Reflections on Bar Examining

by Michael T. Kane, Ph.D.

This is a bittersweet time for me. Working on bar examining as Director of Research at NCBE for the past eight years has been very interesting and enjoyable, and working with all of my friends at NCBE and in the larger bar examining community has been a great experience. I have learned a lot and have had the opportunity to work with many kind and dedicated people. I had intended to retire soon, perhaps gradually, but those plans have changed. Educational Testing Service in Princeton, New Jersey, has asked me to be the first holder of a new position, the Samuel J. Messick Chair in Test Validity, named after Samuel Messick, who led work on validity theory at ETS until his death in 1998. Rather than retiring, I am going off on a new adventure.

The Bar Examiner serves as a forum for the discussion of a wide range of issues in bar admissions, and I have been asked to share some of the lessons I have learned in my work on bar examinations. I will begin with some general comments on licensure examinations, their purpose, and how to evaluate them, particularly with regard to their validity and reliability. I will then review some principles for the design and implementation of bar examinations and how those principles have been used to promote validity and reliability. It never hurts to revisit the basics. Finally, I will briefly consider one main alternative to the use of licensure examinations—the observation of candidate performance in real practice situations.

The Purpose of Licensure Requirements

Licensure requirements are designed to protect the public by ensuring that candidates who are admitted to practice in a profession have met certain basic qualifications.[1] For most professions and in most jurisdictions, the requirements for admission to practice (i.e., the requirements for licensure) include successful completion of an appropriate educational program, the passing of one or more licensure examinations, and some demonstration of good moral character and fitness to practice.

Licensure is not intended to provide a guarantee of excellent performance, nor does it claim to predict how well candidates will perform if admitted to practice. Rather, it certifies that new practitioners have met the basic requirements that are designed to provide the public with some assurance that they are qualified to practice.

Character and fitness evaluations tend to focus on whether a candidate has engaged in any activity (e.g., committed a felony, lied about a significant matter) that would indicate a lack of integrity, or has a problem (e.g., a history of substance abuse)

that might interfere with his or her effectiveness in practice. The procedures used to evaluate character and fitness are not designed to ensure that candidates have stellar characters, and the results of these evaluations are not used to predict future behavior. Rather, the procedures are designed to identify candidates whose past performance indicates that they might present a risk to the public if they were admitted to practice. In the context of licensure, character and fitness evaluations are less concerned with identifying the best candidates and more concerned with weeding out serious risks.

Similarly, the educational and testing requirements are designed to provide assurance that new practitioners have a broad base of knowledge, skills, and judgment (KSJs) relevant to professional practice. The academic requirements do not seek to identify the most accomplished students or to predict future performance in practice, but to ensure that those admitted to practice have achieved a reasonable level of competence in applying professional skills to commonly encountered practice problems. Again, the focus is on protecting the public—in this case, by excluding candidates who lack the KSJs needed in practice to an extent that would pose a risk to clients.

[T]HE EDUCATIONAL AND TESTING REQUIREMENTS ARE DESIGNED TO PROVIDE ASSURANCE THAT NEW PRACTITIONERS HAVE A BROAD BASE OF KNOWLEDGE, SKILLS, AND JUDGMENT (KSJS) RELEVANT TO PROFESSIONAL PRACTICE. THE ACADEMIC REQUIREMENTS DO NOT SEEK TO IDENTIFY THE MOST ACCOMPLISHED STUDENTS OR TO PREDICT FUTURE PERFORMANCE IN PRACTICE, BUT TO ENSURE THAT THOSE ADMITTED TO PRACTICE HAVE ACHIEVED A REASONABLE LEVEL OF COMPETENCE IN APPLYING PROFESSIONAL SKILLS TO COMMONLY ENCOUNTERED PRACTICE PROBLEMS.

Licensure tests generally provide standardized, objective evaluations of the cognitive skills involved in applying professional principles to practice situations. According to the STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING,[2]

> tests used in credentialing are intended to provide the public, including employers and government agencies, with a dependable mechanism for identifying practitioners who have met particular standards. The standards are strict, but not so stringent as to unduly restrain the right of qualified individuals to offer their services to the public.[3]

By ensuring that candidates admitted to practice have achieved a reasonable level of competence in applying professional skills to commonly encountered practice problems, the bar examination requirement and educational requirements are expected to control one threat to the overall quality of practice—the threat posed by practitioners who lack basic competencies. The character and fitness requirement seeks to control other threats to the quality of practice, but none of these sources of information about candidates is expected to provide predictions of future performance in practice.

EVALUATING TESTING PROGRAMS

The technical quality of testing programs is evaluated mainly in terms of two general criteria: *validity* and *reliability*.

Validity

In developing and evaluating any testing program, validity is the primary concern.[4] According to the STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING, "validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests."[5] Basically, validity analyses address the question of whether proposed interpretations and uses of the test scores make sense and are justified.

