The Role of Licensure Tests

by Michael T. Kane, Ph.D.

Licensure is a governmental function designed to protect the public by ensuring that candidates who are admitted to practice in the various professions have met certain basic requirements. In the United States, licensure programs are administered by states and other jurisdictions. For most professions and in most jurisdictions, the requirements for licensure include the successful completion of educational requirements, the passing of one or more examinations, demonstration of character and fitness to practice, and various administrative requirements (age, residence, submission of appropriate documentation and fees, etc.). The specific mix of requirements varies from one profession or occupation to another, and from one jurisdiction to another, but most involve these four elements.

Generally speaking, the goal of licensure programs is to protect the public from unnecessary risks. Licensure is not intended to provide a guarantee of excellent service, nor is it intended to rank candidates in terms of the quality of services they are expected to provide. Instead, licensure programs focus on specific characteristics that are clearly related to the safety and effectiveness of practitioners. These characteristics are assessed in some manner and reasonable standards for admission to practice are established for each characteristic. The requirements for admission to the bar typically include educational, character and fitness, and testing requirements.

The character and fitness requirements are designed to weed out candidates who are considered to be untrustworthy or unable to practice for one reason or another. In most cases, the evaluation of character consists of checks on whether candidates have done anything (e.g., committed a felony, lied about a significant matter) that would indicate a lack of integrity. The procedures used to evaluate character and fitness are not designed to identify candidates with especially good character or an especially high level of fitness, and the results of these evaluations are not expected to provide accurate predictions of future performance. Their function, rather, is to identify candidates whose past performance indicates a serious lack of character or fitness and who therefore represent a clear risk to the public.

Similarly, the educational and testing requirements are designed to provide protection against the risks inherent in admitting candidates who lack the basic knowledge, skills, and judgment necessary for safe and effective practice. They are designed to ensure that the candidate has achieved a reasonable level of competence in applying professional knowledge, skills, and judgment to practice problems. Without such competence, new practitioners would likely make mistakes that could put their clients at risk.

The educational and testing requirements operate on different levels and in somewhat different ways in ensuring basic competence. The evaluations of achievement in educational programs tend to occur over several years, to involve various kinds of skills assessments, and to be designed and implemented by a number of faculty. As a result, they can be more thorough than an evaluation based on an examination lasting a day or two. However, the implementation of educational requirements cannot be highly standardized, and therefore those requirements may not be very consistent across or even within institutions.

Licensure tests generally provide more standardized, objective evaluations of candidates, but these tests are also more limited in the skills they can assess. For example, an educational program could require students to complete a project that takes months; typically, licensure tests are limited to a few hours or days. Paper-and-pencil tests (whether multiple-choice or essay) are effective in assessing the cognitive skills involved in applying professional principles to practice situations described in the test questions. Educational requirements and licensure tests are complementary in the sense that they cover somewhat different but related skill domains in different ways, and therefore, together, provide stronger assurance of basic competence than either one would alone. According to the STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999):[1]

> 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. (STANDARDS, p.156)

Like other licensure requirements, bar examinations are designed to provide protection against risks that would be posed by professional practitioners who lack the knowledge, skills, and judgment expected of entry-level practitioners.

VALIDITY OF LICENSURE EXAMINATIONS

A candidate's readiness for practice, or competence in an area of practice, can be defined as the extent to which the candidate is prepared to handle the various issues and problems that arise in that area of practice. This is a very fundamental definition of competence (McGaghie, 1991; LaDuca, Engel, and Risley, 1978; Kane, 1992, 1994). In law practice, clients need professional help in solving certain problems, and the practitioner is expected to provide such help. Therefore, lawyers are competent, or ready for practice, to the extent that they can handle the kinds of issues and problems encountered in law practice. Competent practitioners are expected to achieve this goal by using their legal knowledge and skills to produce effective solutions to client problems.

