

PRACTICE-BASED STANDARD SETTING

by Michael T. Kane, Ph.D.

In June of 1998, twenty-nine million Americans acquired a chronic medical condition. The condition is seldom fatal in itself, but in its more severe forms it can be disabling and it predisposes its sufferers to a number of serious illnesses, including heart disease, diabetes, and some forms of cancer. The condition? Being overweight.

This epidemic was not caused by an orgy of overeating; it was caused by a change in the National Institutes of Health cutscore relating to the body mass index (BMI), a measure of the body's percentage of fat. Under the new weight guidelines, the criterion for an overweight 5'6" adult went from a BMI of 28 or about 170 pounds to a BMI of 25 or about 155 pounds (Greenberg 1998). The change was motivated by evidence linking higher weight levels to the incidence of various illnesses. The adoption of the new guidelines added twenty-nine million people to the rolls (no pun intended) of the overweight (Shapiro 1998) as they were "transformed overnight from fit to fat" (Cimons 1998).

A change in the passing score on a licensure examination can have a similar, dramatic effect on pass/fail decisions. All else being equal, an increase in the passing score will lead to a decrease in the passing rate, and a decrease in the passing score will lead to an increase in the passing rate. And depend-

ing on where the passing score is in relation to the score distribution for a population, even modest changes in the passing score can produce dramatic changes in pass rates. In contrast, the impact of other changes in test design (e.g., changing content specifications) is less predictable and the results usually much less dramatic.

Given the high stakes associated with bar examinations, it is important that the passing scores on these examinations be as defensible as possible. Within measurement theory, test-based decisions are evaluated in terms of their "validity," which is defined as the extent to which the decisions are supported by evidence for the appropriateness of the proposed interpretation and use of the test scores. The evidence to be considered in the evaluation of a testing program includes both supporting evidence and evidence that might cast doubt on the proposed interpretation. In fact, validity can be viewed as the extent to which a proposed interpretation and use of test scores can withstand thoughtful criticism:

The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it. (Cronbach 1980, 103)

It is important from psychometric, legal, and ethical points of view that the passing scores on licensure examinations be defensible in the sense that they are capable of withstanding critical scrutiny (Norcini and Shea 1997).

One of the most serious challenges to standard setting has been the charge that the resulting passing scores are arbitrary (Glass 1978). An effective response to charges of arbitrariness is a demonstration of the relationship between the passing score and the goals of the testing program in which they function. The avowed purpose of licensure is to protect the public, and therefore, it seems appropriate that passing scores on licensure examinations be grounded in practice requirements.

This article addresses general issues involved in developing standards for licensure examinations, and concludes that (1) performance standards for these examinations should be grounded in practice requirements, and (2) a close link to practice requirements can be provided most easily by employing examinee-centered standard-setting methods based on standards of practice. A second article that will appear in the next issue of *THE BAR EXAMINER* will discuss specific design aspects of examinee-centered standard-setting procedures, and suggest ways to relate standards to practice requirements.

PERFORMANCE STANDARDS AND PASSING SCORES

Standard-setting procedures are designed to establish and justify passing scores. They address the basic policy issue of how good a performance must be in order to be considered good enough. Today, a formal standard-setting study may involve an elaborate data collection design with several rounds of data collection followed by sophisticated statistical

analyses (Norcini and Shea 1997; Kingston, Kahl, Sweeney and Bay 2001; Zieky 2001). Yet, at its core, standard setting is an attempt to establish a reasonable basis for making decisions based on test scores. Statistical analyses can provide useful information, but ultimately, those responsible for setting policy must decide on the standard to be used.

Standard setting has two components, a performance standard and an associated passing score (Kane 1994). The *performance standard* is a qualitative description of the required level of competence, based on the intended purpose of the test. For licensure examinations like the bar examination, the purpose is to provide the public with assurance that those admitted to practice have demonstrated some required level of competence. The performance standard would describe what it means to be ready for practice. The description may be very general or it may describe the requirements in some detail.¹

The passing score (sometimes called a “cutscore”) is a specific point on the score scale that is used in making pass/fail decisions about candidates. A candidate passes 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 passing score is intended to differentiate those who have achieved the performance standard from those who have not.

