The Testing Column
What Everyone Needs to Know about Testing, Whether They Like It or Not

by Susan M. Case, Ph.D.

Like it or not, the minimally competent adult needs to know a few things about measurement in general and testing in particular. You can't listen to NPR or pick up an issue of *Sports Illustrated* or *USA Today* without finding references to measurement concepts: sampling error, margin of error (plus or minus some number of percentage points), predictions of success or failure. These stories relate to social polling ("Which is more popular: ketchup or salsa?"), sports ("Which golf club is more consistent?"), politics ("Which candidate is more likely to win the election?"), and almost everything else in our day-to-day lives.

At NCBE's recent Annual Bar Admissions Conference in Savannah, I outlined four measurement concepts that everyone involved in bar admissions should understand: sampling, reliability, validity, and scaling.

Sampling

Sampling refers to using a representative subset of a larger group (of questions, interview subjects, etc.) to gain information that can be generalized to the larger group. Because you can't test everything a new lawyer needs to know, the bar exam asks as many questions as is logistically possible and economically feasible. The assumption is that the scores would generalize to a different set of questions so that if you were to test the same group of examinees again using a different form of the test (such as the February exam instead of the July exam), each examinee's total scaled score would be virtually the same on both forms.

The broader the content domain, the more questions are required. For instance, testing children on their skill in multiplying two-digit numbers requires only a small number of questions in order to gain a good idea of the level of proficiency of each child. Other skills might require a larger set of questions. But for a given topic or set of topics, all else being equal, the larger the sample of questions the more likely you are to have a good estimate of knowledge and skills.

Reliability

Reliability is closely related to sampling. Reliability estimates the extent to which a group of examinees would be rank-ordered the same if a second similar test was administered. In other words, can you trust the score that the examinee received as being representative of that examinee's level of knowledge and skills in the area tested? As is true with sampling, all else being equal, the more questions you ask, the higher the reliability.

Reliability of Written-Component Scores

If more questions provide greater reliability, it follows that reliability is reduced when fewer questions

are used. A score on a single essay does not predict very well how an examinee is likely to perform on a second essay; some examinees are going to be lucky in terms of how well one or both of the essay questions correspond to their knowledge of the topic, and some are going to be unlucky. As you increase the number of essay questions, the reliability of the score increases, indicating that your estimate of how well that examinee would perform if you had asked a different set of questions is more precise. The problem with written-component (essay and performance test) scores is that bar exams generally have a small number of such questions, and scores based on a small number of questions do not have sufficient reliability for high-stakes tests.

Reliability is also reduced when there is inconsistency. In the case of written-component tests, overall question difficulty varies from one test administration to the next; grader stringency varies from one test administration to the next; and grader stringency also varies from one grader to the next. Statistical equating of essay scores, which would adjust for varying levels of difficulty (as discussed later) is not feasible because essay questions are not reused.

Reliability of Multiple-Choice Scores

While written-component tests have their limitations, the MBE is not a panacea; multiple-choice questions have an image problem. No matter how high the quality of the questions, some people continue to believe that multiple-choice questions are just too far from the real world to be useful, and furthermore that providing the examinee with options to choose from makes the assessment challenge even less realistic. However, the relationship between scores on the written components and scores on the MBE is quite high (correlations usually range from the 0.60s to the 0.80s), indicating that those who do well on the written components tend to also do well on the multiple-choice component, and vice versa. The reliability of the total score constructed by adding together the total written-component score and the total MBE score (equally weighted) is large enough to meet minimum reliability standards for high-stakes tests.

The advantage of the MBE is that the total scaled score is very reliable, assuring that if you were to retest the examinees using a similar exam, the rank-ordering of examinees would be very similar. In addition, scores are not affected by grader traits such as grader inconsistency across time, leniency/stringency, and inconsistency with other graders. Another advantage of the MBE is that content is broadly sampled; luck doesn't play much of a role when each examinee has questions covering 200 cases to answer. And the final advantage is that scores are equated over time to ensure that equivalent levels of performance are required to achieve a passing score. If a particular MBE is slightly more difficult than the last one, the scores are adjusted to take this varying difficulty into account. This adjustment is called *equating*.