There are two major components to definitions of professional competence in any profession. One component is the professional's ability to deal with a range of possible client problems that he or she might encounter, and the other involves the knowledge, skills, and judgment (competencies) that the professional is expected to bring to bear in addressing these problems. Combining these two components, a candidate's level of competence in an area of practice can be defined as his or her *ability to use professional knowledge, skill, and judgment to solve the kinds of problems encountered in practice*.

This definition provides us with a conceptual criterion for evaluating the strengths and weaknesses of different approaches to measuring competence or readiness for practice.

In developing and evaluating any assessment procedure, validity is the primary concern. 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." (AERA et al., 1999, p. 9). Basically, validity analyses address the question of whether the proposed interpretation and use of the test scores is justified.

Bar examination scores are expected to provide a dependable estimate of competence in the use of general legal knowledge and skill needed for effective performance in entry-level practice. If we could observe each candidate's performance in a representative sample of the situations likely to be encountered in entry-level practice, could accurately evaluate the performances in these situations, and had a large enough sample to yield reliable results, the interpretation of the results in terms of professional competence would be clearly valid.

For reasons discussed below, it is not usually possible to implement this kind of fully representative performance assessment. Rather, it is necessary to employ more structured assessments and more complicated chains of reasoning. The scores generated by actual bar examinations tend to be based on performance on standardized tasks under standardized conditions, but the tasks included in the test generally require the candidate to make the kinds of decisions that need to be made in practice (e.g., identifying the legal issues raised by a fact situation).

As is true for any licensure examination, the validity of scores on a bar examination depends on the plausibility of the chain of inferences from a candidate's observed performance to conclusions about the candidate's readiness for entry-level practice, and this in turn depends on the evidence supporting this interpretation and any evidence that tends to refute the interpretation (Cronbach, 1971; Messick, 1989). In evaluating the validity of any test score use, including professional licensure, it is important to look for possible flaws in the chain of inferences from the results of the measurement procedure to the conclusions being drawn and the decisions being made (Cronbach, 1980). In the next section, several alternative approaches to the assessment of readiness for practice will be briefly examined. One of these approaches, the compentency-based model, is then examined in more detail.

THREE APPROACHES TO ASSESSING READINESS FOR PRACTICE

This section provides analyses of three possible approaches to the development and validation of bar examinations: (1) as direct assessments of performance in practice, (2) as predictions of performance in practice, and (3) as assessments of specific competencies that are considered critical for performance in practice.

1. __Assessments Based on Performance in Practice__

In thinking about how to assess readiness for practice, it is natural to consider approaches in which the candidate's performance in actual practice situations is evaluated. If this approach could be implemented for a licensure test, it could yield a fairly straightforward interpretation, but the substantial difficulties inherent in developing and implementing this kind of performance-based assessment make it impractical for broadly defined professions like law.

Under the performance-based model, evaluations of observed performance in a sample of situations are generalized to the expected performance in entry-level practice. There are essentially two inferences made under this model: an evaluation of the observed sample of performance and a generalization from the sample to expected performance over the range of problems that tend to arise in entry-level practice.

Evaluation

The complexity of real practice situations tends to make the evaluation of performance difficult. Performance in solving real problems in real practice contexts generally involves complex interactions between the professional practitioner, one or more clients, and other professional and nonprofessional participants. It is not possible to specify rules in advance for evaluating such performances. The scoring rules are necessarily quite general and therefore the evaluations are necessarily subjective.

Scoring rules are easiest to develop when the desired performance has clearly defined outcomes that can be specified in advance. This may be the case for a particular job that involves a few specific tasks, but law practice involves a large number of activities in a wide variety of situations. A candidate's performance is likely to vary from one client to another, and expert graders may disagree on what the best choices are in a particular situation.

> Performance-based assessments are also potentially vulnerable to some kinds of bias that more objective assessments tend to eliminate. In evaluating a candidate's performance in a real practice situation, the evaluator is likely to be aware of the candidate's age, gender, race, accent, appearance, etc.