The passing score and the performance standard are like the two sides of a coin. The passing score is the operational version of the performance standard; the performance standard articulates the proposed interpretation of the passing score. To validate the decision being applied is to show that the passing score reflects the requirements in the performance standard and that the performance standard is reasonable and appropriate, given the decision to be made.

THE CHALLENGE IN ESTABLISHING DEFENSIBLE STANDARDS

As noted above, even modest changes in the passing score can have a dramatic impact on test-based decisions. If a passing score is raised, some examinees who would have passed under the original passing score will fail under the new passing score, but everybody who would have failed before will still fail under the new passing score. As a result, the passing rate will go down. Similarly, decreasing the passing score will increase the pass rate. If the passing score is in the middle of the score distribution where candidate scores tend to be concentrated, even a small change in the passing score can produce a substantial change in the pass rate. Once the distribution of scores is known (or predicted), the pass rate is an entirely predictable function of the passing score.

If we have two groups of candidates with different score distributions, a change in the passing score will often have a greater impact on one of those groups. For example, if the passing score is close to the middle of the score distribution for the lower scoring group, even a modest increase in the passing score can substantially increase the failure rate for this group. Assuming that the passing score is in the lower end of the score distribution for the higher scoring group (where there are fewer scores), an increase in the passing score would have a smaller impact on the failure rate for this group.

Discussions of passing scores in personnel selection have frequently tended to focus on adverse impact and have been concerned about the legal defensibility of standards. For example, in discussing physical ability tests for employment, Campion (1983) states that:

The conceptual link between the job requirements and the cut-off scores chosen for the selection tests must be made explicit, and it must be documented and defensible. Physical abilities tests do have adverse impact against females; they probably will be legally challenged; and the cut-off scores determine the degree of adverse impact. (p. 545)

Given the impact of passing scores for high-stakes testing, and their potential for adverse impact, the rationale for the performance standard and the corresponding passing score should be clear and persuasive. Bahls (2001), Klein (2001), Merritt (2001), and Merritt, Hargens and Reskin (2001) have recently discussed this issue as it applies to bar examinations.

Two kinds of ambiguity have been recognized in standard-setting efforts, one associated mainly with the performance standard and one with the corresponding passing score. The first source of ambiguity is the lack, in many cases, of any clear basis for defining the performance standard. The goals of testing programs are often stated in very general terms, and as a result, the performance requirements associated with the goals may not be clear. The second kind of ambiguity is the uncertainty associated with the estimation of the passing score, given a determined performance standard.

AMBIGUITY IN THE PERFORMANCE STANDARD

The performance standard is intended to provide an answer to the question of how much is enough (i.e., what level of performance evidencing the skill or knowledge being assessed by the test is to be considered adequate for a particular purpose); such questions are often hard to answer.

Most high-stakes testing programs have very general goals for which the performance requirements are not easily specified, and therefore it is difficult to make the case that the required performance level should be set at any one point rather than another. Competence as a lawyer is not a binary variable with two categories, competent and incompetent. There is a wide range of levels of competence in any activity, ranging from individuals with obvious and serious deficiencies to individuals with a thorough mastery of all aspects of the activity. Given that a higher level of competence is generally preferred to a lower level of competence, where should we draw the line? How good is good enough? In most cases, there is no clear and simple answer to this question.

But the situation is more complicated than this. Different individuals generally vary in their patterns of competency across different activities. Given any two activities, A and B, one candidate might be especially skillful in A but somewhat lacking in B. Another candidate might be especially skillful in B but lacking in A, while other candidates might be more consistent in their mastery of both A and B. What kind of summary evaluation should we give to each of these candidates? The practice of law and other professions involves a wide range of activities, and on each of these activities, candidates can exhibit a wide range of competency. Under these circumstances, defining a clear and generally acceptable performance standard is a formidable task.

