Dimensionality of the Maintenance of Certification for Family Physicians Examination: Evidence of Construct Validity

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

The American Board of Family Medicine (ABFM) Maintenance of Certification for Family Physicians (MC-FP) examination is designed to measure a single construct: clinical decision-making abilities within the scope of practice of family medicine. Implied in the construct of clinical decision-making abilities is the ability to recall relevant elements from a large fund of pertinent medical knowledge. While clinical decision-making abilities could be perceived as comprising several separate constructs (based on, for example, clinical categories or organ systems), that approach would require the development of multiple assessment scales with a passing criteria specific to each. Instead, the overarching construct of clinical decision-making ability, which encompasses those more specific areas, has been selected by the ABFM because it more closely mirrors the pass/fail decision process used to discern which candidates receive certification. In any instance, the construct that the ABFM attempts to measure needs to be sufficiently unidimensional to produce precise, error-free estimates of a candidate's performance. This brief article will discuss the dimensionality of the MC-FP examination and its implications for construct validity, namely the validation that the examination in fact accurately measures the ability of family physicians to make appropriate clinical decisions.

Dimensionality

Why is dimensionality important? Simply put, it is desirable to measure only one thing at a time. Just as physical measurements attempt to measure one thing at a time (eg, a patient's blood pressure reading should not be biased by his or her height, weight, or sex), psychometricians, the measurement experts that help design the ABFM's examinations, also aspire to measure one latent trait at a time. It is only when dimensions are clearly isolated that one can understand the meaning of the measure and make a valid inference about an examination score.

Conflict of interest: The authors are from the ABFM.

Dimensionality of the MC-FP Examination

As we have mentioned previously, the psychometric model that the ABFM employs to score its examinations is the Rasch model, a one-parameter item response theory measurement model. The Rasch model converts raw scores to linear measures and controls for the difficulty of the version of a test a candidate received.[1] In addition to using typical fit indicators, the most effective way to detect multidimensionality in the analysis of data based on Rasch measurements is to use a principal components analysis (PCA) of standardized residual correlations.[2] In short, the Rasch model uses ordinal data to construct a one-dimensional measurement system. Of course, real data are never perfectly unidimensional, so the presence of more than one latent dimension in the data always exists to some extent. When the data perfectly fit the Rasch model (this includes all items and persons examined), all systematic variation is explained by a single dimension. Data that are not in perfect accord with the model leave behind residuals that have a random normal structure and predictable variance.[2]

To evaluate the dimensionality of the MC-FP examination, we perform the aforementioned industry standard tests of fit and PCA of standardized residual correlations. An investigation of how the data fit to the model, both overall and by individual item analysis, can help us discern whether multiple dimensions are present and exactly where in the dataset these dimensions might be. To demonstrate this, let us share an analysis we performed using the core portion of the 2010 examination. The dataset included 3697 examinees and the 423

test items that appeared across the multiple forms of the core portion of the MC-FP examination. Fit statistics indicated perfect overall data-to-model fit, with infit and outfit mean square statistics of 1.0 for both persons and items. Values of 1.0 are ideal for these analyses,[3] and the acceptable range is between 0.80 and 1.20.[4] Fit statistics for individual items then were evaluated. Only8 of 423 items deviated from the ideal range. The most overfitting item had a mean square value of 1.27, and the most underfitting item had a mean square value of 0.77, meaning that less than 2% of the items appearing on the MC-FP examination had fit statistics that fell outside the ideal range for dichotomous data. These statistics indicate excellent item fit with minimal off-variable noise.

Next, the slight noise that was detected in the measures was evaluated using a PCA of standardized residual correlations. The candidates who complete the MC-FP examination each year are quite homogeneous: they are highly educated physicians with expertise in family medicine. Therefore, a great deal of variability across person measures (mean score, 469; standard deviation, 98) and item measures (mean score, 297; standard deviation, 168) does not exist, considering the reported range of scores is from 200 to 800. This lack of variation naturally leads to an inability to explain a great deal of the variance.[5] Data from this MC-FP examination explained just 11.2% of the variance. The test items explained the vast majority of variance (7.5%). The strongest secondary dimension detected explained 1.2% of the variance. The ratio of the overall primary dimension and the secondary dimension was 11.2:1.2; the ratio of the primary item dimension and the strongest secondary dimension was 7.5:1.2. These ratios are accepted universally in the measurement literature as being sufficiently unidimensional.[6,7]

From a dimensionality perspective, the most polarizing items that appeared on the examination were identified by the PCA analysis and reviewed by content experts. The nature of these items pertained to issues of prevention at one extreme and issues of treatment at the other. The items underwent a psychometric evaluation, and all psychometric indicators confirmed the items functioned properly and were indeed good, quality items. Family physicians are expected to be knowledgeable about both the prevention and treatment of illnesses, and therefore the substantive nature of the detected secondary dimension seemed to be rather inconsequential.

Conclusion

The MC-FP examination is intended to measure the single construct of clinical decision-making ability within the practice of family medicine. Results of the dimensionality analysis described above indicated that the MC-FP examination is highly unidimensional from a psychometric perspective. That is, the data accorded well with the model's expectations and the internal structure of the data was correlated in such a way that the same construct was being measured consistently throughout the examination. Review of the substantive content of polarized dimensions by experts provided additional assurance of the unidimensional nature of the examination.

What do these results mean with regard to the validity of examination scores? Renowned measurement scholar Samuel Messick[8] conceptualized construct validity as a uniform concept that required multiple pieces of evidence. He identified 6 aspects of construct validity: content, substantive, structural, generalizable, external, and consequential. When evaluating the results of the analysis of our examination from Messick's framework, psychometric evidence is available that speaks to the content, substantive, and, in a limited way, structural aspects of construct validity. We previously have provided some evidence that speaks to the generalizeable aspect of validity as well.[9] Collectively, these results should be reassuring for candidates because they provide additional evidence of the psychometrically sound nature of the MC-FP examination. Of course, test takers also should be assured that the MC-FP examination yields valid inferences about their scores as well.

References

1. Royal KD, Puffer JC. Understanding the "sum of subtest to overall score discrepancy" on the maintenance of certification for family physicians examination. J Am Board Fam Med 2012;25:260–1.
2. Linacre JM. Detecting multidimensionality: which residual data-type works best? J Outcome Meas 1998;2:266–83.
3. Linacre JM, Wright BD. Dichotomous mean-square infit and outfit Chi-square fit statistics. Rasch Meas Trans 1994;8:360.
4. Wright BD, Linacre JM. Reasonable mean-square fit values. Rasch Meas Trans 1994;8:370.
5. Linacre JM. Variance in data explained by Rasch measures. Rasch Meas Trans 2008;22:1164.
6. Linacre JM. Data variance: explained, modeled and empirical. Rasch Meas Trans 2003;17:942–43.

7. Linacre JM. Dimensionality: contrasts and variances. Winsteps help for Rasch analysis. 2011. Available from http://www.winsteps.com/winman/index.htm?principalcomponents.htm. Accessed March 2, 2012.
8. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psych 1995;50:741–9.
9. Royal KD, Puffer JC. The reliability of American Board of Family Medicine examinations: implications for test-takers. J Am Board Fam Med 2012; 25:131–3.