The validity of a proposed test score interpretation and use depends on the plausibility of the proposed interpretation and the reasonableness of the decisions based on the scores, given all of the available evidence. A proposed interpretation and use of test scores is considered valid if a strong case can be made for the claims incorporated in it. Assuming that the test scores are used to assess current competence in some area of activity, as is the case for licensure exams, the validity analyses focus on the appropriateness of the testing materials and procedures as measures of competence in the area of activity to which the license applies.

Any interpretation and use of test scores is likely to involve a number of inferences and supporting assumptions, and if one or more of the inferences or assumptions is questionable, the validity of the proposed interpretation and use is questionable. This conception of validity implies that it is a good idea to keep the interpretations of test scores as simple as they can be, given their intended uses. If the interpretation is ambitious, involving many claims or involving particularly strong claims, it will tend to be hard to validate, but if the claims are kept modest, the need for evidence to support the claims also tends to be modest. It is therefore not advisable to add unnecessary claims to test score interpretations.

ANY INTERPRETATION AND USE OF TEST SCORES IS LIKELY TO INVOLVE A NUMBER OF INFERENCES AND SUPPORTING ASSUMPTIONS, AND IF ONE OR MORE OF THE INFERENCES OR ASSUMPTIONS IS QUESTIONABLE, THE VALIDITY OF THE PROPOSED INTERPRETATION AND USE IS QUESTIONABLE. THIS CONCEPTION OF VALIDITY IMPLIES THAT IT IS A GOOD IDEA TO KEEP THE INTERPRETATIONS OF TEST SCORES AS SIMPLE AS THEY CAN BE, GIVEN THEIR INTENDED USES.

By analogy with employment testing, it is sometimes suggested that scores on licensure examinations should be evaluated in terms of how well they predict future performance in practice. This model tends to be unrealistic for several reasons. First, it is difficult to define clear criteria for success in professional practice, which includes a wide range of professional activities over a wide range of contexts; for employment tests used for specific jobs, in contrast, it can be relatively easy to define productivity criteria. Second, even if we can agree on a definition of success in practice, it tends to be very difficult to develop fair, standardized measures of success; most of the readily available sources of data (e.g., disciplinary actions, financial success) depend more on character and fitness or personal characteristics than on the kinds of professional competence measured on licensure examinations. Third, candidates who fail a licensure examination do not have the opportunity to practice, and therefore it is not possible to collect the most crucial predictive data (the difference in performance between passing and failing candidates).

For these and other reasons, it is generally suggested that licensure exam scores be interpreted and validated mainly in terms of "judgments that the test adequately represents the content domain of the occupation or specialty being considered."[6]

Reliability

The reliability of test scores is defined in terms of their consistency (or dependability, or reproducibility) over repeated measurements. Reliability addresses the question of whether we would get approximately the same scores or make the same decisions if we repeated the assessment process in more or less the same way. If the test scores fluctuate widely from one set of questions to another, or from one grader to another, or from one day to the next, it is hard to interpret the scores for any particular test administration as an indication of a candidate's qualifications. So, for example, with the MBE measuring the ability to apply basic legal principles to realistic fact situations, we expect candidates' scores on the morning section of the MBE to be strongly related to (or correlated with) their scores on the afternoon section, and this is found to be the case. Similarly, we expect candidates' scores on one set of essay questions scored by a particular group of graders to be positively correlated with their scores on another set of essay questions scored by a different group of graders, and this is also found to be the case.

We have some standard statistical indices for the reliability of test scores. The most basic of these is the *standard error of measurement*, or the standard error, which provides an indication of how much variability we would expect to see in a candidate's scores if he or she took comparable forms of the test at about the same time.[7] Using reasonable statistical assumptions, standard errors of measurement can be estimated using data from regular test administrations, and estimates of standard errors provide useful indications of the consistency of the test scores over repeated measurements. Standard errors are always greater than zero, and it is desirable that they be small. We don't want the results to bounce around just because we change some of the conditions of observation.

A *reliability coefficient* can be defined in terms of the average magnitude of the standard error relative to the variability across individuals. Reliability coefficients have values between 0.0 and 1.0, with a reliability of 1.0 indicating perfect agreement between scores on the two forms of the test (or two sets of graders), and a reliability of 0.0 indicating a complete lack of agreement between them. The reliability is large if the standard error of measurement is small compared to the score differences among the candidates taking the test, and it is small if the standard error of measurement is large compared to the score differences. Therefore, the reliability coefficient provides a good indication of the consistency with which the testing program rank-orders the candidates taking the test.