Concerns about this kind of ambiguity led Gene Glass (1978) to claim that all standards are arbitrary and to suggest that it is “wishful thinking to base a grand scheme on a fundamental unsolved problem” (p. 237). A number of those who disagreed with Glass acknowledged that standards are arbitrary in the sense of being judgmental, but argued that they

need not be capricious (Popham 1978; Hambleton 1978; Shepard 1980).² Recently, Linn (2000) has suggested that, “The problem of setting standards remains as much a fundamental unsolved problem today as it was 20 years ago.” (p. 11)

Since there are situations where it is possible to set clear, defensible standards, I believe that Glass’s blanket rejection of standard setting is not justified. But it is also true as claimed by authors from Glass (1978) through Linn (2000) that there are serious fundamental problems in setting standards in many contexts, and these problems should be taken seriously.

AMBIGUITY IN THE PASSING SCORE ASSOCIATED WITH THE PERFORMANCE STANDARD

The second kind of ambiguity is associated with the inevitable lack of precision in setting a passing score on a score scale, even after the performance standard has been defined. For example, even if we accept a general requirement that lifeguards be able to swim some distance and swim back with an unconscious person, we might disagree about how far they should have to swim, how fast they should be able to do so, and how heavy the person to be brought back might be.

In employing a passing score on a test, we are imposing a sharp distinction where none previously existed. Wherever we set the passing score, there will not be much substantive difference between the candidate with a score one point above the passing score and the candidate with a score one point below the passing score. Judgments have to be made about the specific passing score to be used, and the panelists called upon to make these judgments are usually not in perfect agreement. Shepard (1980) argued that:

There is always error attached to the selection of cutoff scores. Individuals immediately on either side of the standard will be virtually indistinguishable from one another. With a good test, valid distinctions can be made between those who are well above or well below the standard; but pass-fail distinctions near the cutoff will have poor validity because a continuum of performance has been “arbitrarily” dichotomized. (p. 448)

Thus, even if the performance standard is fixed, the passing score selected may vary from one group of panelists to another or from one occasion to another. Much of the procedural complexity and statistical sophistication in formal standard-setting procedures is designed to minimize this variability in the passing score for a particular performance standard.

Lack of precision in setting a passing score is potentially serious, especially if it is substantial, but it is a less fundamental problem than ambiguity in the performance standard. If we are confident about the performance standards we can probably tolerate some uncertainty in the determination of the passing score. However, if we have not defined a clear performance standard, we have no clear basis for determining the passing score, even approximately.

THE DEFENSIBILITY OF PERFORMANCE STANDARDS AND PASSING SCORES

As noted in the preceding section, in many cases it is not easy to define performance standards that are clearly justified. But in some cases, performance standards seem to be clearly determined by the nature of the decision to be made. In this section, I will consider a few examples in which performance standards seem to be well founded and some in

which the standards seem fairly arbitrary, in order to get a clearer sense of what makes performance standards most defensible.

Many routine safety standards are designed to guard against specific risks, and these risks and the context determine the standard within fairly narrow limits. For example, a requirement that a building in an earthquake zone be strong enough to withstand the kinds of stresses typically produced by earthquakes in that area has an obvious justification in terms of public safety. The requirement is clearly and directly related to safety, and the stress to be withstood by the building is determined by the stresses produced by typical earthquakes. Similarly, a performance standard that a lifeguard must be able to swim a certain distance and swim back with a limp or struggling body bears a direct relationship to the main reason for having lifeguards at a beach or pool. Even in these two examples, the context and purpose allow for some flexibility in the standard. For example, an argument could be made for requiring that buildings in the earthquake zone be strong enough to withstand stresses greater than those produced by typical area earthquakes. If we are to opt for this extra margin of safety, how much stress should the buildings be able to withstand? In general, the goal and context of the decision do not determine the performance standard and the passing score precisely, but in some cases, they strongly constrain the range of reasonable choices.

Jackson (1994) describes a number of examples of performance standards for jobs requiring specific kinds of physical performance as a major component. For example, in some jobs, workers are required to lift heavy items. A job analysis in which the requirements of the job are documented can serve to identify the types and weights of the objects

to be lifted. If a job regularly requires the lifting of items weighting up to 50 pounds, a job requirement saying that new employees must be able to lift objects weighing up to 50 pounds without undue fatigue or risk of injury seems reasonable. A commonly used procedure is to estimate the strength requirements of the job (i.e., in a job analysis study) and then simulate the tasks in a battery of tests, with cutoff scores set at or somewhat above the maximum requirements of the job (Campion 1983). There are potential complications even here (e.g., how often is the 50-pound maximum reached, and is equipment available that could be used for heavy lifting?), but the basic standard is grounded in the requirements of the job. It does not seem arbitrary.