As an aside, equating is done with all standardized tests. Some tests do not provide as much information to the general public as the bar exam does, so you might be unaware of this. For example, all standardized tests that children take in school or that are used for admission into college or graduate school use equating and report only scaled/standardized scores; all licensing and certification exams in other professions follow the same practice.

VALIDITY

Validity in testing refers to the extent to which the test score reflects the attribute you are intending to measure. In the bar exam, validity means ensuring that you are testing what a newly licensed lawyer needs to know. Multiple testing methods are used because each method has strengths and weaknesses, and

each is designed to test somewhat different skills, each of which is believed to be important for the practice of law. The pass/fail standard is set at a level that is believed to protect the public from applicants who lack the requisite knowledge and skills to be licensed to practice. Scores that are unreliable cannot be valid. However, validity requires more than just reliability; it is not enough to be consistent if you are consistently measuring the wrong thing. The fundamentals of reliability and validity are described in a previous Testing Column entitled "Back to Basic Principles: Validity and Reliability."[1]

SCALING

The fourth important concept is *scaling*. Scaling written-component scores to the MBE involves an algebraic process that places the written-component scores on the same scale as the MBE. This process "equates" the written-component scores and assures that the scores mean the same thing across test administrations. Scaling eliminates the variability in essay question difficulty from one test administration to the next. Scaling also eliminates the variability in grader stringency from one test administration to the next. Several previous Testing Columns have been devoted to scaling. One, entitled "Frequently Asked Questions About Scaling Written Test Scores to the MBE,"[2] answers common questions about scaling, and another, entitled "Demystifying Scaling to the MBE: How'd You Do That?"[3] describes the algebra behind actually doing the scaling.

Scaling written-component scores to the MBE corrects for changes in overall question difficulty or grader severity from one test date to the next. So, scaled written-component scores more accurately reflect examinee competence (on skills measured by the essays and performance tests) rather than characteristics of the questions and graders. The rank-order of examinees' written-component scores will remain exactly the same before and after scaling. The examinee with the highest written-component score before scaling to the MBE will still have the highest written-component score after scaling. Some individuals will score higher on the MBE and lower on the written component, but over the entire group the average MBE score and the average written-component score will be the same.

This last feature is what makes total scaled test scores transferable across jurisdictions regardless of the grading scale used on the written component. For example, if the average MBE score for a jurisdiction is 140 and that jurisdiction adds up its written-component scores and finds that its total written-component score average is 60, the 60 would be scaled to become a 140. If, on the other hand, another jurisdiction with an average MBE score of 140 had an average written-component score of 450, the 450 would be scaled to become a 140. Scaling to the MBE is a transformation that puts everything on the MBE score scale, while at the same time taking advantage of the equating that is possible with a multiple-choice test and not with a written test.

Is it an overstatement that familiarity with these four principles of testing will make your whole life clearer? Maybe, but it will certainly make you more effective at carrying out your responsibilities in the realm of bar admissions.

NOTES

1. Susan M. Case, Ph.D., *Back to Basic Principles: Validity and Reliability*, 75(3) THE BAR EXAMINER 23–25 (Aug. 2006), *available at* http://www.ncbex.org/assets/media_files/Bar-Examiner/articles/2006/750306_testing.pdf.
2. Susan M. Case, Ph.D., *Frequently Asked Questions About Scaling Written Test Scores to the MBE*, 75(4) THE BAR EXAMINER 42–44 (Nov. 2006), *available at* http://www.ncbex.org/assets/media_files/Bar-Examiner/articles/2006/750406_Testing.pdf.
3. Susan M. Case, Ph.D., *Demystifying Scaling to the MBE: How'd You Do That?*, 74(2) THE BAR EXAMINER 45–46 (May 2005), *available at* http://www.ncbex.org/assets/media_files/Bar-Examiner/articles/2005/740205_testing.pdf.

SUSAN M. CASE, PH.D., is the Director of Testing for the National Conference of Bar Examiners.