test or assessment).
      1. Examinees' responses to individual items or tasks,
      2. Examinees' responses to blocks of items or sections of assessments, and
      3. Examinees' responses to an entire assessment or test.
   D. Scored performances of examinees.
      1. Presented as a single performance score, and
      2. Presented as a profile of performance scores across tasks.
II. Judgment task presented to the panel
   A. Judgments associated with task materials from assessment or test.
      1. Estimation of conditional difficulty,
      2. Estimation of conditional mean score, and
      3. Classification of task material by category.

B. Judgments associated with examinees.
   1. Classification by category (e.g., Below Basic, Basic), and
   2. Selection of examinees on the boundaries of the performance categories.

C. Judgments associated with work products of examinees.
   1. Classification by category (e.g., Below Basic, Basic),
   2. Assignment by position within a performance category (e.g., on the lower boundary of a category, central to a category, on the upper boundary of a category), and
   3. Location in relation to a category boundary (using graphical stimulus materials).

D. Judgments associated with performance scores of examinees.
   1. Classification of profiles of performances across tasks.

(The other four dimensions are: the judgmental process itself, composition and size of the panel, validation of the resulting standards, and the nature of the assessment.) These two dimensions alone can create a wide array of approaches for setting standards. Reckase (2000) offered a different classification scheme that is also useful.

Methods for Setting Performance Standards on Complex Assessments

The methods for setting performance standards reviewed here are organized around four categories. These categories, based on Dimension I above, are: (1) items, tasks, and scoring rubrics (extended Angoff; estimated mean, expected score distribution; item mapping); (2) the examinees themselves (contrasting groups); (3) examinee work (examinee paper selection, holistic or booklet, analytical); and (4) score profiles (dominant profile, judgmental policy capturing, direct judgment). (For other emerging methods, see Cizek, in press; Reckase, 2000.)

Extended Angoff Method

Description. This method was developed for polytomously scored assessments (Hambleton & Plake, 1995; Loomis & Bourque, in press). Often, rubrics are used to guide the scoring of such questions. Panelists estimate the typical score that a borderline examinee will earn on a question. The average of the panelists' performance estimates for each question on a test are determined. The performance standards on the total score scale are then determined as an aggregation of these per-question averages. The averages can then be summed to set each performance standard for the assessment, or they can be weighted by estimates of an individual question's importance/value or by some other weighting function.

Pros and cons. One advantage of the extended Angoff method is its simplicity. The weighting process allows the panelists to differentially value the questions in the test, giving more weight to those questions they feel are more important. However, panelists sometimes feel that the method is too atomistic, breaking the assessment into small, isolated parts. Even though the weighting activity is aimed at providing a "broader view" of the entire assessment, panelists sometimes express the concern that the extended Angoff method might fail to take into account the holistic nature of the performance. Also, questions have been raised about the capability of panelists to make the required ratings.

Estimated Mean, Expected Score Distribution Method

Description. This method has some similarities to the extended Angoff method and was studied by ACT in setting performance standards on the NAEP (Reckase, 2000). Panelists are required to not only estimate the minimum number of score points for borderline examinees (as in the extended Angoff method), but also to estimate the distribution of scores of borderline examinees at the Basic, Proficient, and/or Advanced levels. The method was used by the National Assessment Governing

Board (NAGB) in its work to set performance standards on the NAEP (e.g., Loomis & Bourque, in press).

Pros and cons. One advantage of this method, in principle, is that additional relevant information about the performance of borderline examinees is extracted from panelists. Panelists who expect the standard deviation of the score distribution for borderline examinees to be low are indirectly also expressing considerable confidence in the placement of the performance standard. Higher standard deviations for the performance of borderline examinees likewise correspond to less confidence on the part of panelists about the proper location of the performance standard.