A fairly high reliability (above 0.8; preferably above 0.9) is expected for testing programs that are used to make high-stakes decisions about individuals. Decisions are considered to be "high stakes" if they have important consequences and cannot be changed easily or quickly. Licensure examinations (including bar examinations) are high-stakes tests, because they have serious consequences.[8]

Reliability is a necessary but not sufficient requirement for validity. If the test scores fluctuate widely for individuals (i.e., if the scores are unreliable), it will not be possible to give the scores any coherent interpretation. However, consistent, reliable test scores do not necessarily support the proposed

interpretation; a test can provide consistent, reliable scores but not be appropriate for the intended interpretation and use. A stopped watch is very consistent, but it gives the right time only twice a day.

An adequate level of reliability is important for any testing program, but for licensure programs and many other high-stakes testing programs, a related concern, *decision consistency*, is as important or more important. A testing program is said to have high decision consistency if the decisions based on the test scores are consistent over repeated applications of the testing procedure. A testing program is said to have high decision consistency if the candidates who passed on one test administration would also have passed if we had given them an alternate but similar test, and the candidates who failed on one test administration would also have failed if we had given them an alternate but similar test. Decision consistency depends on the reliability of the tests being used to make the decision but, as discussed later, it also depends on how the test scores are combined to make the decision.

AN ADEQUATE LEVEL OF RELIABILITY IS IMPORTANT FOR ANY TESTING PROGRAM, BUT FOR LICENSURE PROGRAMS AND MANY OTHER HIGH-STAKES TESTING PROGRAMS, A RELATED CONCERN, *DECISION CONSISTENCY*, IS AS IMPORTANT OR MORE IMPORTANT. A TESTING PROGRAM IS SAID TO HAVE HIGH DECISION CONSISTENCY IF THE DECISIONS BASED ON THE TEST SCORES ARE CONSISTENT OVER REPEATED APPLICATIONS OF THE TESTING PROCEDURE.

Evaluating the Validity of Bar Examinations

Licensure tests provide a check on one aspect of readiness for effective performance in practice, and therefore they serve an important but limited purpose. They measure a set of cognitive competencies that are needed in practice, but they do not attempt to measure all of the characteristics required for good practice and therefore are not expected to provide predictions of a candidate's future performance in practice. According to the STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING,

> 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. . . .[9]

Professional expertise is used to identify competencies that are critical for effective performance in practice, and the test questions and tasks are developed to assess these competencies. It is reasonable to require that candidates for licensure demonstrate mastery of these competencies, or KSJs, before being licensed to practice.

As is true for any licensure examination, the validity of the scores on a bar examination depends on the plausibility of the proposed interpretation and use of the scores, as defined by the chain of inferences leading from a candidate's test score to conclusions about the candidate's readiness for entry-level practice; and the plausibility of the interpretation and use depends on the evidence for and against the claims being made. It is therefore important to be clear about the proposed interpretation and use and, in particular, about the claims being made.

The knowledge base of the profession includes KSJs that are critical in the sense that serious deficiencies in a candidate's mastery of these competencies would make it difficult for the candidate to practice effectively. Mastery of the competencies does not ensure success in practice, but a lack of adequate mastery of the competencies would interfere with effective performance in practice. For example, knowing the law of contracts does not, in itself, make for an effective practitioner, but ignorance of these laws on the part of lawyers would clearly put their clients at risk.

The Bar Examination as a Measure of Competence

Following this approach, scores on bar examinations would be interpreted as measures of level of competence in a domain of KSJs generally required for effective legal practice. There are a number of points packed into this statement that merit some discussion.

First, this interpretation defines professional competence in terms of the extent to which the candidate has the legal knowledge and skills needed to solve commonly occurring legal problems. Clients need professional help in solving their legal problems, and lawyers are expected to have the professional knowledge, skills, and judgment to help clients deal with their legal problems. Candidates who lack basic legal knowledge and skills are not considered ready for practice.

Second, a candidate's score is interpreted as a measure of the candidate's current level of achievement on the KSJs. The interpretation does not purport to provide predictions of individual candidates' *future* levels of performance in practice. It would be great if we had measures that would provide accurate predictions of performance over the course of a candidate's career in legal practice, or even of the expected performance over the first few years of practice, but this is not a realistic scenario.

Licensure tests are relevant to effectiveness in practice in that they assess competencies that are needed for effective performance in commonly occurring practice situations. They do not necessarily ensure that passing candidates will do well in practice or that candidates with higher scores will do better than candidates with lower scores. Rather, candidates with low scores (particularly scores below the passing score) will have substantial difficulties performing adequately in practice because they lack KSJs that are needed in practice. That is, competency in the KSJ domain is necessary but not sufficient for effective performance in practice.

All of the bar exam components (MBE questions, essay questions, performance tasks) are designed to measure competence in identifying the legal issues in fact situations and in applying general legal principles to these situations. It is not necessarily the case that candidates with adequate, or even very high, levels of competence on the KSJs will be effective in practice, particularly if they are deficient in other characteristics required in practice (e.g., interpersonal skills, conscientiousness, honesty). However, it is anticipated that individuals who have not achieved a reasonable level of competence in the KSJs would have great difficulty functioning effectively in many practice situations even if they have the other characteristics needed in practice.

