

# WHAT THE BAR EXAMINATION MUST ACHIEVE: THREE PERSPECTIVES

*by Michael T. Kane, Ph.D.*

**B**ar examinations are expected to meet accepted criteria for testing procedures, but they are also integral parts of the overall process of admitting candidates to the practice of law and therefore have to meet the requirements entailed by this use. In this article I will address three complementary perspectives on bar examinations: the measurement perspective, the decision-making perspective, and the candidate or test-taker perspective. The first two perspectives can be characterized as follows:

- In measurement theory, a candidate's test score is thought of as an estimate of that candidate's "true" score (the expected score over an infinite number of replications), and the emphasis is on the validity (or accuracy) and reliability (or precision) of this estimate.
- In evaluating examinations as tools for making licensure decisions, the emphasis tends to be on the appropriateness and fairness of the examination process, in addition to accuracy and precision.

In designing and evaluating bar examinations, it is important to consider the measurement perspective, the more pragmatic decision-making perspective, and to the extent possible, the candidate perspective.

We can imagine a conversation between a person with a strong measurement perspective (MP) and a person with a decision-making perspective (DP):

DP: You look perplexed. What's up?

MP: This candidate's score is only one point above the passing score, and given this small difference, he could easily have failed.

DP: Was some mistake made?

MP: There's no indication of any irregularity, but his score is so close to the passing score that if we repeated the testing, he might fail.

DP: Why should we repeat the testing? He passed and has no reason to test again! As a matter of fact, he is not allowed to take the test again.

MP: His score is above the passing score, but it is only an estimate of his "true" score over all possible replications.

DP: He took the test on this test date, and he passed. Concerns about what might have happened on a different test date are irrelevant.

MP: But the reliability is not perfect, and he would probably get a slightly different score if tested again.

DP: We want the reliability to be high, but the decision is based on his score on this test administration.

The point here is not that one perspective is better than the other. Rather, they are complementary, as I will explain.

## THE FUNCTION OF LICENSURE EXAMINATIONS

Boards of bar examiners and Courts are responsible for evaluating candidates for admission to the practice of law, and in meeting this responsibility, they generally rely on standardized tests of some kind as one way to evaluate competence. The tests function as measurement instruments, but they are also integral components in administrative decision-making procedures, and they have to fit into and support the overall goals of the program.

Bar examinations and other licensure examinations are intended to enhance the quality of professional practice and thereby to protect the public by identifying candidates who are not adequately prepared for entry-level practice; the exams are not intended to rank-order the candidates or to identify the best candidates. In particular, the examinations are designed to assess candidates' levels of competence in some target domain of knowledge, skills, and judgment (KSJs) that are considered important in entry-level practice. For licensure examinations, the focus tends to be on a fairly broad array of KSJs that would be relevant to the range of practice situations and responsibilities covered by the license. Even though most practitioners are likely to specialize to some extent even in their early years (and more so later), licensure examinations focus on the broad range of KSJs needed in entry-level practice and do not include advanced topics and esoteric content required for advanced, specialized practice.

Note that licensure examinations do not evaluate all of the attributes needed for effective practice over a career, or even in the first few years of a career. Traits like honesty, conscientiousness, and diligence are clearly essential for a candidate to practice effectively, as are adequate levels of social skills, physical health, and mental health. In assessing candidates

for the bar, some of these non-cognitive traits are evaluated in character and fitness evaluations, where the emphasis also is on identifying individuals with serious limitations rather than rank-ordering candidates or identifying the best candidates.

## THE THREE PERSPECTIVES

Each perspective from which licensure examination programs can be evaluated—the measurement perspective, the decision-making perspective, and the candidate perspective—reflects a legitimate set of goals, and an examination program has to respect all of these goals to be effective. It is necessary for the testing program to

- meet the requirements embedded in the measurement perspective in order to be considered technically adequate,
- meet the requirements for a sound decision-making procedure in order to be considered fair and effective, and
- provide a level playing field for candidates in order to be considered acceptable.

