Originally appearing in THE BAR EXAMINER*, Vol. 71, No. 4, November 2002. Reprinted with permission of the National Conference of Bar Examiners.*

CONDUCTING EXAMINEE-CENTERED STANDARD-SETTING STUDIES BASED ON STANDARDS OF PRACTICE

by Michael T. Kane, Ph.D.

This is the second of two related articles on standard setting. The first, which appeared in the last issue of THE BAR EXAMINER (Kane 2002), addressed the general issues involved in developing standards for licensure examinations. It argued that performance standards for bar examinations and other licensure examinations should be grounded in practice requirements and that a close link to practice requirements is most easily achieved by employing examinee-centered standard-setting methods. This article discusses specific design aspects of examinee-centered standard-setting procedures, and suggests ways to relate performance standards to practice requirements.

Bar examinations are designed to provide the public with assurance that those admitted to practice are ready for practice at an entry level. The ***performance standard*** for a bar examination provides a description of minimal standards for entry-level practice. The corresponding ***passing score***, which is intended to differentiate those who have achieved the performance standard from those who have not, is a specific score on the bar examination. A candidate passes the test if his or her test score is at or above the passing score, and fails if the test score is below the passing score. The performance standard provides a conceptual definition of readiness for practice. The passing score provides an operational definition of readiness for practice.

Most of the existing standard-setting methods can be categorized as ***test-centered methods*** or as ***examinee-centered methods*** (Jaeger 1989; Zieky 2001). In test-centered methods, participants in the study review the items or tasks in the test and decide on the level of achievement on each task or item (e.g., the probability of getting an answer right or the expected score on an essay question) associated with just meeting the performance standard.

In the examinee-centered methods, participants in the standard-setting study evaluate actual candidate performances relative to the performance standard. For example, in the borderline-group method (Livingston and Zieky 1982; Zieky 2001), the participants identify candidates whose overall performances are right around the specified performance standard. These "borderline candidates" are marginal in the sense that their performances just meet the performance standard. The passing score is then set equal to the median test score for these candidates.

In another examinee-centered method, the contrasting-groups method, the participants categorize candidates into two groups, an upper group who have clearly met the standard, and a lower group who have not met the standard. The passing score is chosen so that it differentiates between these two groups as well as possible (Livingston and Zieky 1982; Zieky 2001).

As indicated in the previous article (Kane 2002), an examinee-centered approach has strong advantages over a test-centered approach for setting practice-based standards for a licensure examination. To the extent that the participants have experience in applying standards of practice, it is in applying them to the actual performance of practitioners in real practice situations, and the examinee-centered methods involve this kind of judgment. In examinee-centered methods, the participants review actual candidate performances on practice-relevant tasks and decide whether they are acceptable or not. In contrast, in the test-centered methods, the participants review test items or tasks, and evaluate how difficult they will be for the marginal or borderline candidates.

Examinee-centered standard-setting methods all have certain features in common. A group of participants evaluates the performance of a sample of individuals who have taken the test (and for whom scores are known) relative to the performance standard. The participants are individuals with the experience necessary to evaluate performances relative to the performance standard. The candidate performances that are evaluated are referred to as "criterion performances."

For licensure and certification tests, the criterion performances are likely to involve candidate performances on some part of the test. For example a candidate for admission to the bar might be evaluated in terms of his or her overall performance on the essay test or on several of the essays. However, unlike the procedures that are typically used to score each essay, the participants would evaluate the candidate's overall performance on the essay test or part of the essay test relative to the performance standard and would categorize each candidate's criterion performance as "clearly failing", "borderline," or "clearly passing."

A recently developed group of examinee-centered methods have participants rate the criterion performances on an evaluative scale defined in terms of the performance standard. Using the rating scale, the participants in the study would rate each candidate's performance on the criterion measures. For example, participants can be asked to rate each performance using the following scale (Cohen, Kane and Crooks 1999):

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 candidate's performance seems to be just consistent with the performance standard (i.e., borderline), neither clearly above nor clearly below the standard, the paper would be given a rating of 5.

If a participant thinks that a particular performance is clearly better than that specified in the performance standard, the participant would assign the

performance a rating of 7. Similarly, if a participant thinks that the performance clearly does not satisfy the performance standard, in the sense that the performance provides little indication of achievement of the performance standard or provides evidence indicating a failure to meet the performance standard, the participant would assign it a rating of 3. 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.

After the participants have rated each candidate's criterion performance, their ratings would be averaged to provide a single, overall rating for each candidate. In general we would expect candidates who get high ratings on the criterion performances to have relatively high scores on the test as a whole and candidates with low ratings on the criterion performances to have relatively low scores on the test as a whole. As a result, using a standard statistical method (in particular, regression analysis) it is possible to establish a mathematical relationship between the participants' overall ratings of the candidates' criterion performances and the candidates' scores on the test. The test score associated with the rating-scale value indicating that a performance just meets the performance standard (e.g., 5 in the example above) is taken as the passing score. Similar examinee-centered methods have been developed by Faggen (1994), Cohen, Kane, and Crooks (1999), Kingston, Kahl, Sweeney, and Bay (2001), and Jaeger and Mills (2001).