Basic Principles for Ensuring Validity

Some general principles for maintaining and enhancing the validity of bar examinations can be derived from this competency-based approach to validation. These principles may seem obvious, and they are

already being applied to bar examinations, but it is useful to review basic principles from time to time.

1. The competencies being tested should be clearly related to entry-level practice.

As noted earlier, licensure is designed to protect the public by ensuring that admitted practitioners have met certain basic requirements. According to the STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING, licensure tests are designed to determine whether essential knowledge and skills have been mastered by the candidate. To achieve this, "[p]anels of respected experts in the field often work in collaboration with qualified specialists in testing to define test specifications, including the knowledge and skills needed for safe, effective performance, and an appropriate way of assessing that performance."[10] The content covered by the test is expected to focus on core knowledge and skills that are widely applicable in the practice of the profession and that are often critical to effective performance in practice.

DEFINING THE CONTENT OF A LICENSURE EXAMINATION INVOLVES A NUMBER OF TRADE-OFFS. THE TIME AVAILABLE FOR TESTING IS ALWAYS LIMITED, AND THEREFORE THE NEED FOR BREADTH OF COVERAGE IS IN CONFLICT WITH THE DESIRABILITY OF THOROUGHNESS IN EVALUATING KNOWLEDGE AND SKILLS IN VARIOUS AREAS. FOR LICENSURE TESTS, THE EMPHASIS TENDS TO BE ON BREADTH, WITH A FEW QUESTIONS ON EACH OF A WIDE RANGE OF CONTENT AREAS.

So, the test content should emphasize general principles and widely applicable skills, and it should avoid delving deeply into specialized and/or esoteric areas of content. More advanced topics are not necessarily to be avoided completely, because all practitioners (especially entry-level generalists) need to be able to recognize issues that could cause problems if not dealt with appropriately. New practitioners are not necessarily expected to be able to deal with all of the issues that might arise in practice situations, but they need to know enough to recognize the issues and take appropriate action (e.g., researching the issue more fully or referring the client to a colleague with expertise in the area).

Defining the content of a licensure examination involves a number of trade-offs. The time available for testing is always limited, and therefore the need for breadth of coverage is in conflict with the desirability of thoroughness in evaluating knowledge and skills in various areas. For licensure tests, the emphasis tends to be on breadth, with a few questions on each of a wide range of content areas. Licensure covers a wide range of practice areas, and licensure tests are designed to cover the knowledge and skills needed for entry-level practice in the range of practice areas covered by the license.

The validation of licensure examinations "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."[11] The test should cover the specified domain but emphasize those areas that are most critical for safe and effective performance in practice.

2. The test should cover as wide a range of critical competencies as possible.

It is not necessary that licensure tests cover all of the KSJs relevant to the practice of the profession; this would be impossible. Some personal qualities that are not amenable to testing (e.g., integrity,

conscientiousness) are evaluated in character and fitness assessments, and many skills that cannot be adequately assessed through standardized testing (e.g., those involved in extended performances) are evaluated in law school. So, bar exams tend to focus on the cognitive skills involved in applying legal principles to practice situations.

However, if the exam is to be useful in identifying candidates who have the competencies expected in entry-level practice and in differentiating them from candidates with serious gaps in their KSJs, it is important that the test cover a substantial subset of the KSJs required in practice. In order to achieve broad coverage of the KSJs required in practice, most bar examinations make use of several different testing formats.

The MBE provides broad coverage of the domain by including a large number of multiple-choice items, each of which requires candidates to apply a legal principle to a briefly stated fact situation. The MBE provides a good overall assessment of candidates' ability to identify appropriate conclusions and actions in a wide range of situations and six content areas: constitutional law, contracts, criminal law and procedure, evidence, real property, and torts.

Essay questions require candidates to analyze more complex fact situations in greater detail and at more length. In addition, candidates are required to present their analyses in an organized and coherent way.

Performance tasks, which are now included in most bar examinations, extend the range of skills being assessed by requiring candidates to carry out a realistic simulated legal task, which does not require much specific knowledge but does require a high level of analysis.

3. The test questions should focus on the ability to apply KSJs to practice situations.

As noted above, licensure examinations are intended to provide assurance that admitted candidates have the knowledge and skills needed for safe and effective performance in general, entry-level practice. The content covered by a licensure examination is designed to be critical for effective performance in entry-level practice, and successful candidates are expected to be able to apply their knowledge and skills to practice situations.

The most direct way to evaluate candidates' ability to apply critical knowledge and skills to practice situations is to ask them to do so. So, the questions included in licensure examinations should typically involve the description of a realistic practice situation followed by a question about what to do to resolve some problem posed by the situation (e.g., whether to object to a question) or about the implications of certain aspects of the situation (e.g., whether certain evidence is admissible in a case).