The requirements imposed by considering all three perspectives are not too burdensome, because there is a lot of overlap in the criteria valued by each perspective. The differences among the perspectives tend to be more a matter of emphasis given to certain criteria than they are about the criteria themselves.

### The Measurement Perspective

As I explained earlier, in measurement theory, a candidate's test score is thought of as an estimate of a "true" score for the candidate, and the emphasis is on the validity (or accuracy) and reliability (or precision) of this estimate. For measures of overall competence in some performance domain, the major concerns are the extent to which the test

adequately covers the domain (validity) and the extent to which the score would be stable for each candidate over possible replications of the testing procedure (reliability).

The measurement perspective assumes that the attribute being measured has a definite value for each test taker. We do not generally know what this value is, and in a sense, we can never know exactly what it is, but we can try to estimate this value as accurately as possible. Within the measurement perspective, the main concern is the accuracy of the test scores as estimates of the “true” value of the attribute being measured. For bar examinations and other licensure tests, the accuracy of the test scores is evaluated in two ways:

1. in terms of the plausibility of the claims that are based on the test scores (particularly the claim that the examination provides an accurate indication of overall mastery of the KSJ domain), and
2. by taking into account any external factors that might distort the results.

The conceptualization tends to be abstract and technical, although the underlying principles are simple and intuitive.

A second and supporting concern is the reliability (or precision) of the scores, which is evaluated in terms of the extent to which candidates’ scores would be likely to stay more or less the same if the examination were repeated at about the same time but using different questions (and in the cases of essay questions and performance tasks, with different graders). Reliability is considered a necessary (but not sufficient) condition for validity, because test scores that are not reasonably stable across different conditions (e.g., different graders or different

questions covering the same content domain) cannot be interpreted in any consistent way.

The measurement perspective assumes that it is at least conceptually possible to repeat the measurement over and over again on the same person without changing the person. This is obviously not the case in practice, but like assumptions about perfectly smooth surfaces in physics, it can be a useful assumption for some purposes.

### *Two Methods for Measuring Validity*

#### *Criterion-Based Validity Studies and Their Problems for Licensure Exams*

The notion of validity as agreement of scores with the “true” value of the attribute being measured may seem simple enough at first, but it can get complicated. Assuming that we don’t know the value of an attribute for a test taker, how can we determine how closely a test taker’s scores approximate her or his true score for the attribute? In some cases, we may have an alternate measure of the attribute that is thought to be very accurate, and in these cases, we can compare the scores on the test to those on this (presumably more accurate) criterion for some sample of test takers; if the test scores agree with those on the criterion measure, we have evidence for the validity of the exam scores as measures of the attribute. This is the approach used for employment tests, where test scores are compared to measures of performance on the job, and in college admissions tests, where test scores are related to performance in college (e.g., first-year GPA). When a reasonably good criterion is available, this criterion-based approach can be very effective, but in many contexts, it is difficult to identify or develop an acceptable criterion measure.

It is generally not possible to conduct adequate criterion-based validity studies for licensure exams

for three reasons. First, the practice of a profession like law covers a wide range of activities, clients, and settings, even in a single area of practice, and it would be very difficult (if not impossible) to evaluate performance consistently across these activities, clients, and settings in order to get an accurate evaluation of performance in practice. The general measures of professional success that are most readily available (awards, publications, income, etc.) do not focus on the protection of the public and therefore are not particularly relevant to the validation of licensure programs. The development of general measures of the quality of practice in terms of client outcomes is probably not possible for most professions.

Second, to the extent that a decent criterion measure could be developed, it would probably need to be context-specific (e.g., performance in handling criminal prosecution or defense). Ratings of performance that would be applicable to different areas of practice would need to be quite general and therefore judgmental and prone to concerns about bias of various kinds. To the extent that the evaluations of performance are limited to specific areas of practice, it would be necessary to study many separate areas of practice in order to get a comprehensive assessment of validity.

Third, and most fundamental, candidates who fail the examination are not allowed to practice as lawyers, and it is therefore not possible to evaluate the relationship between passing the bar examination and effective performance in practice, even if we could develop an adequate criterion measure of performance in practice.

