The Testing Column

Back to Basic Principles: Validity and Reliability

by Susan M. Case, Ph.D.

Today's world is full of tests, but some are so important that special care must be taken to ensure that the results accurately reflect the examinee's level of performance. Such tests are labeled "high-stakes examinations," and include examinations used in the broad category of licensure and certification. The central purpose of these examinations is to identify examinees who are sufficiently competent to practice in the area covered by the license.

The following statement is excerpted from *The Standards for Educational and Psychological Testing*, a set of guidelines that are developed and periodically updated by a joint commission representing the three main associations of testing professionals. *The Standards*, as this document is known, is developed to provide criteria for the evaluation of tests, and is generally accepted by measurement experts.

> The primary purpose of licensure or certification is to protect the public . . . [and] to provide the public . . . with a dependable mechanism for identifying practitioners who have met particular standards. The focus of test standards is on levels of knowledge and skills necessary to assure the public that a person is competent to practice . . . [and] to help ensure that those certified or licensed meet or exceed a standard or specified level of performance. (*Standards*, p. 63-64.)

Validity—Does the Test Measure What It Is Meant To?

The Standards includes a number of requirements specifying the need to ensure that scores are both valid and reliable. Validity of test scores for a particular test is not something that can be easily quantified, but rather involves amassing a body of evidence that would lead reasonable people to conclude that the scores reflect what the examination was intended to measure. One appropriate form of evidence relates to examination content and might involve collecting judgments from various people about the extent to which the content of the examination is directed at assessing the knowledge and skills needed by a new practitioner. A second area of investigation would focus on the extent to which extraneous factors, including a broad range of issues from handwriting to time limits, impact examination scores. Typically these studies of extraneous factors would be exclusionary; for example, one would hope to find no evidence that handwriting affects scores.

Reliability—Do the Scores Accurately Reflect the Examinees' Proficiency?

Reliability is a prerequisite for validity; a set of scores with low reliability cannot be valid. In examinations like those used for bar admissions, reliability has two main components: consistency across forms and graders, and consistency across time. For high-stakes examinations, it should be irrelevant whether an examinee was tested on Test A or Test B, whether the test questions were scored by Grader Set 1 or Grader Set 2, and whether the test was administered on Occasion 1 or Occasion 2.

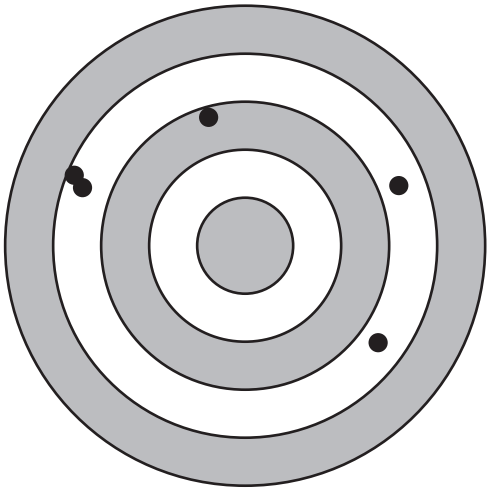
Consistency Across Forms and Graders

Consistency across forms and graders refers to the extent to which the rank-ordering of a group of examinees would be similar if they were tested again using a set of similar, but not identical, questions that were graded by a similar, but not identical, set of graders. Almost all tests include only a subset of the questions that could have been asked, and a test with a small number of questions is likely to advantage some examinees and disadvantage others due to the specific sample of questions included on the test. As the number of questions increases, this luck factor is reduced. Thus, reliability is very directly related to the number of questions asked.

