The Reliability of American Board of Family Medicine Examinations: Implications for Test Takers

Kenneth D. Royal, PhD, and James C. Puffer, MD

A common theme among family physicians who have performed poorly repeatedly on the American Board of Family Medicine (ABFM) Maintenance of Certification (MC-FP) examination is the complaint that they received a score that was identical, or almost identical, to their score on a previous administration of the examination. From their perspective, why they received the exact same score (or a very similar score), despite additional study time and preparation, is a mystery. Often, physicians assume a mix-up has occurred and ask if it is possible that results have been provided erroneously from their previous attempt. After a psychometric review, it is clear that there is no mistake at all. In fact, we anticipate many test takers will receive a comparable score on future successful attempts at taking the examination. We base this anticipation on the psychometric concept of reliability.

Overview of Reliability

The notion of reliability is perhaps one of the oldest yet most misunderstood notions in the measurement and assessment arena. Researchers of all experience levels commonly assert that their instruments are reliable. The truth is there is no such thing as a reliable instrument. Only the scores produced from an assessment have the property of reliability. All tests are dependent on the characteristics of the test, the test administration, and the group of examinees. It is the interaction among these 3 elements that determine the reliability of results for any test.

Let us briefly discuss each of the 3 major elements. Test characteristics typically include test length, item type, and item quality. Generally speaking, longer tests produce more reliable scores than shorter tests. With regard to item type, objective items such as multiple-choice questions typically produce more reliable scores than subjective items such as essays. Item quality is also important because poor quality items tend to reduce reliability. Also, good-quality items should vary sufficiently in difficulty so that they effectively discriminate among examinees. Discrimination is useful in that it helps identify which examinees possess the knowledge necessary to answer an item correctly. Those who possess the most knowledge will have the greatest probability of answering difficult items correctly. Over the course of a lengthy examination, distinctions between examinees become clearer, and we are better able to determine how much knowledge an examinee possesses.

Conditions of administration also are important. These include physical conditions (eg, temperature levels and noise in the testing room), examination instructions, and time limits. Our testing vendor goes to great lengths to ensure these factors remain as constant as possible across multiple administrations of our examination. Variation in these conditions could affect some examinees differently and result in scores that vary for reasons other than an examinee knowing more or less about the content. The ABFM acknowledges that disruptions such as excessive noise or other distractions can introduce additional error into one's score, thus potentially invalidating results. We have policies in place to rectify situations when this occurs. However, other administration factors such as instructions and time limits are imposed equally on everyone, unless a disability is documented, in which case extra time

Submitted 17 October 2011; revised 17 October 2011; accepted 19 October 2011.

From the American Board of Family Medicine, Lexington, KY.

Funding: none.

Conflict of interest: none declared.

Corresponding author: Kenneth Royal, PhD, 1648 McGrathiana Parkway, Suite 550, Lexington, KY 40511 (E-mail: kroyal@theabfm.org).

and possibly other accommodations may be permitted.

Finally, the characteristics of the group of examinees are important. As mentioned earlier, a good test should contain a considerable number of items with varying degrees of difficulty. But what happens when a good test is attempted by a very homogenous sample, say, a group of high achievers with similar levels of knowledge? Although the test may be sound psychometrically, the sample of examinees varies so little that scores cannot be differentiated reliably. When this happens, low reliability estimates are produced and many researchers quickly dismiss the instrument (or assessment) as being of poor quality. It is for this reason that reliability estimates are not *the* measure of examination quality, but rather *a* measure of examination quality. For a test to produce reliable scores, the ability of examinees also must vary sufficiently. When there is a great range of ability in a group, reliable distinctions between what an examinee knows and does not know can be made.