#### *Content-Based Validity Studies:*

##### *An Effective Choice for Licensure Exams*

The *Standards for Educational and Psychological Testing*<sup>1</sup> suggest that *content*-based validity analyses, rather than *criterion*-based analyses, be used for credential-

ing examinations (i.e., those used for licensure and certification) mainly because “criterion measures are generally not available for those who are not granted a license” and as a result, “[v]alidation of credentialing tests depends mainly on content-related evidence, often in the form of judgments that the test adequately represents the content domain of the occupation or specialty being considered.”<sup>2</sup>

This approach is consistent with the intended purpose of licensure exams, to protect the public by excluding from practice candidates who lack a level of competence in the specified KSJ domain needed for safe and effective practice. According to the *Standards*, “Tests used in credentialing are designed to determine whether the essential knowledge and skills of a specified domain have been mastered by the candidate. The focus of performance standards is on levels of knowledge and performance necessary for safe and appropriate practice.”<sup>3</sup> That is, the goal is to get a good indication of each candidate’s level of competence in the KSJ domain by asking the candidate to respond to a set of questions or tasks that are representative of the KSJ domain. The sample of questions or tasks should cover the domain pretty well and should be large enough to yield reliable results (i.e., results that will not vary much over an independent sample of the same size and representativeness).

#### *The Use of Reliability Analyses to Address Variations in Performance*

Reliability analyses evaluate the stability of the scores over replications of the testing procedure and thereby support the proposed interpretation of competence in the KSJ domain by indicating that candidates’ scores are not affected much by the particular sample of questions/tasks or the particular graders who rate the responses. Score variations across

replications of the testing program are referred to as “errors of measurement” or “random errors.”

In measurement theory, each test taker’s observed score on a particular testing date is assumed to equal the “true” value of the attribute for the test taker, plus an unknown component (the error of measurement) that reflects specific aspects of the testing situation and that varies randomly from one score to another. These “errors” are not associated with any mistakes or violations of procedure; rather, they reflect variations in performance across samples of questions/tasks and natural fluctuations in the grading of responses. Evaluations of the reliability of licensure test scores tend to focus on two especially significant sources of variation in scores on written tests.

#### *Grader Variability*

The scores given to a particular performance on the written (essay and/or performance task) portion of the bar examination can vary from one grader to another, even if the graders are well trained and calibrated, and these variations can be a significant source of error in the essay or performance task scores. This source of variation can be controlled by training the graders thoroughly and monitoring grader performance. It is also highly desirable to include multiple graders for each candidate, in order to “average out” variations among graders that remain after training and calibration. This can be done efficiently by including multiple essay questions and/or performance tasks and having each essay answer or task performance evaluated by a different grader (or set of graders).

#### *Question Variability*

A second potentially important source of “random error” is the variability that occurs across questions because a candidate knows some subject areas better than others, misreads a question, or applies a principle incorrectly. It is essential that the questions be written as clearly and accurately as possible. It is also possible to control this source of error to a significant extent by including a large number of questions.

#### *How Standardization Supports Reliability and Validity*

Standardization of testing materials and procedures tends to enhance both reliability and validity. It improves reliability by eliminating various sources of irrelevant score variability that would occur if the testing materials, procedures, and conditions were allowed to vary from one candidate to another or from one administration to

another. By standardizing as many aspects of the examination program as possible, random fluctuations are reduced and reliability is increased.

Standardization also tends to improve validity by controlling the sources of random variability that are considered under the heading of reliability, and by controlling some additional potential sources of variability (test administration procedures, time limits, test site conditions) that are not usually addressed under the heading of reliability. In addition, standardization tends to promote both reliability and validity by making it possible for the standardized materials and procedures to be carefully designed so that they yield consistent results



while giving each candidate a good opportunity to demonstrate his or her level of competence in the KSJ domain.

