THE TESTING COLUMN

BEST PRACTICES WITH WEIGHTING EXAMINATION COMPONENTS

by Susan M. Case, Ph.D.

Nearly every jurisdiction administers a bar examination with more than one component, where each component is designed to assess a different set of skills. The jurisdiction then scales the scores of the written test components to the MBE, combines the scores from these component tests, and makes a pass/fail decision by comparing the single combined score with the established passing score. Making pass/fail decisions on a single combined score is preferred over making separate decisions on each component because decisions based on a single score are more reliable than those based on individual component scores. The issue of scaling written scores to the MBE is not discussed in this column. For more information on scaling, see the May 2005 and November 2005 Testing Columns.

The question arises regarding the appropriate weight to assign to each component. While this might seem like an esoteric topic, Google yields over one million references to weighted scores. Entries include a huge array of tests, ranging from the APGAR score used to evaluate the health of newborn infants to the credit score that determines whether or not you obtain the loan you requested at an optimal rate. Your total credit score, for example, includes five components: a score reflecting your payment history, which is weighted 35%; a score based on your debt amount, which is weighted 30%; the length of your credit history, which is weighted 15%; and two variables related to new credit and credit mix, each of which is weighted 10%.

In each instance of weighted scores, the weights are developed to make the total score as reliable and valid as possible. The term "reliable" refers to the extent to which the score is precise and consistent. The term "valid" means that the score is measuring what you want to measure. In the case of your credit score, researchers have evaluated the various components and determined the weighting percentages that provide the best indication of total creditworthiness.

In the case of the bar examination, we have conducted research to determine the weighting that provides the most reliable and valid total score. Reliability is quantifiable in a direct manner; there is a formula that can be used to calculate the precision of the score. The MBE total score is highly reliable. This means that if you were to test a particular group of examinees again using similar test questions, the score for each examinee from the first test to the second would be very similar. The large number of questions means that the content is sampled very broadly, and the overall exam content varies little from one MBE to the next. In contrast, scores

from written tests have much lower reliability. If you were to test examinees again using similar essay questions, scores would vary quite a bit—some examinees would be lucky in the particular selection of essay topics and some would be less lucky. Why is this important? If an examinee received two quite different scores on the essay tests, we would not know which score better reflected the examinee's level of proficiency. As a result of the very different levels of reliability of the two test components, reliability of the total score is maximized by giving a relatively large weight to the MBE.