Standard-setting studies generally take from two to three days, with most of the first day being devoted to orientation and training. At the beginning of the training process, the participants would agree on a preliminary version of the performance standard. After a satisfactory statement of the performance standard is developed and the participants are comfortable with the rating materials and process, they would evaluate a substantial number of criterion performances in batches, with some discussion of their ratings after each batch. As they conduct these evaluations, they would have the opportunity to revise the performance standard in order to correct any gaps or limitations that are identified as the study progresses. The ratings of candidate performances based on the performance standard can then be related to scores on the MBE scale, using regression models, and a proposed passing score can be identified.

The goal would be to come out of the study with a performance standard and passing score that are closely tied to each other, and that are directly related to current standards of performance in practice. These results are then submitted to the appropriate policy-making body which decides on the performance standard and passing score to be adopted.

Defining the Performance Standard

The policy-making body for a licensure examination has the responsibility of determining the performance standard for passing, but for pragmatic reasons, tends to do so in general terms (e.g.,"readiness for effective practice"). The policy makers typically appoint a standard-setting panel, which is assigned two tasks, the fleshing out of the performance standard and the estimation of the corresponding passing score (Corneille 2001).

Most of the literature on standard-setting focuses on the second of these two tasks, the estimation of a passing score given a performance standard. The emphasis is on the statistical problem of finding the passing score that corresponds to an existing performance standard. Although it is usually not given

much attention in discussions of formal standard-setting studies, the specification of the performance standard is a critical component in applications of these methods. The performance standard presumably guides the participants in their evaluation of candidate performances, and thereby determines the passing score to a large extent.

How should the participants in the panel set about defining the performance standard? Given that the purpose of licensure is to protect the public, it seems evident that the participants should focus their attention on the requirements of entry-level practice in their jurisdiction and on standards of practice in this context. According to the ***Standards for Educational and Psychological Testing*** (AERA et al. 1999),

> The validity of the inference drawn from the test depends on whether the standard for passing makes a valid distinction between adequate and inadequate performance. . . . Standards must be high enough to protect the public, as well as the practitioner, but not so high as to be unreasonably limiting (p. 157).

In particular, the panel should not give too much attention to other considerations, such as: the current supply of new practitioners, the performance standards adopted by other jurisdictions, what is or is not taught in law schools, what students should know after a specific course, or what would make for financial and professional success in practice. All of these considerations can have some relevance to the policy decision to be made, but the primary focus should be on readiness for entry-level practice.

For a bar examination, the performance standard would presumably describe the kinds of situations that the entry-level practitioner would be likely to encounter, the problems that they would be expected to address in these situations, and the kinds of performance to be expected of them. The minimally competent entry-level practitioner would presumably not be expected to address the more complex and specialized problems, but would be expected to deal effectively with common problems in commonly encountered legal situations.

As indicated below, the definition of the performance standard and the specification of the corresponding passing score typically occur in tandem. The participants start the process by agreeing on a preliminary version of the performance standard, which is then used to evaluate candidate performances relative to the standard. As the participants encounter ambiguities in their evaluations of candidate performances relative to the performance standard, they can revise or expand the statement of the performance standard.

Choice of Criterion Measure

Assuming that an examinee-centered method is to be used, it is necessary to have some criterion performances evaluated for a sample of candidates for whom scores on the bar examination are available.

Since the performance standards are to be practice-based, it is desirable that the criterion measure involve performances in the kinds of situations and on the kinds of tasks that arise in practice. For professional licensure examinations, it is generally not feasible to use measures of actual performance in practice for this purpose, and therefore, some standardized measure of performance is used as the criterion (e.g., performance on essay questions based on realistic practice situations).

In general, it should be easier and more natural to apply the standards of practice to such perform-

ances than to performances (e.g., answers to multiple-choice items) that are not very similar to actual performance in practice. From this point of view, ratings of candidate performances on essays or performance tasks probably provide a more appropriate criterion measure than performances on the multiple-choice part of a bar examination.

The participants should be asked to review a substantial sample of performance for each candidate included in the study. Research on performance testing in many contexts has suggested that performance tends to vary from task to task, and therefore a number of task performances is needed to get a reliable indication of a candidate's general level of performance. Further, the inclusion of some range of performances for each candidate may help to offset any tendency of the participants (who probably have a high level of experience and expertise in some area of practice) to evaluate a candidate's performance on a particular question in terms of the expected performance of a specialist in that area.