For aspects of testing that cannot be standardized, statistical adjustments can sometimes be used to achieve some of the benefits associated with standardization. For example, in order to maintain the security of the examination and to keep test content current, it is generally necessary to change the questions from one administration to another. Nevertheless, it would be desirable from a measurement perspective to control possible changes in difficulty from one sample of questions/tasks to another by standardizing the questions/tasks. Equating procedures are designed to adjust for any differences in the statistical characteristics (in particular, the overall difficulty of the questions/tasks) from one administration of the examination to another.

### **The Decision-Making Perspective**

Decision makers generally take a more pragmatic view of testing. Public officials and other institutional decision makers who are charged with making decisions that impact people's lives generally operate under mandates that put a high value on objectivity and fairness.<sup>4</sup> They also typically operate under tight budgets. As a result, they try to make decisions quickly, efficiently, and fairly, and they want the process to be viewed as being efficient and fair. In high-stakes testing contexts like bar examinations and other licensure tests, the measurement perspective's concerns about validity and reliability are important, but the more general goal of making fair and defensible decisions about candidate competence is the central concern.

#### *Tests as Objective Tools*

From a decision-making perspective, the tests are tools that can be helpful in making reasonable and

fair decisions in a timely way. Bar admissions programs and other licensure programs employ well-defined, systematic procedures, mainly to promote fairness by eliminating various potential sources of bias that can arise in more face-to-face assessments and to achieve the relatively high level of reliability that is made possible by standardization. In addition, standardized examination structures and procedures tend to make the decision-making process more efficient, because large numbers of candidates can be assessed at the same time.

Theodore Porter, a professor in the UCLA Department of History who specializes in the history of science, has suggested that testing is a common approach to high-stakes decision making in the public arena, and that objectivity (defined in terms of not being subjective, personal, or capricious) is highly valued in decision making because it "provides an answer to a moral demand for impartiality and fairness."<sup>5</sup> Test scores and other quantitative measures tend to provide the ultimate in objectivity and, as a result, are commonly used in making licensure and certification decisions. Objectivity and standardization are seen as effective ways to promote fairness as well as validity and reliability:

Mechanical objectivity has been a favorite of positivist philosophers, and it has a powerful appeal to the wider public. It implies personal restraint. It means following the rules. Rules are a check on subjectivity: they should make it impossible for personal biases or preferences to affect the outcome of an investigation.<sup>6</sup>

The decision-making perspective values the technical qualities emphasized by the measurement perspective, because these qualities are expected to enhance the validity, reliability, and fairness of the decision. They are not ends in themselves.

### *A Focus on the Specific Test Score*

From the decision-making point of view, the candidate is evaluated in terms of his or her performance on a specific test administration, without considering any other variables (e.g., the candidate's grades in law school; the school attended; the candidate's appearance, race, or gender). The emphasis is not on how well a candidate might do on an infinite sequence of possible replications of the testing procedure, but on how well the candidate does on the current examination. The focus is on one specific score rather than on the expected score over a hypothetical sequence of possible replications.

The decision-making process is specified in some detail, and much of it is automatic; if a candidate gets a score at or above the passing score, she or he passes, and otherwise, the candidate fails and is not admitted to practice. The procedures are designed to be as explicit as possible, thereby minimizing the need for judgment, subjectivity, and possible bias of any kind.

### *Judgments Inherent in the Decision-Making Perspective*

However, judgment cannot be excluded from the overall decision-making process. Judgments are made at a more general level in designing the testing program and the decision rules before these procedures and rules are applied consistently to all candidates. Among the considerations are the specification of the KSJ domain to be assessed, the procedures for developing the tests based on the domain, the length of the tests, the scoring procedures, the passing score, and the procedures for appealing

decisions and requesting accommodations. Many of these issues can be informed by measurement theory, but they are all basically policy issues.

### *Specification of the KSJ Domain*

The design of the KSJ domain generally involves a series of compromises between a desire to include a broad array of KSJs needed for effective entry-level practice and the need to be able to test this KSJ domain reliably and validly within a reasonable period of time. The specification of the KSJ domain depends mainly on professional definitions of the scope of practice, but the experts' judgments can be supported by empirical surveys of the client problems that are likely to be encountered in entry-level practice and by evaluations of the kinds of KSJs needed to handle these problems.