One disadvantage of the method is that it is complicated for panelists. ACT found that often panelists were not capable of distributing exactly 100% of the examinee score distribution across the possible score points for a task. Sometimes they would exceed 100% and other times they would be below 100%; consequently, the ratings would need to be reset and valuable panelist time would be wasted.

Item-Mapping (Bookmark) Method

Description. This new method presents panelists with a scale for reporting achievement; it highlights the performance of examinees on the assessment material at different regions on the reporting scale (Mitzel, Lewis, Patz, & Green, in press). Nellhaus (2000) reported that 18 states used some variation of this method to set performance standards. The item-mapping method is accomplished with item response functions (Hambleton, Swaminathan, & Rogers, 1991), assuming that examinees with more ability or proficiency will be able to perform well on more of the assessment material than examinees with lower ability. The important question for panelists is to decide the level of performance expected of Basic, Proficient, and Advanced examinees. This judgment is made easier by the ordering of assessment material by its level of difficulty.

Pros and cons. Panelists respond positively to this method—item or task level ratings are avoided, and the method can handle both selected and constructed item formats. Panelists decide how much knowledge and skills would be reflected by, e.g., Basic, Proficient, and Advanced examinees on the test or assessment. This is done by marking the assessment material where they think the borderlines should be placed to distinguish among the performance categories.

One of the unknowns in this method is the role of the approach to displaying the performance data on the reporting scale in the final determination of performance standards. For example, items might be identified on the reporting scale by the ability level at which an examinee has a 50% chance of success. In one variation, items might be identified on the reporting scale by the ability level at which an examinee has a 75% chance of success. There is a concern that this simple variation might considerably impact performance standards. More research on this aspect of this SSM is needed.

Other desirable research on this method would include: the impact of unclear descriptions at the beginning of panelists' work (these descriptions are typically prepared at the end of the process), the impact of mixing performance tasks and MC items (whether the performance tasks exert more weight on the placement of the performance standards than their contribution to overall test score), and the impact of the definition of what it means to know/do something on the resulting performance standards (currently this definition is often set at 67%, but it would seem that this choice influences the final performance standards).

Contrasting Groups Method

Description. This method involves judgments about individual examinees, traditionally on the basis of information external to the test for which performance standards are to be set (Cohen,

Kane, & Crooks, 1999). The performance standards are established by comparing the distributions of test scores for examinees who are classified in each performance category. In its original form, panelists with knowledge of the capabilities of individual examinees (e.g., teachers knowledgeable about the mathematical abilities of their students) would be asked to classify them with respect to the standards (e.g., doing work that is below grade level, doing grade-level work, doing work that exceeds grade level). Once examinees were classified without knowledge of their test scores, the score earned by each examinee classified in a particular category would be tabulated to form a frequency distribution. The resulting frequency distributions—one per category—would then be compared to determine their degree of overlap. The score points that optimally distinguish between adjacent categories would be selected as performance standards.

Pros and cons. One advantage of the contrasting group method is that it requires panelists to make decisions about individuals with whom they are familiar. This is a task that is likely to be familiar to teachers. The determination of misclassification rates is also possible, because the overlap of score distributions can be observed directly. However, teachers might be reluctant to place their own students into lower (failing) categories if they have given those students passing grades. Score distributions of students placed into different categories might overlap substantially, making determination of a performance standard either difficult or impossible. Perhaps the most serious concern has to do with the generalizability of the resulting performance standards to the population of examinees for whom the assessment is intended (Kane, 1994). The performance standards are very much group dependent; thus, if the sample of examinees used in the analysis is not representative, the resulting performance standards might not be suitable.

Examinee Paper Selection Method

Description. This method is typically used with assessments involving polytomously scored performance tasks. It is analogous to the borderline group method used with dichotomously scored data, except that panelists or judges consider examinee test responses rather than using their independent ratings of the examinees. Panelists select, from previously scored examinee work, the papers that they feel represent the work illustrative of borderline examinees for each question. The average of the scores on the selected papers for each performance standard is used as the "minimum passing value" for each question. These minimum values are summed to determine each performance standard for the test.