At the other extreme, the passing scores on high school graduation tests are not attached to any particular real world contingency. Politicians from the President down to the local school board are in favor of high standards, or “world-class” standards. Who would be in favor of low standards? But in spite of all of the political rhetoric, there is no obvious basis for deciding on how high the standards should be in various content areas. We may all agree that it would be better for high school graduates to know more math, rather than less math, but what kind should they know and how much is enough?

The problem is that high school graduation requirements are not tied to any specific real-world contingency, and as a result, there is no clear answer to the question of how much mathematics a high school graduate needs to know. The vocational and life plans of the students are too varied to provide a solid basis for standard setting in terms of personal consequences. High school graduates planning to pursue careers in science or engineering need a strong background in mathematics, including alge-

bra, geometry, trigonometry, and perhaps calculus. Many people get along fine in life with little more than a working knowledge of arithmetic. Should the performance standard in mathematics on a high school graduation test be set at a level appropriate for college-bound students (and if so, should it be set for those planning to major in physics or engineering or those planning to major in English or sociology?), or should the focus be on those planning to go directly into the world of work? The goal of raising standards may be laudable, but in itself, it provides essentially no help in setting standards, because standards can be raised indefinitely. We can and do set standards for high school graduation tests, but they are necessarily quite arbitrary, in the sense that there is no compelling reason for setting them at one level rather than another.

In general then, performance standards tend to be more defensible when they are based on well-defined, real-world performance requirements. But the performance standards’ simply being related to job requirements is not enough, especially if the decisions being made are high-stakes (e.g., licensure decisions). In addition, it also seems necessary that a failure to achieve the performance standard poses some specific risk. It is easier to define specific requirements and risks of failure for particular tasks (e.g., lifting 50 pounds) than it is to define the requirements and risks for a job that involves a wide range of tasks like the practice of law. If a job cannot be defined in terms of a limited number of specific tasks, it may be necessary to use generic requirements, but it is harder to define performance standards for such generic requirements.

Employment standards for police, firefighters, and correctional officers have received a lot of attention in the literature on employment testing,

mainly because of litigation involving adverse impact on women candidates (Cascio, Alexander and Barrett 1988). One perceived requirement in these occupations is physical strength and endurance. Firefighters are often called upon to engage in strenuous physical activity under difficult conditions for extended periods. Police officers may also need to engage in strenuous physical activity under some circumstances. In the past, this perceived need for strength was often addressed through height and weight requirements. These requirements, however, were at best only loosely connected to job requirements and risks (Jackson 1994). Campion (1983) summarized the results of a number of judicial decisions in which these height and weight requirements were thrown out because their relationship to job performance was not demonstrated:

It is only under rare circumstances that these types of standards have withstood legal scrutiny, such as when a minimum height is necessary of a pilot in order to see properly and reach all of the controls in an airplane cockpit (*Boyd v. Ozark Airlines*, 1977). (p. 529)

In the face of adverse impact, specific job requirements “must be shown to be necessary to safe and efficient job performance” (Campion 1983, 529). The basic problem with height and weight requirements for police or firefighters was that they were only loosely connected to the requirements and risks of the job (Jackson 1994).

The height and weight requirements have been replaced by assessments and passing scores that are more clearly related to the work requirements of firefighters and police. For example, Jackson (1994) summarizes research showing that “fire suppression work tasks have a substantial aerobic component”

(p. 71). Sothmann, Saupe, Jasenof, Blaney, Donahue-Fuhrman and Woulfe (1990) provide an estimate of the minimum level of aerobic capacity (33.5 mg/kg/min) required for firefighting activities. Even such research-based requirements run into complications in practice. Aerobic capacity declines with age, and therefore it may be advisable to hire individuals who exceed the minimum requirement by a substantial margin to allow for the subsequent decline. Nevertheless, a requirement for aerobic capacity of about 33.5 mg/kg/min is not arbitrary. It is not, like some of the older height and weight requirements, so loosely related to the job requirements as to be suspect.

