DOCUMENT RESUME

ED 411 304 TM 027 348

AUTHOR	Kane, M.; Crooks, T.; Cohen, A.
TITLE	Justifying the Passing Scores for Licensure and Certification Tests.
PUB DATE	1997-03-00
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Certification; *Cutting Scores; *Decision Making; *Licensing Examinations (Professions); Online Systems; *Pass Fail Grading; Standards; Test Interpretation; Test Use
IDENTIFIERS	*Angoff Methods; *Standard Setting

ABSTRACT

Passing scores for licensure and certification tests are justified by showing that decisions based on the passing score achieve the purposes of the credentialing program while avoiding any serious negative consequences. The passing score is justified to the extent decisions based on it have positive consequences. More specifically, the standard should be high enough to provide adequate protection for the public, and not so high as to restrict the supply of qualified practitioners unnecessarily or to exclude competent candidates from practicing. This paper begins by examining the intended outcomes of licensure and certification programs and by outlining the interpretive argument that is typically used for written credentialing examinations. Some criteria are developed for evaluating standard-setting methods in terms of how well they serve the goals of protecting the public, maintaining an adequate supply of practitioners, and protecting the rights of candidates. Finally, these criteria are applied to the Angoff procedure and to a generalized examinee-centered method. The Angoff procedure appears to suffer from inadequate controls on the judges' natural tendency to set high standards. The use of feedback on candidate performance would probably be a more effective control on this tendency to set high standards if it were given before or during the first round of ratings. (Contains 25 references.) (Author/SLD)

Justifying the Passing Scores for Licensure and Certification Tests

M. Kane
University of Wisconsin-Madison

T. Crooks
University of Otago, Dunedin, NZ

A. Cohen
University of Wisconsin-Madison

Paper presented at the Annual Meeting of AERA/NCME, Chicago, March, 1997

Correspondence:

Michael Kane
Department of Kinesiology
University of Wisconsin-Madison
2000 Observatory Drive
Madison, WI 53706
Telephone: (608) 265-2891
FAX: (608) 262-1656

ABSTRACT

Passing scores for licensure and certification tests are justified by showing that decisions based on the passing score achieve the purposes of the credentialing program while avoiding any serious negative consequences. The passing score is justified to the extent decisions based on it have positive consequences. More specifically, the standard should be high enough to provide adequate protection for the public, and not so high as to unnecessarily restrict the supply of qualified practitioners or to exclude competent candidates from practicing. This paper begins by examining the intended outcomes of licensure and certification programs and by outlining the interpretive argument that is typically used for written credentialing examinations. Some criteria are developed for evaluating standard-setting methods in terms of how well they serve the goals of protecting the public, maintaining an adequate supply of practitioners, and protecting the rights of candidates. Finally, these criteria are applied to the Angoff procedure and to a generalized examinee-centered method.

Setting standards on written tests is not necessarily difficult; we have a wide range of standard-setting options available. What is challenging is to choose an appropriate standard-setting method and to develop confidence that the resulting passing score is at least approximately, at the "right" level.

Such assurance is developed by justifying, or validating, the overall decision process in which the passing standard plays a central role. As in all cases of validation, the validation of a passing score does not constitute a proof that the passing score is true or correct; rather validation supports the plausibility and appropriateness of the passing score (Kane, 1994), or the credibility (Norcini & Shea, 1997) and defensibility (Norcini, 1994) of the passing score. Passing scores on credentialing (i.e., licensure and certification) tests are justified by showing that the decision process in which the passing score is used achieves the purposes of the credentialing program while avoiding any serious negative consequences. The standard should be high enough to provide adequate protection for the public, and not so high as to unnecessarily restrict the supply of qualified practitioners or to exclude competent candidates from practicing.

If we could evaluate performance in professional practice, or on some fairly thorough and demonstrably valid performance test or simulation, we might choose to use a criterion-related method to set the passing score on the written test. Assuming further that we could decide whether each candidate passes or fails on the criterion measure (i.e., that we have established performance standards for the criterion), and that we know the candidates' scores on the test, we could decide on an optimal passing score on the test. In fact, we would be using a version of what has been called the contrasting-groups method.

But a good external criterion with established performance standards is not usually available. In this paper, therefore, we will focus our attention on what we can do when we have nothing but the test and a record of candidate performances on the test for some cohort of candidates.

We begin by examining the intended outcomes and some possible unintended outcomes of credentialing programs. We then outline the interpretive argument implicit in getting from test scores to credentialing decisions and evaluate some of the strengths and weaknesses of the argument in light of intended and unintended consequences. The fact that written tests are indirect indicators of professional competence makes the interpretive arguments for written certification tests somewhat more complicated than they might otherwise be. The lack of direct measures also makes it harder to evaluate the intended and unintended consequences of the decision process based on the passing score.

THE GOALS OF LICENSURE AND CERTIFICATION EXAMINATIONS

The two main kinds of professional credentialing programs, licensure and certification, are designed to make decisions about whether individuals are or are not prepared to perform effectively at some level of practice. Licensure focuses on general, entry-level practice and on safety; it is designed to exclude individuals who would be ineffective or unsafe in entry-level practice. Certification focuses on levels of practice that are beyond entry-level, and seeks to identify candidates who have demonstrated readiness for advanced practice in some area of certification.