Questions that call for simple recall may be considered relevant to practice, but they are clearly less relevant than questions that require candidates to apply their knowledge and skills to a particular situation in a way that solves a realistic problem.

4. The question formats should be as simple and straightforward as possible, given the competencies being tested.

For a licensure test to be considered valid, it should provide an evaluation of each candidate's command of the knowledge and skills required in practice, and the candidate's scores should reflect his or her level of competency. To the extent that the test scores reflect any other candidate characteristics (e.g., race, gender) or any competencies other than

those included in the test content domain (e.g., familiarity with complex item types), the validity of the test can be questioned.

So, the format of the test and the format of individual questions should be as simple and straightforward as possible, given the competencies to be measured. There should be no tricks or excessively fine distinctions (e.g., a statute of limitations that is missed by one day because it's a leap year). The situations included in the questions should generally be as common and familiar as possible and should be described clearly and succinctly. Long, complicated stories with lots of extraneous facts are to be avoided, although some irrelevant facts may be necessary (e.g., if the point of the question is to sort the relevant from the irrelevant). Rhetorical flourishes and amusing names should be excluded. Basically, candidates who have the KSJs being evaluated should be able to answer the questions, and candidates who lack some or all of these KSJs should not be able to answer the questions.

5. For essay questions and performance tasks, the scoring rules should be specified in advance and should focus on the competencies being assessed.

For any test, the validity of a proposed interpretation of the test scores in terms of certain KSJs clearly depends on whether the questions focus on these KSJs. For essay questions and performance tasks, the scoring rules prepared for each question/task and the training of the graders are equally important. If a candidate's score on the test is to be interpreted in terms of how well the candidate can apply general principles to specific fact situations, the scoring rules and procedures should focus on how well the candidate applied the relevant principles to the situation, and not on any other factor (e.g., ability to state the law or principles in an area, or legible handwriting).

One issue that comes up in the scoring of responses to essay questions and performance tasks is the extent to which the quality of the candidate's writing should play a role in the scoring. There is no right answer to this question, but given that the essay questions and performance tasks are designed to evaluate the candidate's ability to analyze an extended set of facts and to provide a coherent analysis of these facts, the logical structure of a candidate's analysis, as reflected in his or her response, is important in evaluating the quality of the answer. On the other hand, given that the candidates are responding under fairly tight time constraints, it is probably not useful to focus on errors in grammar, spelling, or punctuation. At best, the answers are rough first drafts rather than polished samples of prose. A case can be made for paying more or less attention to writing per se, but the extent to which the quality of a candidate's writing is to be considered in scoring should be decided up front, and the graders should be trained to be consistent in the attention they give to this criterion.

6. The passing standard should be high enough to protect the public but not so high as to exclude candidates who could practice effectively.

In general, bar examinations emphasize the ability to apply legal principles to practice situations, and these cognitive skills are important for practice. However, the standards for these competencies should not be higher than the level of ability required for entry-level practice. Although some level of mastery of the competencies included in the test is needed for effective practice, it is not necessarily true that higher levels of mastery will lead to improved performance. The standard should be high enough to provide reasonable protection to the public but not so high as to exclude candidates who are prepared to practice effectively.

Evaluating the Reliability of Bar Examinations

Bar examinations tend to have relatively high reliability, because the components included in most bar exams have high reliabilities, and the component scores are combined in an appropriate way into a single score that is used to make pass/fail decisions. Most bar examinations consist of the MBE and a written component involving essay questions and possibly performance tasks. Some bar examinations involve additional components, but for the sake of simplicity, I will focus on the MBE and the written component.

The MBE has a reliability of about 0.9. The reliability of the MBE varies a little from administration to administration (from about 0.89 to 0.91) but is consistently high enough to meet the reliability requirement by itself. The reliability of the written component is generally lower and more variable than the reliability of the MBE. Assuming that the written component includes 6 to 10 tasks (including essay questions and performance tasks), that the candidate responses to each essay question and/or performance task are graded by a single grader (or a set of calibrated graders who have been trained to apply the scoring rules consistently), and that the overall written component score is the sum or average of the scores on the individual tasks, the reliability will tend to be about 0.7. So, the written components of most bar examinations are not reliable enough in themselves to meet the rule of thumb of 0.8 or 0.9, but when combined appropriately with the MBE, the overall score tends to have a reliability higher than 0.9.

IN MAKING PASS/FAIL DECISIONS, SEPARATE TEST COMPONENTS CAN BE COMBINED IN DIFFERENT WAYS. A PARTICULARLY SIMPLE WAY TO COMBINE THE SCORES IS TO SIMPLY ADD THEM TOGETHER OR AVERAGE THEM, AND THEN COMPARE THE RESULTING TOTAL SCORE TO A PASSING SCORE. THIS APPROACH LEADS TO A *COMPENSATORY SCORING RULE*. THE RULE IS COMPENSATORY BECAUSE A HIGH SCORE ON ONE COMPONENT CAN, TO SOME EXTENT, COMPENSATE FOR A LOWER SCORE ON ANOTHER COMPONENT.