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

Generalization from the Sample of Performance to Practice

Once the observed performances have been scored, the second inference involves a generalization from these observed performances to the expected performance in practice. The fact that a candidate has done well or poorly on a particular set of professional tasks in a particular context is mainly of anecdotal interest unless these results can be generalized to entry-level practice.

There are two major threats to the plausibility of generalizations from an observed sample of performance to conclusions about expected performance in practice. First, the observed sample of performances may not be representative of entry-level practice in various ways, and second, the sample of performances may be too small to provide a dependable estimate of expected performance.

Obtaining a representative sample of performance is especially difficult in professions like law, because the professional activities are quite varied, occur in a wide range of contexts, often take a long time to develop, and can involve risk to clients. Although the purpose of licensure is to protect the public, a candidate is not likely to be tested in situations that involve real risk to clients. This difficulty can be partially overcome by using simulated situations or by having the candidate's performances in high-stakes situations be carefully supervised. However, both of these options make the observed sample of performance less representative of actual practice.

Human performance on even moderately complex activities tends to vary substantially from task to task, even when the tasks involve similar problems and contexts (e.g., a student may do a great job in researching one legal issue, but entirely miss the point on a related issue). This variability, called "task specificity," is consistently found in performance tests. Some task specificity is undoubtedly due to grader variability, but much of it is due to variability in the performances themselves. The variability does not go away even when the different performances are rated by teams of well-trained graders.

Hubbard (1971) reported that after three years of attempting to develop a reliable bedside evaluation of medical students attending to real patients, the National Board of Medical Examiners found 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. In another study, Hoffman (1977) found that an oral examination based on a physician's interaction with a client had a low reliability because of variability in performances from one situation to another. Where such results occur, one must conclude that the ratings are, to a large extent, reflecting characteristics of the graders, the situations, or some other factors, rather than measuring the qualifications of the candidate.

The sample of observed performances will generally need to be fairly large, because the only way to control for the error arising from task specificity is to include a large number of separate performances in the assessment of each candidate. With multiple tasks in different contexts and with different graders, variability in performance across tasks, contexts, and graders tends to cancel out. However, professional activities can take a fair amount of time to complete, and the need for large samples of observed performance tends to make this approach cumbersome and prohibitively expensive to implement.

One way to try to control the measurement error associated with task specificity is to employ an internship or apprenticeship in which the candidates can be evaluated over a longer period of time as they engage in professional activities in real practice settings. The main benefit of this approach is the opportunity to observe a large number of performances for each candidate.

However, there are two major disadvantages to this apprenticeship approach. First, the observations tend to occur in a small set of contexts, and therefore, may not be representative of entry-level practice. In particular, because the supervision in an internship or apprenticeship would be less thorough than it would be in a test, the intern is not likely to be assigned to high-risk activities. Second, because the interns will be working in different places, they will have different evaluators and the rating criteria will be fairly general and subjective. Some evaluators may 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 the candidate work together on a daily basis, it becomes progressively more difficult to provide an unbiased (i.e., neither too lenient nor too severe) evaluation of the candidate's performance.

The performance-based approach is most likely to be successful 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 (i.e., feasible and valid) 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.

2. *Test Scores as Predictors of Future Performance*

Another approach, which is commonly adopted in employment testing, is to interpret a candidate's test scores as predictions of subsequent performance in practice. This interpretation emphasizes the relationship between test scores and some criterion measure of performance in practice, and it does not necessarily put many restrictions on how the test scores are generated as long as they predict the criterion performance reasonably well.

In order to use test scores as accurate predictors of future performance, there must be some appropriate and dependable method of evaluating the future performances of those who pass the examination. The accuracy of the test scores as predictions of future performance in practice can then be evaluated using a number of statistical procedures. This interpretation is quite direct; high test scores are expected to be associated with good performance and low scores associated with poor performance.