Licensure and certification examinations generally require examinees to respond to some set of tasks, often written, assessing the knowledge, skills, and judgment (KSJs) needed in practice. Performance on these tasks provides a basis for evaluating the examinee's readiness for practice at the appropriate level. Performance on the written test (typically multiple-choice) is used to make decisions about the candidates' readiness for practice. Therefore, the assessment tasks need to elicit a sample of performance that provides a good indication of how well the candidate is likely to perform in the kinds of encounters that occur in practice.

A conclusion about readiness for practice based on performance on a written test, relies on an interpretive argument involving a string of inferences (Kane, 1992). First, the evaluation of performance on the test tasks involves an evaluation of the candidate’s performance, leading to an observed score. Second, the observed score is generalized to the universe score for the universe of generalization from which the observed performances are drawn. Third, universe scores are interpreted in terms of level of achievement on the KSJs being assessed by the test. Fourth, an inference is drawn from level of achievement on the KSJs to expected performance in practice.

The negative side of the argument, which applies to failing candidates, is usually easier to justify than the positive side, which applies to passing candidates. The negative side of the argument claims that examinees with scores below the passing score are sufficiently lacking in the KSJs that they should not be licensed: failing scores are associated with poor performance on the assessment tasks, poor performance on the assessment implies a low estimate for the universe score, a low universe score indicates inadequate mastery of the critical KSJs, and inadequate mastery of the KSJs is associated with poor performance in practice.

On the positive side, a passing score indicates adequate performance on the assessment, which indicates an adequate universe score, which indicates adequate mastery of the KSJs required for practice, which in turn indicates readiness for practice. The positive side of the argument is weaker than the negative side because the KSJs assessed by the examination generally constitute a subset (arguably a small subset) of the characteristics required for effective performance in practice. Some candidates who have mastered the KSJs included in the examination may fail to perform adequately in practice because they lack some other characteristics (e.g., motivation, diligence, ethics or some KSJs not measured on the test) that are necessary for effective practice.

The examination scores can be interpreted as measures of KSJs that are <u>critical</u> for practice, in the sense that they are necessary, although perhaps not sufficient, for effective performance in practice. Mastery of the KSJs does not guarantee successful performance in practice, but major gaps in mastery would be a serious limitation in the practice of the profession. Therefore, examinees with low scores are probably not prepared to manage at least some of the kinds of encounters included in the performance domain and are thus considered unprepared for practice.

The interpretive arguments for licensure and certification tests involve at least three major assumptions. First, certain knowledge, skills and judgement are necessary for practice. Second, the test measures an important subset of these KSJs, and therefore, test scores are indications of readiness for practice. Third, the passing score is such that those who pass have a high enough level of mastery of the KSJs to be ready for practice and those who fail are lacking essential KSJs.

The interpretive argument assumes that we know, at least in general terms, the kinds of knowledge, skills, and judgement that are needed in practice and that we know how to assess whether students have mastered these KSJs. Models of practice, based on various kinds of empirical data, provide us with a general definition of the content of practice in terms of the kinds of professional encounters occurring in practice and the kinds of interventions being used in these encounters (LaDuca, 1994). The appropriate way to handle each of these encounters is indicated by the results of research on the effectiveness of various approaches to the delivery of professional services (e.g., how the efforts of different professionals from different professions can be effectively combined), on the kinds of interventions that are effective in various situations, and on how to implement various interventions (LaDuca, 1994; Tamblyn, 1994). Given a working model of practice, expert judgement can be combined with empirical data to identify KSJs that are critical for practice, in the sense that the effectiveness of performance in practice depends on the level of mastery of these KSJs.

A particularly important decision to be made in developing the assessment procedure is the choice of a passing score. Given that the practice domain and the critical KSJs have been identified, and that tasks have been developed to test for mastery of the KSJs, it is still necessary to develop appropriate performance standards and to identify the corresponding passing score. It is necessary to identify a particular level of mastery of the KSJs as reflected in assessment scores, below which a candidate is considered unprepared for practice (see Norcini, 1994). A failing score is interpreted as indicating that the examinee has serious deficiencies in their command of the critical KSJs. A passing score is interpreted as indicating that the examinee has adequate mastery of the KSJs to practice safely and effectively.

STANDARD-SETTING IN LICENSURE AND CERTIFICATION

Standard-setting adds an explicit decision rule to the basic interpretation of the score scale. Candidates with scores above the passing score are considered ready for practice (entry-level for licensure, and a more advanced level for certification) and candidates with scores below the passing score are considered unprepared.

The required level of achievement specified in terms of what candidates need to know and be able to do defines a performance standard for the KSJs. The performance standard is a qualitative description of the level of achievement on the KSJs needed for practice at a particular level. The passing score is the point on the score scale for the test that is used to make the pass/fail decisions.

The performance standard and the corresponding passing score are necessarily judgmental in the sense that they impose a specific cutpoint on a continuous range of achievement. Candidates with scores just above the passing score will not be that much better prepared than those just below the passing score. Yet those with scores at or above the passing score pass and those with scores just below the passing score fail. The continuous (at least approximately) variable describing test performance has been converted into a binary, pass-fail variable.