THE DESIGN OF THE KSJ DOMAIN GENERALLY INVOLVES A SERIES OF COMPROMISES BETWEEN A DESIRE TO INCLUDE A BROAD ARRAY OF KSJS NEEDED FOR EFFECTIVE ENTRY-LEVEL PRACTICE AND THE NEED TO BE ABLE TO TEST THIS KSJ DOMAIN RELIABLY AND VALIDLY WITHIN A REASONABLE PERIOD OF TIME.

### *Test-Development Procedures*

The test-development procedures generally require that content specialists develop test questions that cover the KSJ domain adequately, with the guidelines for adequate coverage set forth in test specifications that indicate the mix of questions to be included in each administration of the examination. To keep things reasonably simple, the measurement models generally assume that the test questions are sampled from the domain, but in fact, in bar examinations and other licensure examinations, the questions have to be written and edited by scholars or practitioners; the questions don't exist until composed and edited. Like many theoretical assumptions, the sampling assumption is a fiction, but it works pretty well in estimating reliability.

### *Length of the Test*

The length of the test is also a compromise. According to measurement theory, reliability tends to get better as the test gets longer (i.e., has more questions). A longer test can also sample the KSJ domain more thoroughly and thereby enhance the case for the validity of the examination as a measure of overall competence in the domain. The decision-making perspective values these measurement characteristics and also values the thoroughness associated with a longer examination. On the downside, a longer test requires more candidate time and expense and also requires more time and money to develop the questions. The decision makers have to decide on the appropriate trade-off between cost and time on the one hand and reliability, validity, and coverage of the KSJ domain on the other.

### *Scoring Procedures*

The scoring procedures are also basically a policy issue, with some technical implications and limitations. For bar examinations involving an objective test component such as the Multistate Bar Examination and an essay and/or performance task component, the weight to be assigned to each component is an important policy issue. Analyses of the expected reliability and validity<sup>7</sup> suggest that in order to optimize these measurement characteristics, about 50% to 60% of the weight should be assigned to the objective component, but any weighting system that assigns at least 40% to the objective component works reasonably well.

The selection of a passing score is also primarily a policy issue. Measurement theory does offer some guidance on how to conduct empirical standard-setting studies, but the results of such studies serve mainly to provide those who make the policy decisions with potentially useful information.

The specification of procedures for appealing score-based decisions and for the granting of testing accommodations is guided by applicable laws and regulations, and beyond that by analyses of fairness and practicality. Testing accommodations are a particularly difficult issue from both the measurement perspective and the decision-making perspective, because both of these perspectives value standardization as a means of enhancing reliability, validity, and fairness. From both perspectives, accommodations that level the playing field are legitimate, but accommodations that yield an advantage to one group over another are not considered acceptable.

### **The Candidate Perspective**

A third perspective, that of the candidate, is more goal oriented. In high-stakes contexts, candidates tend to view tests as opportunities or as hurdles that they want to get through successfully and without too much trouble. The candidate perspective does not pay much attention to abstract notions of validity and reliability of the examination as a measurement procedure. It is more concerned about the fairness and appropriateness of the examination, having a level playing field, and having a reasonable chance of success. Paul Holland, a statistician who has worked at Educational Testing Service and as a professor at the University of California, Berkeley, refers to this point of view as the “contest perspective”:

The basic idea behind the measurement view is that a test *measures* something about the test taker. The basic idea behind the contest view is that a test that matters *is a contest* with winners and losers . . . . These two views can emphasize different goals. *Measurement view*: make a test “reliable” and “valid.” *Contest view*: make a test “fair” and “understandable.”<sup>8</sup>

Candidates are likely to engage in various test-preparation activities, including practice on item



types included in the test. This is a legitimate course of action for the candidates, for whom test results have important consequences. The candidates are encouraged to prepare themselves as well as they can, but they are not allowed to engage in any activities, like cheating, that would short-circuit the intended interpretation of the test scores as indications of overall competence in the KSJ domain. As in sports, the candidates have a right to prepare themselves for the examination using any means that are not ruled out.