The number of essays and the types of essays included in bar examinations vary from state to state. It is not necessary to use all of the essays for each candidate whose performance is being reviewed, but it is probably desirable to use the candidate's answers to several essays if they are available. We want a large enough sample of performance to get a fairly reliable indication of each candidate's level of competence across situations and tasks.

Since the participants are called on to evaluate the criterion performances as a whole, relative to the performance standard, the criterion performances cannot be very long. The goal is to use criterion performances that are long enough to provide a good indication of the candidates' readiness for practice as specified by the performance standard, without being so long as to overwhelm the participants. A criterion consisting of a candidate's answers to three or four essay questions would probably be sufficient.

Samples of Candidate Performances

A fairly large and representative sample of candidate performances needs to be included in the study. Fifty to one hundred candidate performances would probably be sufficient (Cohen, Kane, and Crooks 1999; Jaeger and Mills 2001), but guidelines for the number of candidate performances required by various examinee-centered methods have not been developed.

Presumably, the candidate performances should cover the score range in which the passing score is likely to fall. There are at least two options to consider in this regard. If a passing score already exists, and the purpose of the study is to "check" on this standard and possibly adjust it up or down, it would make sense to have an especially high concentration of performances for candidates with test scores in the region of the score scale around the passing score. In the absence of any a priori information about the approximate location of the passing score, the performances included in the standard-setting study should cover the full range of performance in the population.

Under the first of these two models, the participants would evaluate the performances of candidates with scores around the current passing score (e.g., from 10 or 20 points below to 10 or 20 points above the current passing score on the MBE) in order to determine whether the passing score should be raised or lowered, and if so by how much. The performances to be evaluated would be distributed uniformly across the range of scores under consideration. Under this model, the size and cost of the standard-setting study could be much smaller than

under the second model. This kind of "standard-adjusting" study would make a lot of sense in those cases where the current passing score has been in place for some time and seems to be working reasonably well (i.e., most individuals admitted to practice perform satisfactorily, but the least well prepared seem barely satisfactory). Since drastic changes in the passing score are potentially very disruptive and counterproductive, such changes should not be made if they are not necessary. In cases where large changes in the passing score do not seem justified, it would make sense to adjust the current passing score rather than setting an entirely new standard.

Under the second model, no assumption is made a priori about where the passing score is likely to fall on the score scale, and the participants would evaluate the performances of candidates with scores across the full range of scores. The advantage of this approach is that it does not rely on any presumption about where the passing score should be placed. This approach is the obvious choice if the test is new and no passing score exists. It also makes sense whenever policymakers prefer not to make any assumptions up front about where the passing score is likely to fall.

In cases where the full range of performances is to be included, I am in favor of selecting the sample of candidate performances so that it is representative of the population. The participants are instructed to base their judgments on how well each candidate performance compares to the performance standard, and not to base it on norm-referenced judgments (e.g., assumptions about how many candidates should pass). Nevertheless, the participants' judgments are likely to be influenced by the distribution of performances that they are asked to review. If they are given a selection of poor performances, they are likely to set the passing score low enough so that not everyone fails. If they are given only the best performances, they are more likely to set very high passing scores. Given that the sample of performances used may have an impact on the policy being developed, it seems better to provide samples that represent the distribution of performances across the full range.

Selection of Participants

All standard-setting methods involve judgments and therefore all require qualified participants. The standard-setting participants should have extensive knowledge of the content covered by the examination and of the requirements of entry-level practice. The participants in an examinee-centered study must have enough technical expertise to evaluate candidate performances. Familiarity with the population of candidates for admission to legal practice and with the work of newly admitted lawyers should help to keep the standard realistic. In setting standards for a bar examination, good pools of potential participants would include practicing lawyers who supervise newly admitted lawyers, law school faculty who are also involved in practice, and judges who regularly see the work of newly admitted lawyers.

It is important that the outcomes not depend much on the specific sample of participants in the study (Norcini and Shea 1997). Given the heavy reliance on judgment in standard-setting studies, the results are expected to vary somewhat as a function of who is on the panel. This potential source of error is controlled to some extent through the selection of qualified participants for the study and the thorough training of these participants. In addition, the inclusion of a reasonably large number of participants in

the study helps to control the variability due to the sampling of participants. Raymond and Reid (2001) provide a good summary of research on test-centered standard-setting methods and conclude that 10 to 15 participants would be adequate. The numbers required for examinee-centered methods have not been investigated as extensively as the numbers required for the test-centered methods, but acceptable results have been obtained with 10 to 20 participants in a series of examinee-centered studies for eighth- and tenth-grade tests in various content areas (Cohen, Kane, and Crooks 1999). Using a similar examinee-centered method, Jaeger and Mills (2001) suggested that a total of 15 to 20 participants would be adequate. In order to control for any factors that might influence the results for any group of participants, the participants can be divided into three or four groups, with each of the groups to work independently.