Furthermore, there is no "correct" performance standard or passing score to be discovered. The problem of standard-setting is not a problem of estimation (although we can turn it into one). Rather, the definition of the performance standard is a policy decision. Somebody (i.e., a state board for licensure or the certifying organization) gets to decide how high the performance standard should be, and therefore how high the passing score is to be. The outcome of this decision will depend on many factors including the nature and degree of risk posed by marginal candidates, current levels of training of candidates, prevailing standards of practice, the demand for practitioners in the market, as well as overt political factors. So there is no uniquely right answer; we seek a performance standard that is reasonable and appropriate, or as Norcini and Shea (1997) put it, credible.

In some cases in the past (and less frequently today) the passing score was simply set at some predefined score (e.g., 70% of items answered correctly). This passing score has some comfort associated with it by virtue of familiarity (from grade school?), but has very little else to recommend it. A second approach, which has also declined in popularity, is to set the passing score relative to performance in some population. Such norm-referenced approaches set the passing score at a point such that some percentage of the candidates in the reference population would pass.

The standard-setting methods that are currently most popular conceptually and also most commonly used in licensure and certification exams are criterion-referenced methods that emphasize judgements about the level of performance required for some purpose (e.g., readiness for entry-level practice). Expert judgements about how good the performance has to be in order to be good enough comprise a central component of these criterion-referenced methods (Jaeger, 1989; Norcini & Shea, 1997).

Criterion-Referenced Standard-setting Methods

All standard setting is based on judgement of some kind. Jaeger (1989) has drawn a fundamental distinction between two types of criterion-referenced standard-setting methods, test-centered methods and examinee-centered methods. The test-centered methods are based on judgements about the tasks included in the test. The examinee-centered methods are based on judgements about the performances of individual examinees.

The most commonly used methods for licensure and certification examinations are the test-centered methods, with the Angoff (1971) procedure being the most popular of the test-centered methods (Sireci & Biskin, 1992; Kane, 1994; Plake, 1996), but examinee-centered methods, particularly the contrasting groups method, are also used to some extent (Plake, 1996, Clauser & Clyman, 1994).

In the test-centered methods, judges review the items or tasks in the test and decide on the level of performance on each item or task that will be considered just adequate to meet the performance standard. The judgements about the items in the test are then aggregated in some way to arrive at an overall passing score. For example, in the Angoff (1971) procedure, the judges are asked to envision a borderline examinee and to choose a minimum pass level, or MPL, representing the expected performance of the borderline examinees on each item. In some cases, the judges are asked to imagine a group of 100 minimally competent candidates and estimate the proportion who would answer the item correctly. The MPLs are averaged over judges to get the item MPL, and the item MPLs are summed over the items in the test to get a passing score. The procedures proposed by Angoff (1971), as well as those proposed by Nedelsky (1954), Ebel (1972), and Jaeger (1982) are test-centered in the sense that they require the judges to rate test items or tasks, rather than examinee performances.

In the examinee-centered methods, judges categorize examinees as passing or failing, based on some external criterion or on overall test performance. The passing score is set by identifying a point on the score scale for the test that would be most consistent with these decisions. In the borderline-group method (Livingston & Zieky, 1982), judges identify individuals who are borderline, in the sense that their level of achievement is right around the performance standard. The median score on the test (or some other index of central tendency) for this group of borderline examinees can then be used as the passing score. In the contrasting-groups method (Livingston & Zieky, 1982), the judges categorize a sample of examinees into two groups, those judged to have met the requirements in the performance standard and those judged to have not met this standard. The passing score is chosen so that it discriminates as well as possible between the upper contrasting group and the lower contrasting group.

It seems that neither the examinee-centered nor the test-centered methods are best in all contexts. Each type of standard-setting method has strengths and weaknesses that make it appropriate in some situations and inappropriate in others (Kane, 1996). The test-centered methods seem to be particularly appropriate for use with tests consisting of multiple, independently-scored tasks, designed to assess a wide range of distinct KSJs. Examinee responses to individual tasks are scored separately, and these item scores are combined to yield the examinee's overall score. The test-centered methods parallel this test design by assigning a minimum pass level to each task and then combining these item MPLs into an overall threshold score.

The examinee-centered methods seem particularly appropriate for performance tests and simulations in which examinees are asked to complete a few, relatively long tasks. The examinee-centered methods provide an holistic approach to standard setting that meshes well with assessments that emphasize complex performances, and judges tend to have extensive experience in making this kind of judgement. Judges may find it difficult to apply the test-centered methods to performance tests and simulations, but find it relatively easy to apply the examinee-centered methods to such tests.

In part because most credentialing examinations are multiple-choice tests, the test-centered methods (especially the Angoff) are frequently used. However, as more complex approaches to testing are introduced, the examinee-centered methods are likely to gain in popularity.

The Goldilocks Criteria: Using Consequences to Evaluate Standard-Setting Methods

In this section, we discuss two general criteria for evaluating passing scores and the standard-setting procedures used to generate them. These two criteria will be called the "Goldilocks Criteria" for reasons that should become obvious. The first criterion is that the standard and, therefore, the corresponding passing score, should not be set too low. The standard should be high enough to provide adequate protection to the public. The second criterion is that the performance standard and passing score should not be set too high. The standard should not constitute an unreasonable barrier to practice (or, in the case of certification, to practice at an advanced level). The ideal performance standard is one that provides the public with substantial protection from incompetent practitioners and simultaneously is fair to the candidate and does not unduly restrict the supply of practitioners. We want the passing score to be neither too high nor too low, but at least approximately, just right.