Pros and cons. Panelists (especially teachers) react positively to this method, because they have the opportunity to center their decisions based on actual examinee work. However, the method is difficult to implement with a small number of examinees; it might be difficult to get illustrative papers from the distribution of examinee work. Also, much effort is required for selecting examinee papers, identifying the coding system, organizing the packets of examinee papers for the panelists, and handling the large volume of papers to be evaluated by the panelists. Because there is no differential weighting of the component parts, each question is treated equally in the determination of each performance standard, even though some components might be less valued than others in the entire assessment. Although there is no reason why a weighting system could not be employed with the paper selection approach, this has not be done in practice.

Holistic (Booklet) Method

Description. This method is similar to the examinee paper selection method. This method was suggested by the National Academy of Education (1993) in its review of the standard-setting work of the NAGB and ACT. Recently, the holistic method has been studied extensively by Jaeger & Mills (in press) and Kingston, Kahl, Sweeney, & Bay (in press). Panelists are asked to consider

the complete work (including all exercises or tasks in the assessment) of an examinee and decide which examinee booklets represent those of borderline examinees (or Masters and Nonmasters; or Basic, Proficient, and Advanced examinees). In some cases, a rating scale is provided for panelists to sort examinee papers into categories, such as low, middle, and high Below Basic; below, middle, and high Basic (see Plake & Hambleton, in press-b). This approach also has been suggested as an alternative to the Angoff method with MC items, to counter the criticism that the Angoff method (focused at the item level) loses the overall impression of an examinee's performance. NAGB and ACT have been field-testing this method with NAEP data in several subject areas at Grades 4, 8, and 12 (see Reckase, 2000, for details).

Pros and cons. Jaeger & Mills (in press) have conducted several successful field tests of this method. One major advantage is that panelists use the actual work of examinees. Another advantage is that panelists consider the examinee work holistically. It allows panelists to be "more forgiving" than is possible with task-by-task rating methods. However, a question remains about the amount of examinee information panelists can handle. When booklets become overly long, can panelists make reliable and valid judgments about the quality of examinee work? An additional area of research concerns the method of choice for analyzing data to arrive at the performance standards.

Analytical Method

Description. This method was developed for use with assessments that include polytomously scored performance tasks. Plake & Hambleton (in press-a) were the first to use this method. The analytical method provides performance standards that are based on panelists' classifications of examinee performance evoked by the assessment. It is particularly useful when the assessment is large or comprised of multiple sections, because the panelists rate each major component of the assessment independently.

A collection of examinee responses representing the full range of performance on each section of the assessment is reviewed by panelists. Each performance on the first section is classified individually. Following the initial rating, panelists review their ratings as a group, identifying papers for which there are wide disparities in assigned classification and discussing the reasons for those disparities. Following completion of the second round of ratings for the first section of the test, panelists then rate performance samples on the second section.

In one variation, panelists first place the papers in one of the performance categories used to describe examinee achievement (e.g., Below Basic, Basic, Proficient, Advanced) and then designate the placement of the paper within the category (e.g., high, medium, low). In a second variation, the scale is designed to directly identify borderline papers in a single step (e.g., Below Basic, Borderline Basic, Basic, Borderline Proficient, Proficient). Performance standards are established for each section and summed to produce performance standards for the total assessment (see Plake & Hambleton, in press-b).

Pros and cons. The analytical method is easy to explain to panelists. It has the desirable characteristic of allowing panelists to base their judgments on actual examinee work. The calculation of performance standards, using the average scores of papers assigned to the boundary categories, is straightforward. One disadvantage, however, is the possibility that few papers will be assigned to boundary categories, resulting in less-stable performance standards based on a small number of papers. More sophisticated analyses, such as linear or nonlinear regression, do not suffer from this limitation but are less easily explained to the public. The method is also not well suited to MC tests or subtests. Preparation for the standard-setting study can be time-consuming, because test booklets must be separated into sections for copying. Also, if the discussion of ratings between rounds is extensive, the method can be quite time-consuming.