## THE INTERSECTION OF THE THREE PERSPECTIVES


Bar examinations have to meet accepted standards for validity and reliability in order to be considered defensible, but they also have to meet the requirements for fair and reasonable decision making and ensure a level playing field for all candidates. If the test scores are not reasonably reliable (i.e., consistent, stable), and if they are not valid in the sense of adequately sampling a KSJ domain that is clearly relevant to the decision to be made, they will not justify the claims based on them and therefore will not be seen as providing a reasonable basis for bar admissions decisions.

As a requirement for admission to the bar, the testing program has to function effectively as a fair and objective process in order to be implemented and accepted. All of the elements in the decision-making process have to work together to yield fair and reasonable decisions. The decision-making perspective gives a lot of attention to due process and fairness and therefore provides a useful counterbalance to the measurement perspective, which focuses mainly on accuracy and precision. If critics can make the case that the testing program does not provide a level playing field, the program is not likely to survive.

It is also important that test takers see the measurement procedures and the decision rules as giving them a fair chance to succeed. Test takers who see the process as fair and reasonable are more likely to prepare for the examination in constructive ways and are less likely to try to cut corners by engaging in questionable test-preparation or test-taking activities.

From the decision-making perspective, tests are tools that can help make the decision-making process more fair and effective than it would otherwise be. Tests must be reliable and valid enough to achieve that purpose, but these criteria for good measurement are not of primary concern. The examination is expected to be clearly relevant to the purpose at hand and reasonably reliable, and testing procedures should be transparent and fair.

Measurement theory tends to be abstract and hypothetical, relying on notions like statistical sampling from a domain, replications of the sampling design, and the expected values (and variances) of a candidate's scores over hypothetical replications. The statistical models are quite useful in sorting out the complex choices that need to be made in designing a testing program, but their results have to mesh with the need to make fair and reasonable decisions, and must do so transparently.

Again, the point is not that one perspective is better than the others but that they are complementary. The measurement perspective is especially useful in designing testing programs that will yield reliable and valid information about candidate competence, and it can also be helpful in designing the decision procedures, but it does not provide answers to the wide range of policy questions associated with high-stakes testing programs. Deciding how the score-based information will be used in the operational context of decision making requires a broader and more pragmatic perspective. 

## NOTES

1. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, AND NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING (American Educational Research Association 1999).
2. *Id.* at 157.
3. AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, AND NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION, *supra* note 1, at 156.
4. THEODORE M. PORTER, TRUST IN NUMBERS: THE PURSUIT OF OBJECTIVITY IN SCIENCE AND PUBLIC LIFE (Princeton University Press 1995).
5. *Id.*, at 8.
6. PORTER, *supra* note 4, at 4.
7. Michael Kane and Susan M. Case, *The Reliability and Validity of Weighted Composite Scores*, 17 APPLIED MEASUREMENT IN EDUCATION 221–240 (2004).
8. P.W. Holland, *Measurements or Contests? Comment on Zwick, Bond, and Allen/Donogue*, in PROCEEDINGS OF THE SOCIAL STATISTICS SECTION OF THE AMERICAN STATISTICAL ASSOCIATION 27–29 (American Statistical Association 1994).



MICHAEL T. KANE, PH.D., is the holder of the Samuel J. Messick Chair in Test Validity at the Educational Testing Service in Princeton, New Jersey. He was Director of Research for the National Conference of Bar Examiners from 2001 to August 2009. From 1991 to 2001, he was a professor in the School of Education at the University of Wisconsin–Madison, where he taught measurement theory and practice. Before his appointment at Wisconsin, Kane was a senior research scientist at ACT, where he supervised large-scale validity studies of licensure examinations. Kane holds an M.S. in statistics and a Ph.D. in education from Stanford University.