The standards that are most easily defended are those that are clearly based on the purpose to be served by the decision. Since licensure is intended to protect the public from incompetent practitioners, it would seem that the performance standard for licensure examinations should be directly related to practice requirements. The performance standard should represent the level of performance needed to ensure safety and effectiveness in practice.

USING STANDARDS OF PRACTICE TO ANCHOR PERFORMANCE STANDARDS FOR LICENSURE EXAMINATIONS

As indicated above, the standards for licensure examinations are most defensible if they are linked to the requirements of practice and to the risks associated with a failure of practitioners to meet these requirements. Licensure decisions are most easily justified when they are explicitly linked to public safety.

By definition, professions are expected to define the expectations for adequate performance in practice. The standards of practice involve issues of

both ethics and competence. The ethical requirements tend to be stated explicitly and often in some detail in codes of ethics. The standards for judging competence in practice are not generally written in any one place, and in many cases, they are implicit. While they are constantly evolving and rest mainly on the shared values and understanding of the members of the profession, they provide a good, general benchmark for adequate practice. I will refer to these criteria for identifying acceptable levels of practice as the *standards of practice*.

The idea that performance standards be tied to accepted standards of practice is not new. The Uniform Guidelines, developed more than twenty years ago, require a rationale for the cutscores in employment testing:

The overriding consideration is that this score be consistent with “acceptable proficiency within the workforce.” (Equal Employment Opportunity Commission 1978, 38298)

Licensure decisions differ from employment decisions in a number of important ways, but the expectation that the passing score be tied to the requirements of practice seems as reasonable for licensure examinations as the corresponding expectation for employment tests.

Pyburn’s analysis of legal challenges to licensure examinations discusses the need for a “rational relationship” between requirements for licensure and practice requirements. Pyburn (1990) quotes the 1957 decision of the U.S. Supreme Court in *Schwartz v. Board of Bar Examiners* as follows:

A State cannot exclude a person from the practice of law or from any other occupation in a manner or for reasons that contravene

the Due Process or Equal Protection Clause of the Fourteenth Amendment. . . . A State can require high standards of qualification . . . but any qualification must have a rational connection with the applicant’s fitness or capacity to practice [a licensed occupation]. (p. 6)

The standard embodied in the passing score needs to be reasonable or “rational,” given the purpose of the decision process.

The standards of practice in the profession define acceptable performance, relative to practice requirements and risks. Therefore, by linking the performance standards to the standards of practice, the performance standards can be grounded in practice requirements and risks.

Because the standards of practice are intended to apply to performance in practice, they are most easily applied to such performances. Therefore, if standard setting is to be based on the evaluation of candidate performance relative to the standards of practice, a standard-setting approach that focuses on evaluations of candidate performance seems preferable to one that focuses on test questions per se.

The many standard-setting methods developed over the last 30 years can be divided into two broad categories, *test-centered methods* and *examinee-centered methods* (Jaeger 1989). In test-centered methods, participants review the items or tasks in the test and decide on the level of performance on these items or tasks required to meet a performance standard. For example, in the Angoff (1971) procedure, the participants are asked to imagine a typical minimally competent examinee; they then review the test items, one at a time, and decide on the probability, called a *minimum pass level*, or MPL, that the hypothetical

candidate would answer the item correctly. The passing score for the test is the sum of the MPLs for all of the items.³ The participants do not rate actual candidate performances in setting the standard (although in some modified Angoff procedures designed for extended responses, the participants may be shown samples of performances at different score levels as part of their training).

In the examinee-centered methods, actual performances are evaluated relative to the performance standard. For example, in the borderline-group method, the participants identify candidates who just meet the performance standard, and the passing score is set equal to the median score for these candidates. In 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, and the score that best discriminates between these two groups is taken as the passing score (Livingston and Zieky 1982).

