"These recommendations demonstrate the ability of our Academy and others to look at evidence that may go against some of the established perceptions out there," Blackwelder said. "And while they are obviously not absolutes, owing to the fact that we treat individual patients, they are good evidence-based guidelines."

"For PSA screening in men without symptoms, the data is extremely clear that the test provides very little benefit for patients, along with a significant risk of harm from the diagnostic procedures and the treatments that are performed," he said. "Similarly, in terms of oral contraceptives to women, the data is very clear that unwanted pregnancy carries a much higher risk than the use of these various medications, as well as the fact that pelvic exams and other evaluations are really not necessary before prescribing."

To date, more than 50 medical specialty organizations have joined the effort, identifying a list of more than 160 tests and procedures physicians and their patients should question. Other lists will be released throughout 2013 and 2014.

*Matt Brown*
AAFP News Now

**From the American Board of Family Medicine**

*Ann Fam Med* 2013;578-579. doi:10.1370/afm.1587.

## THE ABFM BEGINS TO USE DIFFERENTIAL ITEM FUNCTIONING

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 upon accurately determining 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 gender information has been collected for some time from examination applicants, we began collecting ethnicity data for applicants taking the MC-FP exam this past spring so that we could begin to conduct these analyses.

DIF procedures are based upon the idea that a test item is biased if individuals from different subpopulations, who are of equal ability, 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 the 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 their difficulty, form a linear construct of less to more. 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 very stable and the difference in pass rates could be due to differences of 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, both for samples from a subpopulation and from 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 as having 3 stages: (1) flagging potentially biased items, (2) examining the flagged questions' content 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 considered to be white and the focal groups are the other ethnicity categories. Using this same reasoning, the reference group for gender is male and the focal group is female. Although the fine tuning of this method to meet the needs of 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. Also, 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 would also 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 were flagged for potential bias. The panel is composed of subject matter experts, ABFM diplomates, who represent a diversity of ethnicity and gender. 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 they permit the item to stand. On the other hand, the panel may decide that there is an identifiable source of DIF. If there is, the panel must determine whether or not 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 Items Final Disposition

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 whether the item 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 differential item functioning (DIF) to detect items that function differently for protected classes.[7] While most other American Board of Medical Specialties boards are not yet collecting this information, the ABFM has begun to collect ethnicity data from candidates applying for its examinations so that this kind of bias detection can be performed. 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] Also, documentation of these processes 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 gender are not used to determine the difficulty of the 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 gender reference group. There are not different passing standards or different scales for the different ethnic groups or genders. There is only 1 scale with a single passing standard that applies.

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

## 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; 1993:4.
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, and Chicago, IL: University of Chicago Press; 1980.
5. Luppescu, S. Graphical diagnosis. *Rasch Meas Trans.* 1991;5(1):136.
6. Linacre JM. Winsteps: Rasch measurement computer program, version 3.68.0. http://www.winstepscom/index.htm. Accessed Jan 17, 2011.
7. McAllister PH. Testing, DIF, and public policy. In: Holland PW, Wainer, H, eds. *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates; 1993:389-396.
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.