However, there are a number of serious problems in using licensure examinations to predict future performance (Shimberg, 1981; Kane, 1985). In particular, a demonstration that test scores predict the quality of future performance in practice requires the development of a valid measure of overall performance in practice to be used as the criterion. Valid measures of performance in practice are not readily available (Swanson, 1990). By definition, law practice involves a wide range of activities performed in a variety of settings. Few of these tasks are entirely routine, and many require a high level of professional judgment for effective performance. The measures of professional achievement that are most readily available (awards, income levels, publications, etc.) do not provide appropriate criteria for validating a bar examination.

For a bar examination, the criterion measure of performance in law practice should focus on the quality of services provided to the public. It is very difficult to develop a measure of the quality of services provided to clients over the many different contexts in which lawyers practice. Some may have a broad, general practice and others may be quite specialized. Any evaluative criterion that applies to a wide variety of practice patterns and settings will necessarily be very general. As a result, the evaluations of performances in practice will tend to be quite subjective and will be susceptible to the sources of bias to which such evaluations are prone (e.g., possible grader biases, differences due to work settings).

In addition to the criterion problem, there are several practical problems in validating a predictive interpretation for licensure test scores. (Shimberg, 1981; Kane, 1982). In particular, it is very difficult to establish an empirical relationship between pass/fail status on the examination and performance in practice when only passing candidates get to practice.

Predictions about passing candidates could in principle be evaluated if a suitable performance criterion were developed, but predictions about failing candidates cannot be examined at all. This is a serious limitation, because the difference in performance between passing and failing candidates is the main issue in evaluating licensure tests.

Furthermore, because performance in practice is likely to be influenced by many variables (e.g., the practitioner's physical and mental health, work opportunities, and life events), which may change over time in unpredictable ways, predictions of future performance are not likely to be very accurate, and it is difficult to say how good the predictions need to be in order to be considered good enough.

In addition to the criterion problem, there are several practical problems in validating a predictive interpretation for licensure test scores. . . . In particular, it is very difficult to establish an empirical relationship between pass/fail status on the examination and performance in practice when only passing candidates get to practice.

3. ***Test Scores as Measures of Basic Competency***

A third approach to the development of licensure tests, including bar examinations, employs a competency-based model, in which the test is designed to assess specific competencies that are considered critical for effective performance in practice.

A bar examination is designed to measure some set of critical competencies (e.g., the ability to apply general principles to fact situations), and scores are interpreted in terms of level of skill in these competencies. The use of the scores as measures of readiness for practice involves three inferences: an evaluation of performance on the test tasks, an inference from the observed performance to conclusions about the candidate's level of skill in the critical competencies, and an inference about readiness for practice based on mastery or non-mastery of the critical competencies. No attempt is made to evaluate performance in actual practice situations, although the competencies are assessed in practice-related tasks or simulations.

For professional licensure tests, the competency model is the standard approach. This is the case because the performance-based model and the predictive model are not considered feasible and because the competency model is consistent with the basic purpose of licensure and can provide an adequate basis for licensure decisions. 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. . . . Panels 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. (*Standards*, p. 156)

Professional practitioners and faculty are asked to identify competencies that are critical for effective performance in practice and the test tasks are developed to assess these competencies.

The interpretation of bar examination scores as measures of certain critical competencies is consistent with various other governmental activities designed to protect the public in that they focus on specific dangers and do not try to guarantee successful outcomes. The pattern is basically the same for all licensure programs. Assuming that a basic set of competencies has been identified as being critical for a profession, it is reasonable for the state to require that candidates for licensure demonstrate mastery of these competencies before being licensed to practice.

> Assuming that a basic set of competencies has been identified as being critical for a profession, it is reasonable for the state to require that candidates for licensure demonstrate mastery of these competencies before being licensed to practice.