Dominant Profile Method

Description. This method can be used for an assessment that consists of a limited number of tasks or exercises, each of which can be scored polytomously. [For an example of an application in setting a single performance standard for a credentialing exam, see Plake, Hambleton, & Jaeger (1997).] Using a consensus-building strategy, panelists are asked to generate the policy (decision) rule that determines which scores across the tasks or exercises that comprise the assessment are the minimum needed to "pass." Compensatory and conjunctive components are feasible for this decision rule, so that the panelists can build as complex a decision rule as is meaningful to qualify the examinee as passing the assessment. This method, however, would be complicated to apply with more than a single performance standard.

After they have been trained on the tasks and exercises that comprise the assessment—including the meaning of the various possible score point values for each of the tasks or exercises—panelists are asked to independently create their initial decision rules for passing. These decision rules are analyzed qualitatively to identify those that represent similar components. These summary decision rules are then presented to the panelists for consideration, discussion, and modification. In some applications, additional rounds of individual decision rules are obtained, with a goal of building consensus in the panelists on a single decision rule. Often, examinee data are presented to inform panelists of the impact of the various decision rules. As the process progresses, it is expected that consensus (or at least a majority opinion) will emerge. This decision rule is voted on for panelists' endorsement and becomes the minimum examinee performance to qualify for pass status; all profiles that have score values that meet or exceed (dominate) this minimum profile are deemed as passing and all that do not meet this decision rule fail.

Pros and cons. This method allows panelists full flexibility in building as complex a decision rule as they deem necessary. At the conclusion of the process, if group consensus emerges, the panelists tend to feel confident that the decision rule is appropriate for making pass/fail decisions. On the other hand, the method does not necessarily lead to a total group consensus and minority positions might be strong. Under those circumstances, some panelists might object seriously to the final decision rule and will sometimes be unwilling to concede to the total group's opinion. This happens when minimum point values, such as not allowing any very low scores on the exercises, are entered into the decision rule. Some panelists might feel strongly that such low performance is inconsistent with certification and will resist relaxing this view. However, other panelists might feel that a "slip" on only one or some of the exercises is tolerable, especially if the overall total is kept at a sufficiently high level. Therefore, one disadvantage is the possibility of having the standard-setting process fail to result in a group consensus (perhaps resulting in a split in the vote between competing decision rules).

Another disadvantage of this method is the potential for conjunctive components to be part of (or even dominate) the decision rule. Because of limited measurement quality (especially reliability of scores) for these types of exercises, endorsement of conjunctive components puts serious measurement demands on the assessment components that might not be warranted by the observed level of measurement quality. This approach allows for (or encourages) conjunctive thinking in the composition of the decision rule for setting the minimum performance standard.

Judgmental Policy Capturing Method

Description. This method involves judgments about profiles of examinees' scores on the exercises in an assessment (see Jaeger, 1995, for an example application). Although it could be applied to any test, practical constraints make the method applicable solely to PAs that contain relatively

few (no more than approximately 10) exercises. Judgmental policy capturing is most useful when each exercise in an assessment is scored on a scale with at least three values, rather than being scored pass/fail. Because many performance exercises are scored this way, this constraint does not diminish the value of the method. The method is particularly useful when the assessment itself is multidimensional and a single performance standard is needed.

To apply judgmental policy capturing, standard-setting panelists must receive extensive training so that they thoroughly understand the meaning of each possible score that an examinee could earn on each exercise. When presented with an examinee's profile of scores on all exercises, panelists can then form a mental picture of the quality of the examinee's overall assessment performance. The score profiles of a large sample of examinees are classified by independently placing each into one of a number of categories with labels that are referenced to the performance standard (e.g., far below, somewhat below, barely below, barely above, somewhat above, far above).