Compensatory and Noncompensatory Scoring Rules

In making pass/fail decisions, separate test components can be combined in different ways. A particularly simple way to combine the scores is to simply add them together or average them, and then compare the resulting total score to a passing score. This approach leads to a *compensatory scoring rule*. The rule is compensatory because a high score on one component can, to some extent, compensate for a lower score on another component.

Alternately, a *noncompensatory scoring rule* requires that a candidate pass each of the separate test components in order to pass the test as a whole, and the candidate will fail if he or she fails on any single component. A compensatory rule has one hurdle, but a noncompensatory rule has several hurdles, all of which must be passed in order to pass the test as a whole. In general, compensatory rules tend to have much higher decision consistency than noncompensatory rules because, basically, a noncompensatory decision procedure is only as consistent as is its least reliable component.

If we adopt a compensatory scoring rule and combine the scores on the MBE (with a reliability of

about 0.9) and the written component (with a reliability of about 0.7) by taking their sum or average for each candidate (and weight the MBE at 50 percent or a bit higher), the reliability will tend to be in the low 0.90s, which is above the high benchmark of 0.90 and well above the more lenient benchmark of 0.80. That is, the total test score obtained by summing or averaging the MBE score and the written component score is more reliable than either score separately, but the reliability of the MBE score alone is almost as high as the reliability of the total score (i.e., the MBE score plus the written component score). The essay questions and performance tasks contribute more to validity than they do to the reliability of bar exams.

There are many alternatives to this compensatory approach to combining scores, but most of the noncompensatory approaches tend to have poorer decision consistency. For a noncompensatory rule to work well, each of the components has to have a relatively high reliability. For example, if candidates have to pass both the MBE and the written component in order to pass the exam as a whole, each of the components would be expected to have a reliability above 0.8 or 0.9, and written components do not generally achieve this level of reliability.

Ultimately, it is not the reliability per se that is important, but rather the dependability of the pass/fail decisions that are made using the test scores, and noncompensatory decision rules that involve multiple hurdles tend to be more error-prone than compensatory rules, because a false negative (failing a candidate who should have passed) on any component leads to a failure of the examination as a whole. With multiple hurdles, there are multiple opportunities to generate false negative decisions. Furthermore, the separate components tend to be less reliable than a single composite score because they are shorter than the total test of which they are a part, and the least reliable components can have a substantial impact on the overall decision consistency.

For complicated noncompensatory rules (e.g., requiring a candidate to get passing scores on particular sets of essay questions), it can be difficult to compute the reliability of the decision rules, but for the reasons discussed above, the decision consistency of complicated noncompensatory rules is likely to be relatively low. So, I would recommend that the scores on different parts of the test be combined into a single overall score, using a compensatory scoring rule, and that this single score be used to make the pass/fail decision.

Basic Principles for Ensuring Reliability

In order to achieve a high degree of reliability for bar exams that include the MBE and a written component, four principles are relevant:

1. The test should be long enough to provide a reliable score.

The reliability of scores across different forms of a licensure exam depends on the quality of the test questions and the number of questions. Assuming that the questions are well written, the reliability then depends mainly on the length of the test. The MBE, with 200 questions, provides a good base for the reliability of bar examinations.

The addition of essay questions and performance tasks to the MBE can do much to enhance the validity of the bar examination if the scores are combined appropriately, but it typically has a relatively modest impact on reliability.

2. Compensatory scoring rules, in which the MBE and written component scores are combined to yield

a single overall score that is the basis for decision making, are generally preferable to more complicated noncompensatory rules.

As discussed above, the reliability of the overall results of a bar examination including several components (MBE, essay questions, performance tasks) tends to be highest if a compensatory scoring rule is adopted. It may seem desirable to require separate passing scores on different test components, or even on specific essay questions or performance tasks, but this kind of noncompensatory rule tends to be highly unstable; a candidate who does very well on most of the test but goes off on a wrong tangent on a specific question—or, even worse, gets scored especially harshly on one question—may fail for that reason alone.

3. The scores of the more reliable components should be given fairly high weights.

In a compensatory scoring system, the reliability of the total score tends to be highest if the more reliable component scores are given greater weight than the less reliable component scores. For most bar exams, this general principle implies that the MBE should generally get a higher weight than the written component.

However, it is important not to sacrifice validity of the total score just to enhance reliability, and given that essay questions and performance tasks address competencies that are not covered by multiple-choice questions, validity would suffer if the MBE were given too much weight. A good balance of reliability and validity can be achieved by assigning a weight of about 0.5 to 0.6 to the MBE and a weight of 0.4 to 0.5 to the written component.

4. For essay questions and performance tasks, the scoring rules should be clearly defined, and the graders should be trained to be as consistent as possible in applying the rules.