The adoption of performance standards and passing scores that are neither too high nor too low is certainly an appealing goal, but how do we achieve it? And how do we know if we have achieved it? How do we know if the performance standard and/or passing score is too high or too low?

Unfortunately, we do not have any method for determining what the standard should be, and therefore, we do not have any obvious way of deciding if the standard is too high or too low. We might be able to agree, however, on criteria for identifying cases in which the standard is clearly too high or clearly too low (Kane, 1994). For example, if within a population of apparently well prepared candidates for practice (i.e., those meeting rigorous educational and experience requirements), we get a very high failure rate, we might be inclined to think that the examination is too difficult or the passing score is too high. Similarly, if there are no failures among a population of candidates with very uneven educational backgrounds and levels of experience, we might be inclined to suspect that the passing score is too low. We are particularly likely to suspect that the passing score is too low if we get frequent indications that recently licensed or certified individuals lack the KSJs covered by the examination. Although such reality checks may be useful in identifying cases for which the standard is clearly out of bounds, they are not very useful in identifying cases where the passing score is a little too high or a little too low and are therefore not very useful in fine-tuning the process.

Given this limitation, how should we design the standard-setting procedure to meet the Goldilocks criteria, and thereby, to give us reasonable assurance that the passing score is in an appropriate part of the score scale. We do not expect to get it exactly right; it is not even clear what it would mean for the standard (a policy decision) to be "exactly right" (since there is no true value). Mainly, we want it not to be drastically wrong in the sense of having unacceptable consequences.

A reasonable way to achieve this goal would be to design the standard setting procedure so that it has characteristics or elements that tend to keep the passing score from being too high and also characteristics or elements that keep it from being too low. The Goldilocks criteria do not give us a sure way to achieve appropriate standards, but they do point to certain questions that we might want to address in developing and/or evaluating a standard setting procedure:

(1) What characteristics or elements of a proposed procedure are likely to keep the standard from getting too high?

(2) What characteristics or elements of a proposed procedure are likely to keep the standard from being too low?

(3) How realistic is it to expect that the combined effect of all of these characteristics and elements will result in a reasonable standard?

(4) In particular, how confident can we be that the combined effect of these characteristics and elements will not result in a standard that is much too high or much too low?

There is a good dose of the traditional mini-max strategy here. We want to minimize our maximum loss. We want to avoid the probability of getting a standard that is substantially off. But we also want to design the system so that it has a good chance of giving us a passing score that is appropriate. In the next two sections, we will examine how the consequence criteria might be applied to the design of a test-centered method (the Angoff) and to an examinee-centered method.

The approach taken here is largely analytic. We use the Goldilocks criteria as general guidelines for evaluating the plausibility or credibility of standard-setting procedures, but do not intend to imply that empirical checks on the results of these procedures are unimportant. Kane (1994) and Norcini and Shea (1997) have catalogued a number of empirical checks on the validity or credibility of standards, and where feasible, these methods should be employed. The analytic use of the Goldilocks criteria provides a complementary perspective on evaluation of the procedures, as well as some guidelines for the design of standard-setting procedures.

A Test-centered Method - The Angoff

As noted earlier, the Angoff procedure has judges envision a minimally competent candidate and estimate the probability (the MPL) that this just-competent examinee would answer each item correctly. The MPLs are averaged over judges and summed over items to get the passing score.

The Angoff raters are intended to have both a thorough knowledge of the content being assessed and to have current experience with practice at the appropriate level. They are usually chosen to be experts on the KSJs and to have enough experience with recently credentialed practitioners to know what can be expected of them.

Current implementations of the Angoff procedure typically involve two or more rounds of ratings, with the judges getting information about the internal consistency of their estimates, about their consistency with other judges, and/or about the empirical p values of the items after each round (Berk, 1986). The judges are asked to reconsider their estimated MPL values after receiving the new information presented in each round.

On the face of it, the Angoff method seems to provide a reasonable and defensible method for setting passing scores on multiple-choice tests used for licensure and certification (Norcini, 1994). It is clearly the most widely used method (Sireci & Biskin, 1992; Kane, 1994; Plake, 1996). Note, however, that Shepard, Linn, Glaser, Bohrnstedt (1993) present some serious criticisms of the Angoff procedure.

Applying the Goldilocks Criteria to the Angoff Method

The Angoff procedure has several characteristics that are likely to keep the passing score from being too low. The task put to the raters (i.e., to determine the probability that a minimally competent candidate would answer the item right) tends to encourage high standards. In evaluating the KSJs, one could reasonably ask why we should accept anything less than perfect performance. All of the questions on the test are relevant to good practice and should be important in some practice situations. Would not the public be safer and better served if passing candidates knew the answers to all such questions, rather than just to 60% or 80% or even 95% of them? Furthermore, the judges who are chosen, in part, for their expertise (Norcini, Shea, and Kanya, 1988) are likely to have a particularly high level of achievement in the KSJs. Such judges are likely to see most of the questions as being quite easy, and so are likely to expect even minimally qualified candidates to have a high probability of answering the questions correctly.