When the panelists' judgments are analyzed, the categories are assigned numerical values. Statistical modeling procedures (e.g., multiple regression analysis) are used to determine a mathematical relationship between examinees' profiles of exercise scores and their assignments to performance categories. This relationship establishes the weights that will be applied to examinees' scores on each exercise and allows for computation of an overall performance score for each examinee.

Once an overall performance score has been computed for each examinee, exercise score profiles are again presented to the panelists, rank ordered from highest to lowest overall performance score. Panelists are asked to specify the lowest overall performance score necessary to pass the assessment. The recommendations provided by all panelists are tabulated, and the median of this distribution is taken to be the performance standard.

Both steps can be replicated after conducting a controlled discussion session in which panelists are given an opportunity to explain the rationale underlying their initial recommendations. Replication of the method following discussion usually results in greater agreement among panelists on the final performance standard and sometimes produces a small change in the performance standard.

Pros and cons. Judgmental policy capturing is one of the few SSMs designed to be used with PAs composed of exercises that are separately scored on a score scale with more than two points. It is a flexible method in that it acknowledges the possibility that some assessment exercises should be accorded greater importance than others when computing an examinee's overall performance score.

One of the important outcomes of judgmental policy capturing is a set of weights that defines the relative importance of the exercises. Although it is possible to use this method with models that approximate setting separate performance standards for each exercise in an assessment, the traditional multiple-regression analysis models most often used with the method are inherently compensatory in their structure. This means that an examinee can compensate for relatively low performance on some exercises by exhibiting relatively high performance on others. Compensatory performance standards have been shown to increase the measurement reliability of classification decisions, which is highly desirable (Hambleton & Slater, 1997).

The greatest disadvantage of judgmental policy capturing is the extensive time needed to train panelists who use it. Training panelists to use the method can be accomplished quite readily, but the need to ensure that panelists understand the meaning of each possible score an examinee could earn on each exercise in an assessment imposes a substantial training burden. If judgmental policy capturing is applied to a complex PA, days of training might be required. As a result, setting a performance standard can be quite costly.

Judgmental policy capturing models the implicit standard-setting policies held by panelists. The procedure models what panelists do in classifying examinees' performances on an assess-

ment, rather than asking panelists to explicitly state their policies. Some panelists find the method frustrating for this reason, because their actions do not mirror their expressed policy preferences. Advocates of the method hold that panelists' actions are a superior indicator of their policy preferences, because not everyone will understand their preferences, and few can do a good job of verbalizing them.

Direct Judgment Method

Description. As is true of judgmental policy capturing, this method involves judgments about profiles of examinees' scores on all of the exercises in an assessment (for a detailed description and example, see Hambleton et al., in press). Practical constraints make it applicable only to PAs that contain relatively few (e.g., no more than 10) exercises. It can be used with assessments containing exercises that are dichotomously scored (e.g., scored pass/fail or 0/1) or scored on a scale with three or more values.

To apply the direct judgment method, standard-setting panelists must receive thorough training so that they understand the meaning of each possible score that an examinee could earn on each exercise. When presented with an examinee's profile of scores on all exercises, panelists can then form a mental picture of the quality of the examinee's overall assessment performance.

The method is composed of two parts: (1) panelists determine the weight that will be assigned to each of the assessment's exercises when calculating an examinee's overall performance score on the assessment, and (2) the overall performance score and associated profile of exercise scores that are necessary to pass the assessment are determined.

Pros and cons. It is a flexible method that acknowledges the possibility that some assessment exercises should be accorded greater importance than others when computing an examinee's overall performance score. One of the important outcomes of this method is a set of weights that define the relative importance of the exercises. The direct judgment method produces an exercise weighting that is inherently compensatory.