For essay questions and performance tasks, the quality and consistency of the scoring is as important in determining the reliability (and validity) of the scores as is the quality of the questions. It is important that the scoring rules (and preferably the outline of a complete satisfactory answer) be specified in advance, and that the graders be trained to apply these rules fairly and consistently throughout the scoring process.

Performance Assessments: An Alternate Approach to Evaluating Readiness for Practice

On the face of it, various kinds of extended performance assessments, such as apprenticeships and internships, in which a candidate's actual performance in real-world contexts is evaluated, could provide particularly effective ways of assessing readiness for practice. A natural way to evaluate a candidate's readiness for an activity (e.g., legal practice) is to observe his or her performance of the activity (e.g., in a sample of practice situations).

However, there are two major disadvantages to this approach. First, such observations tend to occur in a small set of contexts and, therefore, may not be representative of entry-level practice. In addition, the intern is not likely to be assigned to high-risk activities on his or her own. As a result, performance assessments tend to raise serious problems of reliability and validity. Second, because the interns work in different places with different evaluators, the rating criteria need to be fairly general, and general criteria tend to be subjective. In this context, some evaluators are likely to be more severe than others, thus introducing a major source of error. In

addition, the possibility of bias for or against a candidate is likely to be more pronounced than it would be in a testing context; to the extent that the evaluator and candidate work together on a daily basis, it becomes progressively more difficult for the evaluator to provide an unbiased (i.e., neither too lenient nor too severe) evaluation of the candidate's performance.

The Difficulty in Implementing Performance Assessments

The first issue concerns the adequacy of the performance sampling. A performance assessment is likely to be most accurate as a measure of overall competence when the practice domain for which licensure is being awarded is relatively well defined and homogeneous (e.g., operating a particular kind of equipment). As a result, this approach tends to be most useful for licenses with relatively limited scopes (e.g., technicians responsible for a limited range of tasks). The more narrowly defined the scope of practice, the easier it is to observe candidates over comparable, representative samples of performance.

In contrast, the practice of a profession like law necessarily involves a wide range of activities in a wide range of situations. A candidate's performance is likely to vary from one situation to another, and even experienced practitioners may disagree on what to do in a particular situation. So it is hard to grade the performances consistently, and it is risky to make generalizations about readiness for practice from a small sample of performances. A candidate's performance on a few tasks in one setting does not necessarily say much about how the candidate would perform on other tasks in other settings. Performance on even moderately complex activities tends to vary substantially from one task to another, even when the tasks involve similar problems and contexts. This variability, called *task specificity*, is consistently found in performance assessments.

[T]HE PRACTICE OF A PROFESSION LIKE LAW NECESSARILY INVOLVES A WIDE RANGE OF ACTIVITIES IN A WIDE RANGE OF SITUATIONS. A CANDIDATE'S PERFORMANCE IS LIKELY TO VARY FROM ONE SITUATION TO ANOTHER, AND EVEN EXPERIENCED PRACTITIONERS MAY DISAGREE ON WHAT TO DO IN A PARTICULAR SITUATION. SO IT IS HARD TO GRADE THE PERFORMANCES CONSISTENTLY, AND IT IS RISKY TO MAKE GENERALIZATIONS ABOUT READINESS FOR PRACTICE FROM A SMALL SAMPLE OF PERFORMANCES.

To get a good overall estimate of a candidate's competence, a fairly large number of separate performances is required; however, professional performance assessments take time, and the need to observe a number of separate performances tends to make this approach cumbersome, time consuming, and prohibitively expensive to implement.

The Challenge of Ensuring Reliable Performance Evaluations

The second issue concerns the reliability and fairness of the performance evaluations. In the context of licensure, a bedside evaluation of medical students attending to patients, conducted by J.P. Hubbard, revealed that when one observer rated a candidate in one situation and another observer rated the same candidate in a different situation, their agreement was at the chance level (like rolling dice), indicating that the ratings were reflecting characteristics of the observers, the situations, or other extraneous factors, rather than the qualifications of the candidate.[12]

These problems get worse for apprenticeship programs in which the candidates operate in different contexts. In evaluating performances involving complex interactions with a number of people over a range of contexts, the scoring criteria must be quite general and, therefore, highly subjective. Specific, objective scoring rules are easiest to develop when the performance being evaluated is highly standardized and has a small range of outcomes that can be specified in advance.

Performance assessments are also potentially vulnerable to some kinds of bias that more objective assessments tend to preclude. In evaluating a candidate's performance in a real or simulated practice situation, the evaluator is likely to be aware of the candidate's age, gender, race, accent, appearance, and so on. Given that the evaluations require subjective judgments, it is essentially impossible to ensure that such extraneous factors have no influence on the results.