The component of the process that is most directly related to keeping the passing score from being too high is the feedback provided to judges on actual candidate performance, usually in the form of item p values. The judges are likely to realize that even highly qualified candidates are not likely to get 100% on any credentialing examination, and that setting the passing score at 100% will result in everyone failing, but they are not likely to have any precise expectations about candidate performance. Therefore, to aid the judges in moderating their expectations in the light of reality, it is now common practice to provide them with information about the performance of the pool of candidates on each item (Berk, 1986; Norcini, Shea, & Kanya, 1988; Jaeger, 1989; Plake, 1996; Norcini & Shea, 1997).

The data may be provided at once after the first round or in several installments, one after each of several rounds. The introduction of data on examinee performance provides a reality check that is designed to keep the passing score within reasonable bounds (Norcini, 1994; Berk, 1986; Shepard, 1984; Jaeger, 1989). However, in practice, the main concern is that the passing score will be set too high if some check is not put upon the tendency of item-by-item judgements to yield high passing scores. Plake (1996) mentions a similar concern about task-by-task, Angoff-like judgements applied to extended response questions.

This reality check may not be very effective, however, because the effectiveness of feedback on candidate performance provided after the first round is questionable. A number of studies have found that, when information on examinee performance is provided after an initial round of ratings (Busch & Jaeger, 1990; Harker & Cope, 1988: McLaughlin, 1993), the performance data tend to have relatively little impact on the passing score. The variability in the ratings tends to go down and their correlations with the empirical p values tend to go up, but the overall passing score does not change appreciably. It seems that the judges use the feedback mainly to align their ratings with the p values and thereby to make them more consistent, but they do not seem to change the location of the assigned passing score to any appreciable extent.

To the extent that feedback on item p values is to have much impact on the overall passing score, it probably needs to be provided before or during the first round of ratings (Norcini & Shea, 1992). Once the judges have rated all of the items and possibly discussed their ratings during the first round of ratings, they have probably made up their minds to a large extent. Since credentialing examinations can include up to several hundred items, the first round of ratings typically involves a lot of work and takes a long time. It is probably the case that many judges have formed strong impressions about what they expect of minimally competent candidates on the items by the end of the first round, and it would be quite difficult to modify these views substantially after that point. Therefore, if feedback on p values is to be effective in keeping the MPLs from getting too high, it should probably be given early in the process, preferably during initial training.

Potential Impact of Very Hard and Easy Items

In designing a procedure to achieve the kind of balance envisioned in the conseqences criteria, it is important to avoid any feature that might throw the result off, in the sense of yielding an unacceptably low or high passing score. For example, some recent research indicates that the Angoff procedure may be vulnerable to the presence of items with extreme p values (Mclaughlin, 1993; Shepard et al., 1993)

In a study of standard setting results on NAEP, McLaughlin (1993, p. 97), found that the passing scores seemed to be biased by a tendency of the judges to avoid extreme MPLs. As Shepard et al. (1993, p. 58) put it, "While judges estimated higher p-values for easy items than for hard items (as evidenced by correlations between judges' ratings and real-data p-values), judges failed to adjust sufficiently for differences in item difficulty."

Adjustments for task difficulty are an integral, although perhaps implicit, part of any viable standard setting procedure (Kane & Wilson, 1984). Jaeger (1989) and Kane (1994) suggest that the Angoff MPLs tend to have higher correlations with p values than the other test-centered methods. Nevertheless, the results reported by McLaughlin (1993) and Shepard et. al. (1993) suggest that Angoff judges may avoid very high or low MPLs.

A general tendency of the judges to avoid extreme MPLs would be a potential source of bias and instability in the resulting passing scores. If a test happened to include a substantial number of very easy items (and no balancing set of very difficult items), the passing score would tend to be set lower relative to examinee performance than it would otherwise. Similarly, if a test happened to include a preponderance of very difficult items, the passing score would tend to be set higher relative to examinee performance than it would otherwise.

In order to eliminate this potential source of error, it would be desirable to avoid applying the Angoff method to items that are either very hard or very easy for the population of candidates being assessed.

The Angoff procedure generally involves the review of a large and representative sample of the items in the test. It is often the case that all of the items are reviewed, but this is probably overkill (Norcini, Shea & Ping, 1988). Given the large number of items in most multiple-choice tests and the relatively modest numbers of raters used in most standard-setting studies, the errors due to the sampling of raters may be substantially larger than the errors that would result from sampling items. So, instead of having 10 raters assign MPLs to the 300 items in a licensure or certification test, it might be much better to have three groups of 10 raters, with each group rating 50 or 100 items. In any case, decreasing the number of items used for the Angoff ratings is not likely to cause any serious problems. Therefore, eliminating items with extreme p values from the Angoff rating process should not interfere with effective use of the Angoff procedure, and may alleviate one potential source of bias.

Our application of the Goldilocks criteria to the Angoff procedure suggests that the elements tending to keep the standard from being too low are probably effective, but that the elements designed to keep the standard from being too high may be inadequate. To achieve better balance, we suggest that data on candidate performance on items be provided early in the process, either before or during the first round. As an additional precaution, we suggest that items with extreme p values be excluded from the process.