Figure A

EXPECTED RELIABILITY OF THE TOTAL SCORE GIVEN VARIOUS WEIGHTS ASSIGNED TO THE MBE

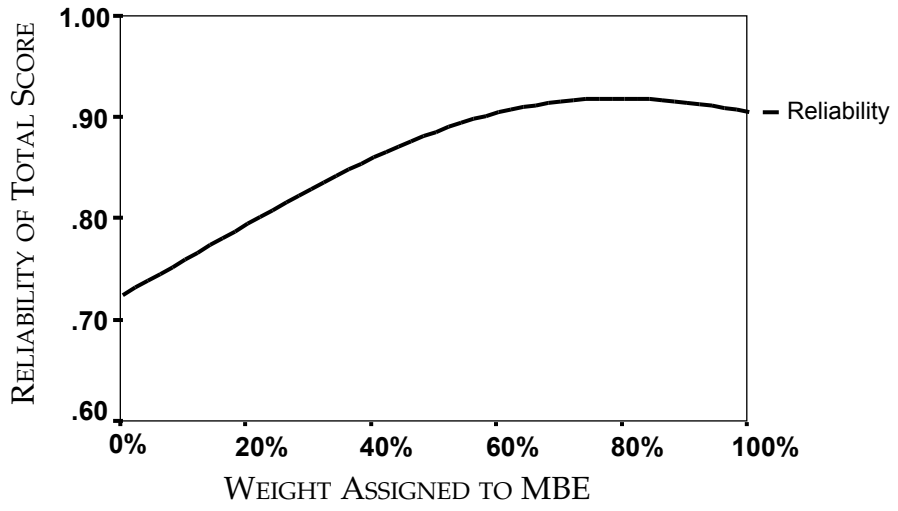
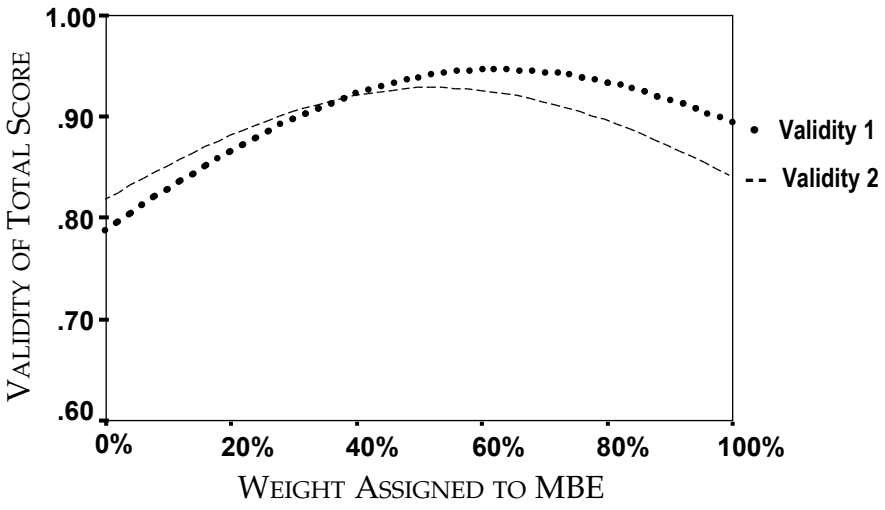


Figure A shows the relationship between the MBE score weight and reliability. The horizontal axis ranges from an MBE score weight of zero (i.e., weighting the essay/performance test component score at 100%) to an MBE score weight of 100 (i.e., weighting the MBE score 100% and not counting the written portion at all). Of course, these extremes are not used; most jurisdictions weight the MBE score between 40% and 50%. The vertical axis shows the reliability that would result from each possible MBE score weight. Because reliability indicates the precision of the score, you want the reliability value to be as high as possible. For high-stakes tests, such as the bar examination, testing standards indicate that the test scores on which decisions are made should have a reliability of at least 0.90. Figure A shows that the total test score comes close to this value as long as the MBE weight is at least 50%. Note that the reliability value rises fairly quickly from a low of about 0.70 up to about 0.90 and then plateaus with a very small decrease after that. This indicates that increasing the weight of the MBE is beneficial up to a point, but that increasing the weight beyond 60–70% does not improve the reliability of the total score.

The above discussion refers only to reliability, which is only half of the equation; we also must be concerned with validity. Validity assesses the extent to which the scores are measuring what they are intended to measure. Some people believe that essay and performance tests assess the most important aspects of readiness to practice law, and would therefore wish to weight those scores very highly. The problem that arises is that these written test components do not produce sufficiently reliable scores on their own. A battery of written tests typically yields a reliability of 0.70, which is not high enough to meet industry standards for scores

Figure B

EXPECTED VALIDITY OF THE TOTAL SCORE GIVEN VARIOUS WEIGHTS ASSIGNED TO THE MBE



Reliability of MBE = 0.90 Reliability of essay test = 0.72 Correlation = 0.60

used to make high-stakes decisions about individuals. Regardless of the weights that seem appropriate from a validity perspective, reliability must also be taken into account, because a score that is not reliable cannot be valid.

To illustrate this issue, consider a situation in which you need to determine the productivity of lakes in Wisconsin. The most valid way of making this determination is to measure dozens of attributes from many areas in each lake, but this process is deemed to be too costly and too error-prone to be acceptable. In situations like this, researchers often use surrogate measures that are more reliable and, because of this reliability, produce a more accurate measure of the target—even though they do not measure the target directly. In this case, researchers have found that measuring walleye growth produces a "score" that is directly related to lake productivity.