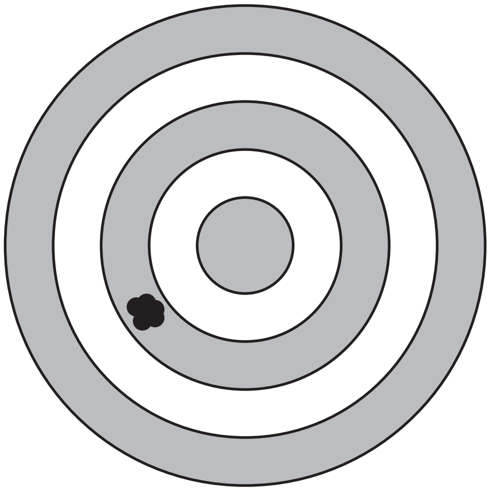
Similarly, everyone knows that graders of essay answers vary in stringency, and examinees are likely to be advantaged or disadvantaged based on the number of relatively harsh or lenient graders that score their answers. If a single grader graded all examinees on a question, and did so consistently, then the examinees would be treated equally for that question—if the grader were harsh, everyone would receive relatively low grades on the question; if the grader were lenient, all would receive relatively high grades. A consistency problem arises when some examinees get a harsh grader for a particular question and others get a lenient grader for that same question. Inconsistency can also cause reliability problems when a grader changes criteria or standards from one paper to the next. Reliability is thus directly related not only to the number of questions asked, but also to the number of graders who grade the set of answers submitted by an individual, and the consistency of each grader across examinees.

Consistency Across Time

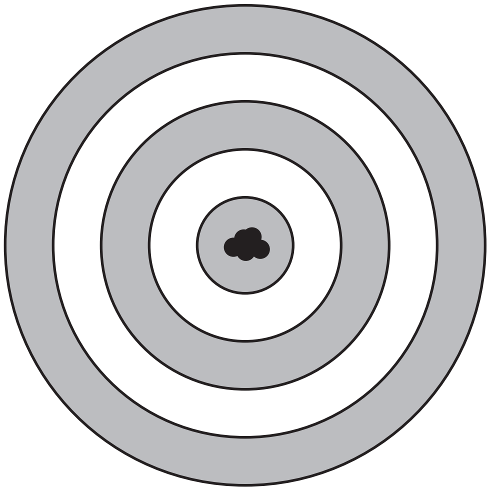
The second main component of reliability is consistency across time. It should be irrelevant to examinees whether they are tested in February or July, and whether they are tested this year or next. Unfortunately, graders of essay answers are not able to calibrate themselves over time, and it is likely that, as a group, the graders will be harsher on one occasion than they are on another. This phenomenon does not occur with multiple-choice tests that are equated, such as the MBE. In order to equate, some previously used questions must be included on every new form of an examination. By analyzing performance of a group of examinees on the reused set of questions, we can determine the proficiency of this group of examinees relative to other groups of examinees who took the same questions in the past. Then, by analyzing performance on the new questions in light of performance on the previously used questions, we can determine the relative difficulty of the new questions. Scores are "scaled" using the equating process to ensure that scores have a stable meaning over time. For example, an examinee with a scaled score of 135 on the July 2004 MBE would demonstrate the same level of proficiency as would examinees with scaled scores of 135 on any other MBE administrations.



This target shows an archer who is neither reliable nor valid; his arrows hit in an inconsistent pattern, and they fail to hit the bull's-eye.



This target shows an archer who is reliable but not valid; he is consistent (reliable) but not accurate (valid).



This target shows an archer who is both reliable and valid; he consistently hits the bull's-eye.

While the MBE and the MPRE are equated by embedding previously used questions into a new examination form, this process is not possible for essay and performance tests like the MEE and the MPT, because the extended-response items are too memorable to be reused. Nevertheless, jurisdictions can ensure that essay scores are consistent across time by scaling essays and performance test scores to the MBE and by basing their standards on the total combined score. Most jurisdictions are now doing this.

Summary

1. Those who develop and administer high-stakes tests such as those used for bar admissions are obligated to ensure that their examination forms and graders produce scores that are sufficiently consistent to be reliable and valid.

1a. This obligation must be met for any score that is used to make a decision about whether an examinee will be admitted to the bar. If separate hurdles are used, and an examinee must pass both a written component and a multiple-choice component, for example, there is an obligation to ensure that each score is based on a large enough sampling of content and is graded consistently enough to be reliable and valid.

1b. If scores on essays and performance tests are combined with the MBE, and if the MBE is weighted at least 50%, the total score is likely to be sufficiently reliable for licensure purposes.

2. Those who develop and administer high-stakes tests are obligated to ensure that their examination scores maintain the same meaning over time. Not doing so creates a situation where the passing standard varies across test dates. The optimal way of ensuring that the standard remains constant is to scale written scores to the MBE, producing scaled scores that are equivalent across time.

Susan M. Case, Ph.D., is the Director of Testing for the National Conference of Bar Examiners.