A Generalized Examinee-Centered Method

As indicated earlier, the examinee-centered methods offer several advantages for standard-setting, especially in those cases in which the assessment involves a relatively small number of independent tasks, each of which involves an extended response or a series of responses (Livingston & Zieky, 1982). However, examinee-centered methods tend to be expensive, in terms of time and resources to implement, because it takes a lot of time to review extended responses (Norcini, 1994). It is, therefore, desirable to use the data derived from these studies as efficiently as possible.

The two most commonly used examinee-centered methods are relatively inefficient. For the borderline-group method, raters identify performances that just meet the performance standard. Given that the judges may need to review a large number of performances in order to identify a relatively small number of borderline performances, and given that each performance may require an extended period for review, we have a large expenditure of time for a relatively modest return in terms of data.

The contrasting-groups method is much better in that the raters are looking for performances in two relatively broad categories, rather than performances right around the performance standard. Even so, we are still likely to have performances reviewed that do not contribute data, because they do not clearly fall into either of the contrasting groups. And, the simple binary classification does not extract much information from each performance reviewed.

In this section, we propose a generalized examinee-centered method that could increase the efficiency of examinee-centered standard setting. This method enjoys the main advantages of the two examinee-centered methods. In particular, the judges review actual examinee performances, which helps to ground the process in reality. Further, judges are asked to engage in a relatively familiar activity, rating performances. In addition, the generalized examinee-centered method incorporates data analysis procedures (i.e., nonlinear regression) that are potentially more efficient than those employed in the borderline-group method or the contrasting-groups method.

The generalized examinee-centered method requires judges to rate each paper using a scale like the following:

8	
7	Clear Pass: performance clearly exceeds the performance standard
6	
5	Just Passing: performance just meets the performance standard
4	
3	Clear Fail: performance clearly fails to meet the performance standard
2	

The scale is anchored by the performance standard. If the examinee's performance seems to be just consistent with the performance standard, neither clearly above nor clearly below the standard, the paper would be given a rating of 5. If the data collected using this procedure were to be used in a borderline-group study, the performances with a rating of 5 could be viewed as the borderline group. So, in a sense, the procedure proposed here can be viewed as including the borderline-group method.

If the judges think that the test performance is clearly better than that specified in the performance standard, the judges would assign the performance a rating of 7. Similarly, if the judges think that the performance clearly does not satisfy the performance standard, in the sense that the performance provides no substantial indication of achievement of the performance standard and/or evidence clearly indicating a failure to meet performance standards, the judge would assign it a rating of 3. If the data collected using this procedure were to be used in a contrasting-groups study, the performances with a rating of 7 could be viewed as the upper contrasting group and the performances with a rating of 3 could be viewed as the lower contrasting group. So, the method proposed here could also be viewed as incorporating a version of the contrasting-groups method.

Exceptionally good performances could be given a rating of 8, and especially poor performances could get a rating of 2. Performances a bit above or below the performance standard would get ratings of 6 or 4, respectively.

The performance standard defining minimally competent performance should be stated as clearly as possible. The judges would need to become thoroughly familiar with the performance standard before starting the rating, because the performance standard anchors the rating scale. In standard setting for credentialing examinations, the judges generally get to specify what is to count as minimally acceptable performance as they flesh out what they mean by expressions like, "readiness for safe and effective practice" at the appropriate level (i.e., entry-level for licensure and a more advanced and specialized level for certification), and this process of defining the standard is interwoven with the training of judges.

In all examinee-centered studies, the choice of the specific performances to be reviewed is an important issue. The papers presented to the judges undoubtedly lead the judges to form some impression of the available pool of performances, and this impression could easily influence the ratings. Given that this is likely to be the case, it seems prudent to give the judges an accurate impression about the range of candidate performances, rather than an inaccurate impression. Therefore, in implementing the generalized examinee-centered method, the papers presented to the judges would be representative of the performances in the candidate pool; this could be accomplished by ordering the papers in terms of total score (with random ordering for tied scores) and taking a spaced sample. Rating a sample of 20 papers selected in this way should give the judges a good indication of the range of performance in the pool of candidates.

Assuming that judges rate 5 or more sets of 20 papers each and that we average the ratings on each paper, we would have at least 100 points to fit the curve. This should provide a rather stable specification of the curve to be fitted, and therefore of the passing score corresponding to a rating of 5. (Note, that if the passing score were near the extremes of the distribution of observed scores, it might be necessary to review additional papers or to use a different method).

Generating a Passing Score

It is reasonable to expect that performances that get high ratings will generally get higher scores than performances that get low ratings. The test scores and the generalized examinee-centered method's ratings are, after all, evaluations of the same performances. If they are not positively related, something must be seriously wrong with either the test scores or the ratings or both.

However, the relationship between the ratings and test scores is not likely to be linear. Papers with the lowest scores are likely to get ratings of 2 or 3. At somewhat higher levels on the score scale, papers will tend to get higher ratings. At some point in this progression, most performances will be getting ratings of 7 or 8. When this happens, it will not be possible for the ratings to improve further as a function of test score, and the curve will flatten out.