Concluding Remarks

I appreciate the opportunity I've had to work with the bar examining community on their examinations. I've seen great progress over the years, and I'm proud to have been a part of the ongoing efforts to improve the examinations. Much has been accomplished. There has been an ongoing effort to expand the range of KSJs being tested. The introduction of a multiple-choice component, the MBE, greatly extended the number of topics that could be included in a single bar exam and improved the reliability of the scores. The MBE enhanced the comparability of scores from one administration to another, through statistical equating, and scaling written component scores to the MBE also improved the comparability of these scores from one administration to another. The introduction of performance tasks increased the range of KSJs being covered, with a particular emphasis on basic practice skills.

In recent years, the procedures used to equate different forms of the MBE and to scale essay and performance task scores to the MBE have been updated and streamlined. The development and evaluation of questions for the MBE and for written components have been improved and made more consistent, and excellent, up-to-date practice materials for the MBE have been made available to candidates at very low cost. New modes of delivery have been developed to meet ADA requirements.

Ultimately, the quality of the examination-based decisions in a jurisdiction depends most heavily on the efforts of those responsible for the process: the court, the board of bar examiners, and their staffs. The boards of bar examiners have worked hard, individually and collectively through the CBAA, to improve administration procedures and, where appropriate, to adjust these procedures to accommodate special needs. Jurisdictions have worked hard to make the scoring of essays and performance tasks as accurate and consistent as possible. Disaster plans have been put in place, and test security has improved.

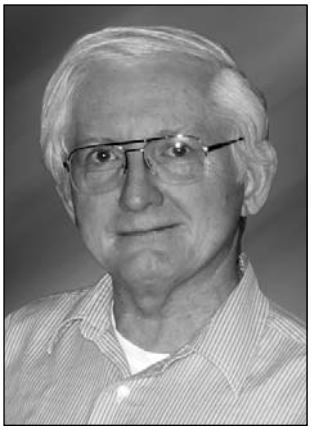
Much has been accomplished, but, as is always the case, there is more work to be done. Bar examination scores, like many other test scores, exhibit substantial differences between various ethnic, racial, and gender groups. These differences have been found to be highly consistent across different tests (MBE, essay tests, performance tests) and testing programs, including the LSAT, law school tests, and bar exams. They are not unique to bar exams and have not been found to be associated with any particular characteristic of bar exams, and therefore bar

examiners cannot eliminate the differences simply by changing the tests; however, bar examiners may be able to ameliorate the impact of the differences by keeping the process transparent and by continuing (and expanding) the availability of inexpensive, readily available practice materials. The availability of such materials will not cause unqualified candidates to pass, but they can make it less likely that otherwise qualified candidates will fail because of a lack of adequate test preparation. Improved communication between bar examiners and law schools and the introduction of online practice exams have been important steps in this direction.

I look forward to reading about future developments in THE BAR EXAMINER. I am changing jobs and residences, but my interest in the progress of bar examining will continue.

ENDNOTES

1. Benjamin Shimberg, *Testing for Licensure and Certification*, 36 AMERICAN PSYCHOLOGIST 1138–46 (1981); Michael T. Kane, *The Validity of Licensure Examinations*, 7 AMERICAN PSYCHOLOGIST 911–18 (1982); William C. McGaghie, *Professional Competence Evaluation*, 20 EDUCATIONAL RESEARCHER 1:3–9; Anthony LaDuca, *Validation of Professional Licensure Examinations: Professions Theory, Test Design, and Construct Validity*, 17 EVALUATION AND THE HEALTH PROFESSIONS 2:178–97 (1994).
2. The STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING were developed collaboratively by three professional organizations with a major emphasis on educational and professional testing in the United States: the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. The standards are revised every 10 to 15 years and represent a professional consensus on good practice in testing.
3. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING 156 (Washington DC: American Psychological Association 1999).
4. Samuel Messick, *Validity*, in EDUCATIONAL MEASUREMENT, 3RD EDITION (R.L. Linn, ed., New York: American Council on Education/Macmillan 1989).
5. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *supra* note 3, at 9.
6. *Id*. at 157.
7. Comparable in the sense that the two forms contain the same kinds of questions, cover the same general mix of content, and are administered under similar conditions. If graders are used, the results are graded in the same way by similar graders.
8. Candidates are not allowed to practice most professions until they pass the licensure examination for the profession, even if they have met all of the other requirements for admission to the profession. Furthermore, candidates are generally not allowed to take the exam until after they have completed their professional education and are ready to start practice, and a candidate who fails may have to wait a long time before retaking the examination. Certification examinations that can be taken multiple times before the consequences come into play (e.g., the MPRE) have lower stakes for this reason.
9. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *supra* note 3, at 156.
10. *Id*.
11. *Id*. at 157.
12. John P. Hubbard, MEASURING MEDICAL EDUCATION: THE TESTS AND PROCEDURES OF THE NATIONAL BOARD OF MEDICAL EXAMINERS (Philadelphia: Lea & Febiger 1971).



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.