Instructions to Participants

The participants should get thorough training on what they are expected to do. They should be introduced to the purpose of standard setting, the method to be used, and the materials to be used. The rating scale to be used would be described, and the participants would get a chance to practice using the scale to evaluate candidate performances and to reach agreement on their ratings and on the performance standard, as described above. Training should continue until both the participants and those conducting the study are satisfied that the participants are comfortable with the performance standard and the process to be used in translating the performance standard into a passing score. Previous experience in scoring examinations is useful but not sufficient.

Early in this process, the participants would develop a general practice-based performance standard based on existing standards of practice in the profession. The initial statement of the performance standard will be refined during the study, but it is important to start with a clear focus on generally accepted standards of practice.

As noted in the previous Bar Examiner article (Kane 2002), there are many problems associated with defining performance standards, and the standard will not be specified with great precision. The practice-based performance standard would not be defined in terms of what it would be "good to know" in some general sense, nor would it be defined in terms of what is needed to be a great success in a major law firm, but rather in terms of minimal requirements for practice. At the start of the standard-setting study, the performance standards can be stated in fairly general terms. As the participants practice using the rating scale to evaluate candidate performances, they will have the opportunity to discuss and revise the performance standard.

Concluding Remarks

Bar examinations should identify candidates who are minimally competent to practice law at the entry level. Standards for bar exams are designed to provide adequate protection to the public, while not subjecting candidates to arbitrary requirements.

In this article, I have described an examinee-centered standard-setting model, as contrasted with a test-centered model. The examinee-centered methods have advantages over test-centered methods with regard to the participants' experience in evaluating candidate performances.

Jurisdictions that contemplate beginning a standard-setting study should remember that standard setting is not an exact science. In addition, the final decisions to be made regarding standards are

matters of policy. While empirical standard-setting studies can provide useful information for policy makers, they cannot provide all of the answers. Thus, those with responsibility for licensure decisions should be involved in the design of standard-setting studies, and in interpreting and making decisions about the results.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. THE STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING.

Brennan, R. 2001. Some problems, pitfalls, and paradoxes in educational measurement. EDUCATIONAL MEASUREMENT: ISSUES AND PRACTICE 20(4):6-18

Cohen, A., Kane, M., and Crooks, T. 1999. A generalized examinee-centered method for setting standards on achievement tests. APPLIED MEASUREMENT IN EDUCATION 12:343-366.

Corneille, M. 2001. Examining passing examination scores. THE BAR EXAMINER 70(4):17-20.

Faggen, J. 1994. Setting standards for constructed-response tests: An overview. Research Memorandum, Princeton, N.J.: Educational Testing Service.

Jaeger, R. 1989. Certification of Student Competence. In EDUCATIONAL MEASUREMENT. 3rd ed. Edited by R. Linn. New York: American Council on Education and Macmillan 485-514.

Jaeger, R., and Mills, C. 2001. An integrated judgment procedure for setting standards in complex, large-scale assessments. In. SETTING PERFORMANCE STANDARDS: CONCEPTS, METHODS, AND PERSPECTIVES. Edited by G. Cizek. Mahwah, N.J.: Lawrence Erlbaum 313-338.

Kane, M. 2002. Practice-based standard setting. THE BAR EXAMINER 71(3):14-24.

Kingston, N., Kahl, S., Sweeney, K., and Bay, L. 2001. Setting performance standards using the body of work method. In SETTING PERFORMANCE STANDARDS: CONCEPTS, METHODS, AND PERSPECTIVES. Edited by G. Cizek. Mahwah, N.J.: Lawrence Erlbaum.

Livingston, S., and Zieky, M. 1982. PASSING SCORES: A MANUAL FOR SETTING STANDARDS OF PERFORMANCE ON EDUCATIONAL AND OCCUPATIONAL TESTS. Princeton, N.J.: Educational Testing Service.

Norcini, J., and Shea, J. 1997. The credibility and comparability of standards. APPLIED MEASUREMENT IN EDUCATION 10(1):39-59.

Raymond, M., and Reid, J. 2001. Who made thee a judge? Selecting and training participants for standard setting. In SETTING PERFORMANCE STANDARDS: CONCEPTS, METHODS, AND PERSPECTIVES. Edited by G. Cizek. Mahwah, N.J.: Lawrence Erlbaum 119-157.

Zieky, M. 2001. So much has changed: How the setting of cutscores has evolved since the 1980s. In SETTING PERFORMANCE STANDARDS: CONCEPTS, METHODS, AND PERSPECTIVES. Edited by G. Cizek. Mahwah, N.J.: Lawrence Erlbaum 19-51.

MICHAEL T. KANE, PH.D., is the Director of Research for the National Conference of Bar Examiners.