So, the curve specifying the relationship between test scores and ratings may tend to have a logistic shape: low and relatively flat for low scores, increasing fairly rapidly in some middle range of scores, and then flattening out again as it approaches its upper asymptote. If this pattern occurs, the threshold score can be set by fitting a two-parameter logistic curve to the ratings and defining the threshold score as the point on the score scale that corresponds to an expected rating of 5. (If the data do not exhibit the form of the logistic function, linear or polynomial regression could be used.)

Curve Fitting

The use of a rating scale that extends from 2 to 8 is likely to be convenient and easy to remember for standard-setting judges, but a different scale is more convenient for curve fitting. The new variable, r_p, can be defined as:

$r_p = .95$ for a rating of 8

$r_p = .80$ for a rating of 7

$r_p = .65$ for a rating of 6

$r_p = .50$ for a rating of 5

$r_p = .35$ for a rating of 4

$r_p = .20$ for a rating of 3

$r_p = .05$ for a rating of 2

The choice of .05, .20, .35, .50, .65, .80, .95 for the values of r_p is, of course, arbitrary. A different set of values, for example .01, .2, .4, .5, .6, and .8, .99 could serve as well. Preliminary analyses suggest that the particular choice does not make much difference, as long as these values are symmetric, have a middle value of .5, and do not get too close to 0.0 or 1.0. The linear transformation from ratings to r_p does not change the empirical content of the data and makes the equations much simpler.

Assuming that the ratings approximate a logistic function of the scores r_p will also approximate a logistic function of the test scores, x_p:

$$r_p = \frac{e^{ax_p - b}}{1 + e^{ax_p - b}} \qquad (1)$$

This relationship can be transformed to:

$$\frac{r_p}{1 - r_p} = e^{ax_p - b} \qquad (2)$$

Taking the natural log of both sides, we have:

$$\ln\left(\frac{r_p}{1 - r_p}\right) = ax_p - b \qquad (3)$$

Equation 3 is basically a simple linear equation with the score, x_p, as the independent variable, and the dependent variable, y_p, specified as :

$$y_p = \ln(\frac{r_p}{1 - r_p}) \qquad (4)$$

Since r_p is never equal to 0 or 1, y_p is always defined. The particular set of values used for r_p will have an impact on the values of the parameters, particularly the "a" parameter, but this effect is expected to cancel out in determining the passing score.

Given ratings of a number of test performances, and scores for these performances, estimates of the parameters, a and b, can be easily obtained using least squares. As noted earlier, a convenient way to draw a sample of papers for the generalized examinee-centered method would be to take a spaced sample across the full range of scores in the population. This kind of sample provides a solid basis for curve fitting and also provides the judges with information about the range of performances in the population.

Once the constants, a and b, have been estimated, they can be put into Equation 1, and used to generate the expected value of r_p for any score x_p. The passing score can then be set equal to the value on the score scale that yields an expected value of .5 for r_p. A value of .5 for r_p is associated with ratings of 5 by the judges, and the performances given ratings of 5 are judged to have just met the standard.

The advantage of using this curve fitting methodology is that it makes it possible to use all of the data for a range of test papers to set the passing score, rather than using one or two selected subgroups, as in the borderline-group method or the contrasting-groups method.

Applying the Goldilocks Criteria to the Generalized Examinee-Centered Method

The generalized examinee-centered method has two features that may serve to keep the standard from being too low or too high. The first feature involves the task assigned to the judges. In all of the examinee-centered methods, the judges are asked to review actual candidate performances and decide whether the candidate has achieved a level of performance consistent with the performance standard.

Since all of the tasks included in the assessment are presumably relevant to good practice. and, since most judges are likely to prefer that practitioners be well prepared to deal effectively with any situations they might encounter in practice, the nature of the assignment given to the judges probably tends to push the passing score up, rather than down. What member of a professional standard-setting panel will want to pass candidates who do not know what to do in any practice situation (e.g., to say that it is OK for passing candidates to provide inadequate services in any practice situation)? So, the expert judges are not likely to set the passing score too low.

Second, the judges are evaluating actual candidate performances that are representative of the range of performance in the population of candidates. Because the judges are evaluating actual candidate performances, they are not likely to set the standard so high that nobody would pass, nor so low that anyone could pass. So, the sample of performances supplied to the judges is likely to have a moderating influence on the outcome, making it less likely that the judges will set the passing score much too high or much too low.

As indicated earlier, the generalized examinee-centered method employs a sample of papers with a distribution of scores paralleling that in the population. If the sample of performances is representative of the population of candidates and if the judges know this, we have a built-in reality check for the standard-setting process.

CONCLUSIONS

Standard setting is an essential component in licensure and certification testing. Given the way licensure and certification tests are used to make decisions, it is necessary to have a specific passing score. We recognize that the performance of a candidate with a score just above the passing score is not much different from that of a candidate with a score just below the passing score, but a decision has to be made about whether or not each individual is to be given a license or to be certified.

Standard setting is not a problem of estimation, but rather, a policy decision, and as such, represents a balancing of competing goals. We want the policy decision to be sensible, reasonable, and defensible. We want it to be informed by good judgement and, to the extent that it is helpful in making reasonable policy decisions, by data. But we cannot expect it to be true or accurate in any absolute sense.