Most standard-setting studies for licensure and certification examinations (particularly those relying exclusively on multiple-choice questions) have involved test-centered methods (usually the Angoff method) in which a group of panelists set the passing score by specifying how well a hypothetical minimally competent candidate would perform on each item. The performance standard is defined in terms of the panelists' expectations about what new practitioners should know and be able to do, but the standard being set is not explicitly tied to practice requirements. The panelists are asked to make judgments about how a minimally qualified candidate would perform on certain tasks, but the hypothetical candidate being invoked is an abstraction, and the task being evaluated is not a sample of practice.

An examinee-centered approach has strong advantages over a test-centered approach for setting practice-based standards. To the extent that the panelists 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 approaches involve just this kind of judgment. The panelists review actual candidate performances and decide whether they are acceptable or not. In the article that will appear in the next issue, examinee-centered methods for bar examinations will be discussed and illustrated.

CONCLUDING REMARKS


Given the way bar examinations are used to make decisions about bar admissions, it is necessary to have a specific passing score. The aim in setting that score is to provide adequate protection to the public while not subjecting candidates to arbitrary requirements, and therefore, the choice of a passing score on a bar examination is a matter of public policy. This policy is articulated in the performance standard and implemented through the passing score.

In this article, I have described a standard-setting model that links the passing score standards to current standards of practice, and therefore builds validity into the performance standard and its associated passing score by making them more relevant to the actual practice of law. Although the standards of practice may not be so clearly defined and may vary from setting to setting, the existing standards of performance for practice provide the best available basis for standard setting.

The methodology of standard setting has improved over the last twenty-five years, but standard setting is still not an exact science. Standard setting is basically a matter of policy, of deciding

how much is enough. Empirical standard-setting studies can provide useful information for policy makers, but they cannot provide simple answers to complex questions involving tradeoffs between public protection, the rights of candidates, and the availability and cost of professional services. As Brennan (2001) has put it:

[P]olicy makers need to understand that their prerogatives and authority come at the cost of being responsible for the inevitable value judgments that are part of standard setting. Measurement professionals cannot provide them with an impermeable shield that will defend them from critics who disagree with their value judgments. (p. 11)

Those with responsibility for licensure decisions should be involved in the design of standard-setting studies, and they must be involved in the interpretation of the results. 

ENDNOTES

1. For example, three sets of performance standards defining performance levels of basic, proficient, and advanced have been applied to the National Assessment of Educational Progress. In general terms, the proficient level is defined in terms of “solid academic performance.” Students at this level “have demonstrated competency over challenging subject matter and are well prepared for the next level of schooling.”

At a more detailed level, proficient performance in mathematics for eighth graders is described as follows (Shepard et al. 1993):

Eighth-grade students performing at the proficient level should apply mathematical concepts and procedures consistently to complex problems in the five NAEP content areas.

Eighth graders performing at the proficient level should be able to conjecture, defend their ideas, and give supporting examples. They should understand the connections between fractions, percents, decimals, and other mathematical topics such as algebra and functions. Students at this level are expected to have a thorough understanding of basic level arithmetic operations—an understanding sufficient for problem solving in practical situations.

Quantity and spatial relationships in problem solving and reasoning should be familiar to them, and they should be able to convey underlying reasoning skills beyond the level of arithmetic. They should be able to compare and contrast mathematical ideas and generate their own examples. These students should make inferences from data and graphs; apply properties of informal geometry; and accurately evaluate and communicate results within the domain of statistics and probability. (p. 35)

2. The term “arbitrary” is used in a number of ways in the literature on standard setting. According to Carson (2001), “A reading of the case law makes it clear that the courts are not using the term arbitrary to refer to the exercise of judgment in making choices among legitimate options. Rather, the term as used by the courts connotes unreasonable and capricious decision making” (p. 430).
3. Usually each panelist decides on an MPL for each item and the panelists’ MPLs are averaged to get the working MPL. Then the MPLs are added to get a passing raw score. (The alternative would be to have the panelists agree on an MPL for each item.)

For example, if the panelists’ MPLs (after averaging) on a four-item test were .5, .3, .4, and .8, the passing score would be 2 and a candidate would have to answer 2 items correctly to pass.

