Understanding the "Sum of Subtest to Overall Score Discrepancy" on the Maintenance of Certification-Family Practice Examination

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

When high-stakes examinations, such as the American Board of Family Medicine's (ABFM's) Maintenance of Certification-in Family Practice (MC-FP) examination, are administered, candidates and diplomates are keenly interested in the accuracy of their test scores, especially when their scores are close to, but below, the pass/fail cutpoint. In some instances, candidates will attempt to reverse engineer their scores using the information provided on the score report in an effort to verify that the "weighted sum of the subtest scores" is congruent with the overall test score. Any discrepancy might become alarming to the candidate, providing a seemingly legitimate reason to believe the overall score was inaccurate, thus prompting a phone call to the ABFM for further investigation and clarification. Historically, such a mistake in scoring has never been found; however, a statistical phenomenon that we will describe below could make it appear so. We would like to explain this phenomenon so that examinees who attempt to reverse engineer their score reports will better understand the "sum of subtest to overall score discrepancy" phenomenon.

Sum of Subtest to Overall Score Discrepancy

Sometimes when examinees attempt to reverse engineer their score reports, the weighted sum of the subtest scores will be higher than the total score. When the reverse happens, we generally do not receive a phone call. For example, some candidates may find the weighted subtests add up to a scaled score of 400 when the overall scaled score was 380. Because the current minimum passing standard is 390, candidates who experience this phenomenon may question the validity of the overall score and ultimately the pass/fail decision. Here, we will attempt to explain (albeit briefly) this rather technical statistical phenomenon.

Diplomates typically view scores as quantities that have additive properties. For instance, in the past, the ABFM presented raw scores on the score report. If one were to add the weighted raw scores of the subtests, the scores would certainly equal the raw score of the total test. Unfortunately, raw scores are not measures. Although raw scores are useful for descriptive purposes, they lack generality because they are specific to the particular test that was taken. Raw scores are counts and are deterministic and exact, but the measures they imply are probabilistic and have some degree of imprecision. The ABFM employs the Rasch model,[1] a 1-parameter item response theory measurement model, to score examinations. The Rasch model converts raw scores to linear measures and controls for the difficulty of the test version one received.

In some instances, the weighted sum of the subtests scores (as determined by the item response theory scoring method) will be greater than the overall score. The primary reasons for this are 2-fold. First, score exchanges have asymmetric nonlinearity. That is, within-person variation increases on subtest areas, making the distribution of subtest measures wider than the distribution for the overall test. This can often make mean measures appear larger. Second, there is an increase in measurement error because of the small number of items available in each subtest area. Consequently, the increase in measurement error also inflates measure variance, thus causing even more inferential instability. It is for these reasons that we report

Submitted 5 January 2012; accepted 5 January 2012.
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).

a standard error with each measure on the MC-FP score report, and it assists the examinee in understanding the stability of each particular measure. For a more detailed discussion on the topic of summing subtest measures, readers are encouraged to refer to Wright.[2]

Additional Insights and Recommendations to Test Takers

What does the statistical phenomenon presented above mean to persons who take the MC-FP examination? First, examinees should know that only the overall scaled score is used to determine the pass/fail decision. This score is based on one's cumulative performance on 350 items; thus the results will be both highly precise and statistically stable. Therefore, subtest scores should be viewed simply as good approximations of one's performance in a particular clinical content area because these scores are often highly unstable because of the limited number of items and larger standard errors.

Next, test takers should be aware that extreme subtest scores are not uncommon because there are a limited number of items in each subtest area. This may cause additional problems with regard to interpretation. For example, the ABFM's reported range for scores is 200 to 800. It is possible that scores actually may be well below 200 or far greater than 800, but in such instances scores are rounded back to fit the range of the scale. Therefore, in most instances in which candidates find that their weighted sum of subtest scores do not equal that of the total test, extreme scores likely are the primary culprit. Examinees who attempt to reverse engineer the score report should be particularly mindful of extreme scores and how scores of 200 or 800 are not necessarily indicative of a true 200 or 800 score.

Examinees also need to be aware that some granularity exists with the reporting of scores. The MC-FP examination provides truncated scores that are reported in increments of 10. For example, an examinee who truly scored a 507 would see a reported score of 500. Although detailed scaled scores are used in the calculation of scores, only truncated scores are reported. This largely is for purposes of clarity and simplicity. However, in no instance is an examinee's score rounded up because all test-takers are expected to meet or exceed a particular passing threshold. Subtle nuances such as these also can have some bearing on the impact of subtest score summations.

Conclusion

It is important to emphasize that only the overall scaled score is used to determine the pass/fail decision on the ABFM MC-FP examination. Subtest scores are less stable because of the fewer number of items and the larger standard errors. Despite the instability of subtest scores, a good bit of inferential value can be gleaned from this information; subtest scores serve as useful approximations for one's performance in various clinical categories on the MC-FP examination. Instances in which examinees attempt to reverse engineer their scores based on the information presented in the blueprint likely will prove to be unproductive because of the statistical phenomenon discussed earlier. Rather than attempting to make the case that one's score result should be corrected because the sum of one's weighted subtest scores is not congruent with the overall score, candidates instead are encouraged to use their subtest scores to improve their medical content knowledge by developing an improved self-directed learning plan, thereby increasing their likelihood of future success on the examination.

References

1. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press; 1960.
2. Wright B. Combining part-test (subtest) vs whole-test measures. Rasch Measurement Transactions 1994;8:376.