Empirical Example and Interpretation

Although no strict guidelines for minimum levels of reliability exist, many measurement experts tend to agree with Nunnally and Bernstein's[1] recommendations. That is, the minimum reliability necessary for a group of test scores is 0.90 if important decisions are going to be made based on those scores. Reliability estimates between 0.80 and 0.89 are considered reasonably reliable. The 2009 ABFM MC-FP examination had a reliability estimate of 0.94. This is considered quite a high estimate of internal consistency. This estimate indicates that an estimated 94% of the observed variance in scores is caused by systematic differences in examinee performance, with 6% due to chance differences. Another way to interpret this estimate is to consider perfect reliability (1.0) minus the observed reliability (0.94). The difference, in this case 0.06 (or 6%), is the amount of observed variance that is caused by measurement error.

Implications for High-Stakes Testing

In many ways high estimates of reliability essentially echo the old adage to test takers: "If you always do what you've always done, you'll always get what you've always gotten." For an examinee who has a history of scoring very high on the examination, this notion will typically work in the examinee's favor. However, it should be made abundantly clear that this is not a guarantee. On the other hand, test takers who have failed an examination previously may find this news disconcerting. This is not to say that one is not capable of making such gains. With a significantly improved approach to examination preparation, most examinees that have failed previously are capable of making the types of gains necessary to pass this examination. It all begins with asking the right question and preparing an effective study plan.

Examinees should not ask themselves, What do I have to do to reach the minimum score necessary for passing?, but rather, How can I become a more knowledgeable physician? For physicians whose goal is to simply pass the test, their intentions, and possibly preparation strategy, are misguided. One's goal should not be to pass the examination, but rather to become a better family physician. With an increased fund of medical knowledge, the chances of passing the examination will improve naturally as a result of actual learning. However, if one's goal is to simply receive a passing score, then the examinee likely will find him or herself in the position of trying to anticipate examination items and otherwise resorting to methods similar to "cramming." Spending exorbitant amounts of time and energy in an attempt to memorize content solely for the purpose of regurgitating it at a later time, or working on improving one's test-taking skills with regard to identifying distracters, do not work well on a high-stakes, criterion-referenced examination such as ours that measures one's fund of medical knowledge.

As we have demonstrated, simply being a good test taker is not likely to improve significantly one's chances of passing a high-stakes certification examination.[2] Also, the scoring methods used for our exams work in such a way that one's ability is estimated based on correct/incorrect responses to items of varying degrees of difficulty. When both person ability and item difficulty are mapped onto a single continuum, it becomes clear from a psychometric perspective what an examinee knows and what he or she does not know.[3] Therefore, only when a physician has taken an improved approach to examination preparation, particularly one that focuses on increasing one's fund of medical knowl-

edge, can he or she seriously expect to advance along that continuum of ability.

Conclusion

It is important to emphasize clearly and directly that an examinee of marginal ability or someone with a history of previous failures is likely to continue to fail the MC-FP examination if he or she continues with the same preparation approach or otherwise utilizes study preparation methods that do not solicit actual and sustained learning. Improving test-taking skills will be of minimal benefit to a test taker because high-stakes examinations are not a measure of test-taking skills. The MC-FP examination is constructed in such a way that the influence of test-taking skills is negligible. Examinees should understand that the only legitimate way to improve one's performance on the MC-FP examination is to increase their fund of medical knowledge and decision-making ability in clinical scenarios—that is what the examination measures. When examinees make real gains in improving these, they are most likely to receive higher scores. It should be noted that the ABFM provides important information on its website about its exams that is intended to help the family physician understand both the type and amount of content he or she might expect to see, as well as tips for developing a study plan.[4] Using this information can assist with improving performance on our examinations.

References

1. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill; 1994.
2. O'Neill TR, Royal KD, Puffer JP. Performance on the American Board of Family Medicine certification examination: are superior test taking skills alone sufficient to pass? J Am Board Fam Med 2011;24(2):175–80.
3. Linacre JM. KR-20 or Rasch reliability: which tells the "truth"? Rasch Measurement Transactions 1997:(11)3:580–1. Available at http://www.rasch.org/rmt/rmt113l.htm. Accessed November 18, 2011.
4. American Board of Family Medicine. Examination descriptions. Available at https://www.theabfm.org/moc/exams.aspx. Accessed November 18, 2011.