Standard setting decisions, like any policy decisions, are to be evaluated in terms of their consequences. We do not ask whether a policy is true or not, or how accurate a policy is. Rather, we ask how well it works, and how well it works depends on the consequences, positive and negative, of using the test scores in conjunction with the passing scores to make decisions. So, to the extent that we are concerned about the validity of passing scores, it is the evaluation of consequences that is of central concern.

In evaluating a standard setting procedure, the Goldilocks criteria suggest that the procedure should be designed such that it will tend not to set the passing score too low nor too high. The aim is to have a high probability of avoiding especially inappropriate passing scores, and to have a good chance of getting a passing score that provides a reasonable balance between the goal of protecting the public and the goals of maintaining an adequate supply of practitioners and protecting the rights of candidates.

The most commonly used test-centered method, the Angoff method, as it is usually implemented, appears to suffer from inadequate controls on the judges' natural tendency to set high standards. The use of feedback on candidate performance (e.g., in the form of p values) would probably be a more effective control on this tendency to set high standards if it were given earlier, either before or during the first round of ratings.

Examinee-centered methods are receiving increasing attention because of growing interest in the use of performance testing. The generalized examinee-centered method described in this paper was designed to include features that would be useful for setting standards for such tests, making more efficient use of the judges' ratings, while at the same time, keeping the passing score from being set too high or too low.

REFERENCES

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

Busch, M.I. & Jaeger, R.M. (1990). Influence of type of judge, normative information, and discussion on standards recommended for the National Teacher Examination. Journal of Educational Measurement, 27, 145-163.

Clauser, B. & Clyman, S. (1994). A contrasting groups approach to standard setting for performance assessments of clinical skills. Academic Medicine, RIME Supplement.

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

Harker, J. & Cope, R. (1988). The effects of several variables on judgementally-obtained cut scores. Paper presented at the annual meeting of the National Council on measurement in Education, New Orleans.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. Educational Evaluation and Policy Analysis, 4, 461-476.

Jaeger, R. M. (1989). Certification of Student Competence (pp. 485-514). In R. L. Linn (Ed.), Educational Measurement, 3rd ed. New York: American Council on Education and Macmillan.

Kane, M. (1992). An argument-based approach to validity. Psychological Bulletin, 112, 527-535.

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

Kane, M. (1995). Examinee-centered vs. task-centered standard setting. In Joint Conference on Standard Setting for Large-scale Assessments, Proceedings, Vol. II, 119-141

Kane, M. & Wilson, J. (1984). Errors of measurement and standard setting in master testing. Applied Psychological Measurement, 8, 107-115.

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. Evaluations and the Health Professions, 17, 2, 178-197

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.

McLaughlin, D.H. (1993). "Validity of the 1992 NAEP Achievement-Level-Setting Process," in Setting Performance Standards for Student Achievement: Background Studies. Stanford, CA: The National Academy of Education.

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

Norcini, J., & Shea, J.(1997). The credibility and comparability of standards. Applied Measurement in Education, 10, 39-59

Norcini, J. (1994). Research on standards for professional licensure and certification examinations. Evaluations and the Health Professions, 17, 2, 160-177

Norcini, J., Shea, J., & Ping J. (1988). A note on the application of multiple matrix sampling to standard setting. Journal of Educational Measurement, 25, 2, 159-164

Norcini, J., Shea, J., & Kanya, D. (1988). The effect of various factors on standard setting. Journal of Educational Measurement, 25, 7-65

Plake, B. (1996). Setting Performance Standards for Professional Licensure and Certification. Paper presented at workshop on NAEP Achievement Levels: Setting Consensus Goals for Academic Achievement, Washington, D.C.

Shepard, L. (1984) Setting performance standards. In R. A. Berk (Ed.) A Guide to Criterion-Referenced Test Construction. Baltimore: Johns Hopkins Press.

Shepard, L., Glaser, R., Linn, R., & Bohrnstedt, G. (1993). Setting Performance Standards for Student Achievement. National Academy of Education, Stanford University, Stanford, CA.

Sireci, S.A. & Biskin, B.M. (1992). Measurement practices in national licensing examination programs: A survey. CLEAR Exam Review, 3, 21-35.

Tamblyn, R. (1994). Is the public being protected? Suboptimal medical practice through training programs and credentialing examinations. Evaluations and the Health Professions, 17, 2, 198-221.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

TM027348

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: JUSTFYING THE PASSING SCORES FOR LICENSURE AND CERTIFICATION TESTS

Author(s): KANE, M. CROOKS, T. COHEN, A.

Corporate Source: UNIV. OF WISCONSIN, MADISON

Publication Date: MARCH, 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☑ Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□ Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Michael T. Kane

Printed Name/Position/Title: PROF. MICHAEL KANE

Organization/Address:
DEPT. OF KINESIOLOGY
2000 OBSERVATORY DR.
MADISON, WI 53706

Telephone: (608)265-2891

FAX: (608)262-1656

E-Mail Address:

Date: May 20, 1997

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064

800 464-3742 (Go4-ERIC)

April 25, 1997

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. We feel you have a responsibility to make your paper readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. If you have submitted your paper, you can track its progress at http://ericae2.educ.cua.edu.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are soliciting all the AERA Conference papers and will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and stet **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 1997/ERIC Acquisitions
The Catholic University of America
O'Boyle Hall, Room 210
Washington, DC 20064

← To

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/E

ERIC® Clearinghouse on Assessment and Evaluation