Interpreting Scores as Measures of Competence

In designing a compentency-based assessment procedure, the goal is to support the three major inferences: evaluations of observed performance, conclusions drawn about critical competencies from test scores, and decisions about licensure based on mastery or non-mastery of the competencies.

Because the test is to be designed to support inferences about the readiness for practice, it is reasonable to start the process of test design and development with an analysis of entry-level practice. The scope of practice provides a preliminary framework for describing entry-level practice; presumably law practice does not go beyond these limits. Expert judgment can refine this description by ruling out areas of practice that are too advanced or specialized to be considered part of entry-level practice.

Information about the contexts in which new practitioners work and the kinds of problems that they work on can be helpful in defining the scope of entry-level practice more clearly and in assigning weights to different areas of practice. Empirical surveys in which new practitioners provide information about their professional responsibilities can also be helpful in delineating the scope of entry-level practice and the weights to be assigned to different areas of practice.

The knowledge base of the profession indicates the knowledge and skills that are needed in dealing with the various problems that arise in practice. The next step is to identify test tasks that measure the critical competencies, and are preferably not influenced much by factors unrelated to performance in practice. For example, a test question could present a fact situation involving a contract dispute, and the candidate would be asked to address some aspect of this dispute.

A sample of these tasks is administered to candidates and a score for each candidate is based on the candidate’s performance. The number of competencies that could be included on a licensure test is quite large, and the number of possible test tasks associated with each competency is also large. It is not possible to have candidates complete more than a small sample of these tasks. But the sample of tasks on the test is expected to be large enough and representative enough to provide a good indication of the candidate's level of achievement in the critical competencies. Fairness is enhanced by having all candidates complete the same tasks, and by designing scoring procedures that are as objective as possible.

Critical Competencies

The intent in determining critical competencies is to identify areas of knowledge and skill 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 law effectively. The competencies are linked to practice in the sense that they are considered necessary for effective practice. A high level of achievement in 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 make one a good negotiator, but ignorance of these laws on the part of lawyers would put clients at risk.

Licensure tests are designed to be highly relevant to the practice domain. If it were feasible to evaluate performance in practice directly, this would probably be the preferred approach; as indicated earlier, however, this is typically not possible. If it were possible to evaluate all of the competencies relevant to success in practice, this would also be a good way to evaluate readiness for practice, but it is also impossible. So the test measures a limited set of competencies that are critical for effective performance in practice. Licensure tests are designed to assess competencies that (a) are considered critical for effective performance in common practice activities, and (b) can be assessed, fairly and consistently, within the test format being used.

The first of these criteria (criticality) tends to rule out routine aspects of practice (scheduling, record keeping, etc.) and to focus on competencies relevant to activities that have consequences if a mistake is made. As Rakel (1979) put it, "There is a justifiable need to test more heavily on problems that . . . fall into the 'uncommon but harmful if missed' category" (p. 93). Given that the purpose of licensure is to protect the public, the "harmful if missed" category is especially important.

The second criterion (measurability) leads to an emphasis on the cognitive aspects of practice (involving knowledge, analytical skills, and judgment) and specific skills, which can be reliably assessed, rather than complex performances, which would require more subjective judgment. The test tasks are developed to assess competence in specific activities that occur fairly often in law practice and are considered critical to successful practice outcomes. In many cases, the specific competencies addressed by the test tasks are sub-components of more broadly defined practice activities (e.g., making specific professional decisions within the context of a practice activity).

The art is to include essential elements of practice activities in tests tasks so that those who succeed on the test tasks are likely to succeed in corresponding practice activities and those who fail on the test tasks are likely to fail in the practice activities. The tasks included in bar examinations typically involve the application of legal principles to fact situations. For example, a multiple-choice question could describe a common practice situation and then ask the candidate to indicate which of several courses of action to recommend to a client.