REFERENCES

- Angoff, W. 1971. Scales, norms, and equivalent scores. In *EDUCATIONAL MEASUREMENT*, 2nd ed. Edited by R. Thorndike. Washington, D.C.: American Council on Education.
- Bahls, S. 2001. Standard setting: The impact of higher standards on the quality of legal education. *The Bar Examiner* 70(4):15-17.
- Brennan, R. 2001. Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice* 20(4):6-18.
- Campion, M. 1983. Personnel selection for physically demanding jobs: Review and recommendations. *Personnel Psychology* 36:527-50.
- Carson, J.D. 2001. Legal issues in standard setting for licensure and certification. In *SETTING PERFORMANCE STANDARDS: CONCEPTS, METHODS AND PERSPECTIVES*. Edited by G. Cizek.
- Cascio, W., Alexander, R. and Barrett, G. 1988. Setting cutoff scores: Legal, psychometric and professional issues and guidelines. *Personnel Psychology* 41:1-24.
- Cimons, M. 1998. As obesity standard drops, dieters’ spirits may follow: Experts fear lower guidelines, which designate 97 million Americans overweight, may cause many to quit trying to get slim. *Los Angeles Times*, June 5, 16.
- Cronbach, L. 1980. Validity on parole: How can we go straight? In *NEW DIRECTIONS FOR TESTING AND MEASUREMENT: MEASURING*

ACHIEVEMENT. Edited by W.B. Schrader. San Francisco: Jossey-Bass Incorporated Publishers.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. 1978. Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. *Federal Register* 43:38290-38315.

Glass, G. 1978. Standards and criteria. *Journal of Educational Measurement* 15:237-61.

Greenberg, D. 1998. Of human poundage. *The Lancet*, June 11, 352(9122):158.

Hambleton, R. 1978. On the use of cut-off scores with criterion-replacement tests in instructional settings. *Journal of Educational Measurement* 15:277-90.

Jackson, A. 1994. Preemployment physical evaluation. *Exercise and Sport Science Review* 22:53-90.

Jaeger, R. 1989. Certification of student competence. In EDUCATIONAL MEASUREMENT, 3rd ed. Edited by R. Linn. New York: American Council on Education and Macmillan.

Kane, M. 1994. Validating the performance standards associated with passing scores. *Review of Educational Research* 64:425-61.

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, NJ: Lawrence Erlbaum.

Klein, S. 2001. Setting bar exam passing scores and standards. *The Bar Examiner* 70(4):12-15.

Linn, R. 2000. Assessments and accountability. *Educational Researcher* 29(2):4-16.

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

Merritt, D. 2001. Raising the bar: Limiting entry to the legal profession. *The Bar Examiner* 70(4):9-12.

Merritt, D., Hargens, L., and Reskin, B. 2001. Raising the bar: A social science critique of recent increases in passing scores on the bar exam. *Cincinnati Law Review* 69(3):929.

Norcini, J. and Shea, J. 1997. The credibility and comparability of

standards. *Applied Measurement in Education* 10(1):39-59.

Pyburn, K. 1990. Legal challenges to licensing examinations. *Educational Measurement, Issues and Practices* 9:5-6.

Popham, J. 1978. As always, provocative. *Journal of Educational Measurement* 15:297-300.

Shapiro, L. 1998. Fat, fatter: But who's counting? *Newsweek*, June 15, 131(24):55.

Shepard, L. 1980. Standard setting, issues and methods. *Applied Psychological Measurement* 4:447-67.

Shepard, L., Glaser, R., Linn, R. and Bohrnstedt, G. 1993. SETTING PERFORMANCE STANDARDS FOR STUDENT ACHIEVEMENT. Stanford, California: National Academy of Education, Stanford University.

Sothmann, M., Saupe, K., Jasenof, D., Blaney, J., Donahue-Fuhrman, S. and Woulfe, T. 1990. Advancing age and the cardiorespiratory stress of fire suppression: Determining a minimum standard for aerobic fitness. *Human Performance* 3:217-36.

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 NJ: Lawrence Erlbaum.



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