Evaluations of the direct judgment method by panelists have produced positive results. Panelists easily learn and understand the method and, contrary to the reactions of some panelists who have used judgmental policy capturing, feel that they can readily adjust and control the weight assigned to each exercise. Although panelists sometimes have difficulty selecting a performance standard (there inevitably seems to be one or two troubling profiles of exercise scores), most panelists express very high confidence in and satisfaction with the performance standard produced by this method. As with the judgmental policy capturing method, the greatest disadvantage is the extensive time needed to train panelists who use it and the resulting cost of that training.

Directions for Future Research

Several of these SSMs are able to handle multiple performance standards. The paper selection, analytical, and entire booklet methods are illustrative of new SSMs that present panelists with examinee papers (other SSMs also do this as part of the judgmental SSM; e.g., the "body of work" method, Kingston et al., in press). The examinee papers serve as concrete illustrations of the range of examinee performance and form the basis for several new and emerging SSMs.

Approaches Based on Examinee Work

For SSMs that present panelists with a collection of examinee examination papers, there are a number of research questions that need to be considered. The number of examinee papers and the distribution of the scores of these papers need to be addressed in a systematic research program.

To date in most applications, a minimum of 50 examinee papers has been used, spanning the full score distribution. Research should consider whether other distributions of scores for the papers would produce higher precision at the performance standards. It seems logical that concentrating papers in the range of scores in which the performance standard(s) is most likely to fall would be a promising strategy. A variety of distribution shapes might have promise in addressing these issues: instead of a uniform distribution, some symmetrical distribution with predetermined central location might be appropriate. Research is also needed to investigate both the number of papers and their distribution to be used in SSMs based on examinee work.

For the analytical judgment method, studies should be conducted on optimum classification strategies. Although early research indicated administrative challenges with the task of sorting papers, panelists reported that they felt more confident when they used the sorting approach than did panelists who used a direct classification strategy (Plake & Hambleton, in press-b). The sorting method, in combination with a reduced total number of examinee papers and differing score distributions, should be the subject of a series of concentrated studies designed to clarify the circumstances under which the sorting (or direct classification) strategies produce the best results.

Another issue is whether examinee scores should be revealed to the panelists and when. It is not clear whether providing the panelists with the scores for the examinees' papers will serve as a biasing factor early in the judgmental process, or whether the presence of scores later in the process would help alleviate random error or unwanted biases. On the other hand, revealing examinee scores might present a bias regardless of when in the process they are shown to the panelists. In some cases, due to the complexity of the task presented to the examinees, panelists feel strongly disadvantaged without access to examinee scores. Having a research base for making informed decisions regarding examinee score availability would be a strong advantage in these situations.

General SSM Research

One area for future research is the role of feedback to panelists during the standard-setting process, including the type and timing of that feedback. Some have argued that providing panelists with examinee performance information is essential because it serves as a "reality check" (Cizek, in press). Instances in which an individual panelist's estimate of examinee performance deviates substantially from actual examinee performance might result in a need to reconsider expectations in light of current, relevant performance information.

On the other hand, some have argued that the presentation of examinee performance information introduces an inappropriate normative element into what is intended to be a criterion-based method (Impara & Plake, 1997). Further, some have argued that the timing of examinee performance data is important. Presenting examinee performance data too early in the decision-making process could be leading and biasing for the panelists. However, presenting this data late in the process might be ineffective (panelists might already be already fixed in their views and might be insensitive to examinee data) or frustrating (panelists might feel that this important information should have been shared earlier, before they spent valuable time pondering unrealistic expectations). In addition to when (and if) such examinee performance information should be shared with the panelists is the question of whether information about panelist judgments should be shared with the panel members between judgmental rounds.

When assessments are composed of both selected-response and constructed-response items or tasks, it is unclear whether it is best to use a generalized approach for both item types or whether different methods should be used and then the results combined to obtain the performance standards. Such decisions could be aided by research that focuses on the differences in final performance standards when a single or multiple approach is used.