Designing the Test and the Scoring Rules

Once we have some understanding of the competencies involved in practice, the question is how to assess those competencies. The procedures used to assess readiness for practice should provide accurate measures of the competencies involved in practice or of a substantial subset of these competencies.

An effective way to develop test tasks that are clearly related to performance on practice activities is to use tasks that involve relatively critical parts of the practice activities. The candidate is not asked to complete the full activity, which might be quite time consuming, but to carry out some essential part of the activity. For example, attorneys defending clients who are accused of a crime are expected to protect their clients' rights; a test task related to this activity might ask a candidate to indicate how they would respond to a potential threat to their client's rights. More generally, any legal activity is likely to involve a number of professional decisions, and the questions on a written test could ask about these decisions. Although the activity (e.g., a trial) might take days, weeks or months, the test task might take a minute (e.g., "Given that the prosecutor has asked the witness question X, which of these four options is the best course of action for the defense attorney?").

Standardized measurement procedures are used for most licensure programs. In the interest of fairness, tasks are standardized so that one candidate isn't facing a very difficult task while another has a very easy task. Everyone faces the same tasks. In the interest of efficiency, relatively little attention is given to routine performances that provide little information about the candidate's competence. Specific scoring rules are developed for each task, and tasks that are too complicated or idiosyncratic to be easily scored are avoided. The tasks, the conditions of observations, and the scoring criteria are all standardized.

The use of standardized tasks to which all candidates respond also facilitates the development of fair and objective scoring rules. For multiple-choice items, the scoring rule (or scoring key) is completely objective; the experts developing and reviewing items go to great pains to ensure that there is a clearly best answer to each question. Since the questions ask candidates to make judgments about fact situations, there may be no single correct answer, but by making the questions clear enough and the options different enough, there can still be a clearly best answer.

For essay tests and performance tests, scoring is more difficult. In these cases, the experts developing the test tasks are also expected to develop a model answer and scoring rules that reflect the main points that need to be addressed in an answer to the question. By working together in applying these scoring rules to sample candidate responses, graders can learn to grade papers with a high degree of consistency; this process is called "calibration." The scoring of essays and performance tasks is always more subjective than the scoring of multiple-choice items, but the graders' evaluations can achieve a high degree of consistency.

As noted above, the tasks included in the test may be very similar to activities that occur in practice, or may constitute critical parts of these activities. To the extent that test tasks are closely related to practice activities, inferences from observed performance on the test tasks to future performance in practice are facilitated.

Licensure Decisions

Standard test development procedures are designed to minimize the chances that candidates who have mastered the competencies being assessed will get low scores or that candidates who have not mastered the competencies being assessed will get high scores. In this vein, it is important that no irrelevant barriers to success be allowed into the test. For example, the language used in the test should not unduly interfere with performance. Except for the use of technical vocabulary, the reading level of the examination should be kept at a moderate level.

Once the test developers have specified the practice domain, identified a set of critical areas of knowledge and skill, and defined the test tasks that will be used to evaluate candidates' level of achievement in the critical competencies, it is still necessary to decide how well a candidate has to perform on the test tasks in order to be considered ready for practice. In other words, it is necessary to identify a particular level of achievement in the critical competencies, as reflected in test scores, below which a candidate is considered inadequately prepared for practice.

Choosing a passing score is a critical issue in licensure testing. The passing score is necessarily based on judgment (i.e., about how good is good enough?), but empirical studies involving panelists with content expertise and familiarity with entry-level practice can provide empirical support for setting a passing score (Cizek, 2001). The passing score is supposed to be high enough to provide adequate protection to the public, but not so high as to unduly restrict access to the profession. The central question is the level of test performance that reflects the level of competence needed in entry-level law practice (Kane, 2002).

