The ABFM Begins to Use Differential Item Functioning

Thomas R. O'Neill, PhD, Michael R. Peabody, MS, and James C. Puffer, MD

The American Board of Family Medicine (ABFM) believes that it is important to have evidence to show that the pass/fail decisions related to its examinations are based on accurate determination of the minimum knowledge necessary to be a board-certified family physician and, furthermore, that these decisions are unbiased against any particular subset of the population. Accordingly, as part of the ABFM's commitment to continuously improve the Maintenance of Certification for Family Physicians (MC-FP) process, the ABFM has started using differential item functioning (DIF) procedures to detect potentially biased items on its examinations. Although data on examination applicants' gender has been collected for some time, in the spring of 2013 we began collecting ethnicity data from applicants taking the MC-FP examination so that we could begin to conduct these analyses.

DIF procedures are based on the idea that a test item is biased if individuals who have equal ability but are from different subpopulations do not have the same probability of answering it correctly.[1,2] Although pass rates are an indicator of whether a particular subpopulation is performing at a level comparable to other subpopulations, it is silent with regard to whether the meaning of the scores is stable across subpopulations. These differences could be due to bias in the items that would effectively destabilize the construct.[3] By this we mean that the items, when ordered by difficulty, form a linear construct of less difficult to more difficult. If some items are more difficult or less difficult relative to the other items for a specific subpopulation, then the construct represented by the test is degraded to the extent that the items are disordered for that subpopulation. On the other hand, the hierarchical construct represented by the test could be stable and the difference in pass rates could be due to differences in socioeconomic status and the potential associated inequities inherent in the educational system. DIF analysis permits us to disentangle item-level bias from differences in ability among subpopulations.

The process of calibrating test questions with regard to their difficulty for samples from both a subpopulation and the overall population is probabilistic. Therefore, this type of DIF study is best used as a screening tool to find biased items. It does not prove that the items are biased. The ABFM DIF process can be viewed in 3 stages: (1) flagging potentially biased items, (2) examining the content of the flagged questions for sources of bias, and (3) determining their final disposition.

Flagging Items

The particular method of DIF detection used by the ABFM is based on the dichotomous Rasch model.[4–6] Using this method, the items are calibrated twice: first using only responses from members of the reference group and next using only responses from members of the focal group. Because the largest self-reported ethnicity among ABFM diplomates is white, the ethnicity reference group is white and the focal groups are the other ethnicity categories. Using this same reasoning, the reference group for sex is male and the focal group is female. Although the fine-tuning of this method to meet the needs of the ABFM is still being developed, the process will largely reflect the procedure described below.

For each item, the 2 calibrations are compared. If the 2 calibrations fall outside of the 95% confidence interval for their mean, then the item is flagged as potentially biased. Please note that the potential bias could be to the advantage or the disadvantage of the focal group. In addition,

Conflict of interest: The authors are from the ABFM.

when using this flagging criterion, it is expected that approximately 5% of the items will be flagged just by chance. Although the criteria could be made more stringent to reduce the number of false positives, it also would reduce the number of false negatives, potentially permitting some biased items to go undetected. The 95% confidence interval seems to be reasonable for use as an initial screening criterion. All items that are flagged as potentially biased in either direction are forwarded to the DIF Review Panel for evaluation. Over time, the screening criteria will likely be better optimized.

Convening a DIF Review Panel

The DIF Review Panel is convened once a year to review the content of items that have been flagged for potential bias. The panel comprises subject matter experts (ABFM diplomates) who represent a diversity of ethnicities and both sexes. The panel also includes a linguist and is moderated by a psychometrician. The panel meeting begins with an explanation of DIF as a concept and the purpose of the panel. The panel is charged with the responsibility of reviewing items for appropriateness for the examination with regard to DIF. The panel may decide that there is no identifiable content that caused the DIF and permit the item to stand. On the other hand, the panel may decide that there is an identifiable source of DIF. If so, the panel must determine whether that source of DIF is related to an important aspect of family medicine. If it is important, then the panel is to let the item stand. If it is not important, then the panel should recommend that the item be deleted or reworked. The items that the panel recommends deleting or reworking are forwarded to the ABFM Examination Committee.

Determining the Final Disposition of the Items

The Examination Committee reviews the recommendations of the DIF panel and makes a final decision on whether an item is sent back to the ABFM content development department for revision/deletion or is permitted to stand. To send the item back for revision/deletion, the Examination Committee should concur that there is likely something in the item causing the difference in relative difficulty that is not an important aspect of family medicine. Of course, the Examination Committee can always send an item back to be reworked or deleted and the reason need not be limited to DIF issues; however, the Examination Committee review is the final step in determining the disposition of an item.

Summary

To defend against claims of discrimination, the certification and licensure testing industry routinely uses DIF to detect items that function differently for protected classes.[7] While most other American Board of Medical Specialty boards are not yet collecting this information, the ABFM has begun collecting ethnicity data from candidates applying for its examinations so that this kind of bias can be detected. The industry generally regards this type of analysis as a best testing practice that makes the meaning of the examination results more stable across subpopulations.[8] Documentation of these processes also can be used to show that a test publisher has made a diligent effort to minimize or eliminate sources of irrelevant variance that might have detrimental effects on subpopulations of interest.

On a final note, it is important to underscore that the ABFM does not release ethnicity information to external parties. Furthermore, ethnicity and sex are not used to determine the difficulty of test items with regard to scoring the examination. The operational item calibrations that are used for scoring are based on responses from the entire group, not a particular ethnicity or sex reference group. There are not different passing standards or different scales for the different ethnic groups or sexes: there is only one scale with a single passing standard that applies.

References

1. Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates;1980: 212.
2. Angoff WH. Differential item functioning methodology. In: Holland PW, Wainer H, eds. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates;1993: 3–23.
3. Suen HK. Principles of test theories. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990: 186.
4. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

5. Luppescu S. Graphical diagnosis. Rasch Meas Trans 1991;5:1–136.
6. Linacre JM. A User's Guide to Winsteps version 3.68.0. Available from: http://www.winsteps.com/index.htm. Accessed January 17, 2011.
7. McAllister PH. Testing, DIF, and public policy. In: Holland PW, Wainer H, eds. Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum Associates; 1993: 389–96.
8. Standards for educational and psychological testing. 5th ed. Washington, DC: American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education;1999: 81.