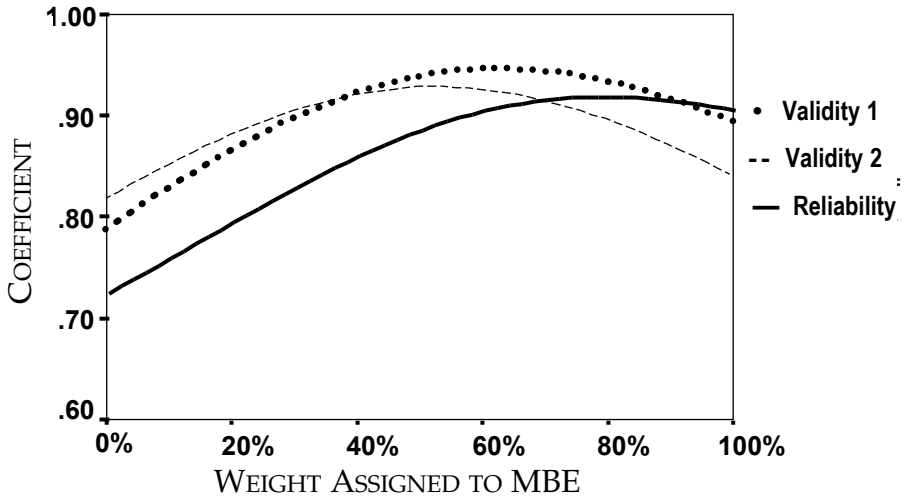
Studies undertaken at NCBE, using data from many jurisdictions, have shown that utilizing the relationship between reliability and validity can aid jurisdictions in making decisions about the optimal weights for exam components. Results show that giving extra weight to the more reliable of the scores improves the reliability of the total combined score (as expected), but they also show that giving extra weight to the most reliable of the scores, up to a point, also improves the validity of the combined score.[1]

The figures show the relationship between reliability, validity, and relative weights. As discussed on the previous page, Figure A shows that the reliability of the total score increases as the weight of the MBE score is increased. For a high-stakes test, where the score can have a major impact on the individual examinee, a reliability of at least 0.90 is viewed as appropriate. Figure A shows that a reliability of 0.90 is achieved as long as the MBE score is assigned a weight of at least 50%.

Figure B on this page shows two validity curves that occur as the weight of the MBE score is increased. The dotted line labeled Validity 1 shows the validity curve that would result if you believe that the written component and the MBE are equally important. Note that the validity values are lowest when the MBE is given the least weight; as the MBE weight

Figure C

EXPECTED RELIABILITY AND VALIDITY GIVEN VARIOUS WEIGHTS ASSIGNED TO THE MBE



Reliability of MBE = 0.90 Reliability of essay test = 0.72 Correlation = 0.60

is increased, the high reliability of the MBE score pushes the overall validity values higher.

The dashed line in Figure B labeled Validity 2 shows the validity curve that would result if you believe that the written component is twice as important as the MBE. Even in that case, the need for a valid score to also be reliable pushes the validity curve upward as the MBE weight is increased. Note that if you believe that the written portion of the test is twice as important, the curve suggests that you should weight the MBE scores less than if you believe that the two components (MBE and written) are equally important. However, under both conditions, validity is optimal with at least a 50% weight given to the MBE scores.

Figure C demonstrates the balance that must be achieved between reliability and validity. Regardless of your beliefs about the relative importance of the MBE and the written portions of the test, the data indicate that the MBE scores should be weighted at least 50%. Because the MBE measures underlying legal knowledge and skills that are also reflected in the written components and measures these skills so much more reliably, MBE scores provide a critically important part of the total measure of legal skills.

As a result of the analyses, NCBE recommends that each jurisdiction scale essay and performance test scores to the MBE, weight the MBE score at least 50%, combine the two scores, and make a pass/fail decision on that single combined score.

A final thought on the lakes example: What did the walleye say when he ran into a wall? "Dam!"

ENDNOTE

1. Kane, M. & Case, S., *The Reliability and Validity of Weighted Composite Scores*, APPLIED MEASUREMENT IN EDUCATION 2004, Vol. 17, No. 3, pp. 221–240.

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