For a licensure examination that measures a set of critical competencies rather than all of the characteristics required for good practice, it is important that the standards on the examination not be set too high. Although some level of mastery of the competencies included in the test is considered necessary for practice, it is not necessarily true that higher levels of mastery will lead to improved performance. In general, bar examinations emphasize the ability to apply legal principles to practice situations because these skills can be measured accurately with written tests. Since cognitive skills are important for practice, such testing is appropriate, but the standards for these competencies should not be higher than the level of ability required for entry-level practice. 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.

Summary Conclusion

To validate a test-score interpretation or use is to show that the proposed interpretation and use are supported by appropriate evidence. This is accomplished by stating the interpretation explicitly, by showing how the decisions based on test scores follow from the scores and the assumptions inherent in the interpretation, and by making the case that the assumptions are reasonable.

Bar examination scores can be interpreted as measures of competencies that are critical for practice in the sense that they are necessary, but not sufficient, for effective performance in practice. A high level of achievement in the critical competencies does not guarantee success in practice, but lack of competence would be a serious impediment in practice and would tend to put clients at risk. If candidates perform poorly on the test, we can conclude that they lack the critical competencies to a substantial degree, and if they lack these competencies, they are not likely to perform adequately in practice. The inferences in this sequence are all fallible and exceptions are possible, but for most cases, the interpretation is reasonable. Therefore, candidates with low scores on the licensure test can be considered inadequately prepared for practice.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

Cronbach, L.J. 1971. Test Validation. Edited by R.L. Thorndike, Educational Measurement. Washington, DC: American Council on Education.

Cronbach, L.J. 1980. Validity on Parole: How Can We Go Straight? Edited by W.B. Schrader, *New Directions for Testing and Measurement—Measuring Achievement: Progress Over a Decade, no. 5.* San Francisco: Jossey-Bass.

Kane, M.T. 1982. The Validity of Licensure Examinations. *American Psychologist*, 7:911-918.

Kane, M.T. 1992a. An Argument-Based Approach to Validity. *Psychological Bulletin*, 112:527-535.

Kane, M.T. 1992b. The Assessment of Professional Competence. *Evaluation and the Health Professions*, 15:163-182.

LaDuca, A. 1994. Validation of Professional Licensure Examinations: Professions Theory, Test Design, and Construct Validity. *Evaluation and the Health Professions*, 17, 2:178-197.

LaDuca, A., Engel, J.D., and Risley, M.E. 1978. Progress Toward Development of a General Model for Competence Definition in Health Professions. *Journal of Allied Health*, 7:149-155.

LaDuca, A., Taylor, D., and Hill, I. 1984. The Design of a New Physician Licensure Examination. *Evaluation and the Health Professions*, 7:115-140.

McGaghie, W.C. 1980. The Evaluation of Competence: Validity Issues in the Health Professions. *Evaluation and the Health Professions*, 3(3):289-320.

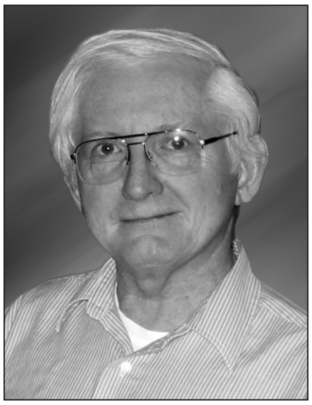
McGaghie, W.C. 1991. Professional Competence Evaluation. *Educational Researcher*, 20(1):3-9.

Messick, S. 1989. Validity. Edited by R.L. Linn, Educational Measurement (3rd ed.). New York: American Council on Education and Macmillan.

Norcini, J. 1994. Research on Standards for Professional Licensure and Certification Examinations. *Evaluation and the Health Professions*, 17, 2:160-177

Shimberg, B. 1981. Testing for Licensure and Certification. *American Psychologist*, 36:1138-1146.

Swanson, D. 1990. Issues in Assessment of Practice Skills in Medicine. *Professions Education Research Quarterly*, 12:3-6.



Michael T. Kane, Ph.D., is the Director of Research for the National Conference of Bar Examiners.