Different SSMs generally produce different results; different panels, using the same SSM, will produce noncomparable results. Therefore, results from an SSM would be strengthened by running simultaneous panels using the same method and by employing a combination of standard-setting approaches. In some cases, it can be predicted in advance whether a method will likely produce under- or over-estimates of the performance standard. For example, because of the nonzero item intercorrelations on MC examinations, it is expected that the Angoff method will produce performance standards more extreme than intended by the panelists (Linn & Shepard, 1997). In addition, due to regression effects, it is likewise expected that the borderline group method will produce a performance standard closer to the distribution mean than the panelists intended. If a standard-setting study used multiple methods (e.g., Angoff and borderline group), the results could be considered boundary points in which a reasonable performance standard could be located. The use of multiple methods in standard setting is an area that should be studied.

Direction for validation efforts for performance standards would be very desirable. Performance standards derived from a standard-setting study must be subjected to validity investigations (Kane, 1994). However, the basis for this validity data is not always obvious or readily obtained. This is especially true in licensure and certification examinations, for which collateral examinee information is traditionally lacking or inaccessible. In school settings, however, examinee information can be more easily assessed and could be used in creative ways to provide validity evidence for the performance standards (Giraud, Impara, & Buckendahl, in press).

Recent innovations in SSMs have been developed because constructed response tasks have been used in examinations for making performance decisions (graduation, licensure, certification). With the development of these new methods has come the need for research studies to identify appropriate uses of the methods and ways to improve the precision of the results. In addition, research questions remain across many SSMs. The decisions that are made based on the performance standards derived from standard-setting studies have serious consequences for examinees. Therefore, it is critically important that research studies investigate ways to improve these performance standards.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). 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. (Ed.). (in press). *Standard setting: Concepts, methods, and perspectives.* Mahwah NJ: Erlbaum.

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education, 12,* 343–366.

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

Giraud, G., Impara, J. C., & Buckendahl, C. W. (in press). Making the cut in public schools: Alternative methods for standard setting. *Educational Assessment.*

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. N. (in press). *Handbook for setting performance standards.* Washington DC: Council of the Chief State School Officers.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8,* 41–55.

Hambleton, R. K., & Slater, S. (1997). Reliability of credentialing examinations and the impact of scoring models and standard-setting policies. *Applied Measurement in Education, 13,* 19–38.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park CA: Sage.

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

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education and Macmillan.

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. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives.* Hillsdale NJ: Erlbaum.

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

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

Linn, R. L., & Shepard, L. A. (1997, June). *Item-by-item standard setting: Misinterpretations of judges' intentions due to less than perfect item intercorrelations.* Paper presented at the Large Scale Assessment Conference, Colorado Springs CO.

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

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

Mitzel, H. D., Lewis, D. M., Patz, R. J., & Green, D. R. (in press). The bookmark procedure: Cognitive perspectives on standard-setting. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives.* Hillsdale NJ: Erlbaum.

National Academy of Education. (1993). *Setting performance standards for student achievement.* Stanford CA: Author.

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.

Parshall, C. G., Davey, T., & Pashley, P. J. (2000). Innovative item types for computerized testing. In W. J. van der Linden & C. Glas (Eds.), *Computerized adaptive testing: Theory and practice.* Boston: Kluwer.

Plake, B. S. (1998). Setting performance standards for professional licensure and certification. *Applied Measurement in Education, 11,* 65–80.

Plake, B. S., & Hambleton, R. K. (in press-a). A standard setting method designed for complex performance assessments: Categorical assignments of student work. *Educational Assessment.*

Plake, B. S., & Hambleton, R. K. (in press-b). The analytic judgment method for setting standards on complex performance assessments. In G. J. Cizek (Ed.), *Standard-setting: Concepts, methods, and perspectives.* Hillsdale NJ: Erlbaum.

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.

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

Author's Address

Send requests for reprints or further information to Ronald K. Hambleton, University of Massachusetts, Hills South, Room 152, Amherst MA 01003, U.S.A. Email: rkh@educ.umass.edu.