COMMENTARY

A Linguist's Perspective on the American Board of Family Medicine's Differential Item Functioning Panel

Jennifer Cramer, PhD

As part of their continuing efforts to create higher parity levels in the Family Medicine Certification Examination, the American Board of Family Medicine has established procedures to explore bias in certification examinations by establishing a differential item functioning (DIF) analysis process and panel review. The review panel consists of a diverse group of family physicians and a linguist who is charged with determining whether items from the examination contain bias unrelated to the practice of medicine. It is the objective of this commentary to explain the panel process itself and to promote the inclusion of a linguist in similar panels. I argue that the inclusion of a linguist on a DIF panel can aid in determining where language itself is the source of bias. (J Am Board Fam Med 2022;35:387–389.)

Keywords: **Certification, Continuing Medical Education, Family Medicine, Linguistics, Psychometrics**

As part of their continuing efforts to create higher levels of parity in the Family Medicine Certification Examination (FMCE), the American Board of Family Medicine (ABFM) has established procedures to explore bias in certification examinations through the establishment of a differential item functioning (DIF) [1–3] analysis process. DIF is "a collection of statistical methods utilized to determine if examination items are appropriate and fair for testing the knowledge of different groups of examinees."[4] The goal of this process is to identify questions on the FMCE that indicate potential bias based on the race/ethnicity and sex of the examinee. While there are certainly other factors[5–7] that might lead a subgroup of family physicians to perform better or worse than another, it is necessary to also explore how the examination questions themselves, separate from those other factors, might contribute to this difference. The full approach has been described,[2] and the results of the first 8 years of DIF analysis show that very few questions on the FMCE had to be reworked or deleted because there was an identifiable source of bias that was not related to family medicine in an important way. This study also showed that the number of questions that showed potential bias was small and that the number of questions that advantaged the reference group was nearly the same as the number that advantaged the focal group, which suggests that individuals across these categories are generally not disadvantaged.

After items have been identified for potential bias using a DIF flagging procedure, a DIF Review Panel is convened. This review panel consists of a diverse group of family physicians and a linguist who is charged with determining whether these flagged items contain bias unrelated to the practice of medicine. A linguist is an expert in the structure and use of language; many linguists specialize in the analysis of lexemes, syntactic structures, and sound systems without regard to context, but it is a sociolinguist (or other applied linguists, like linguistic anthropologists), one who engages in the exploration of the two-way impact of language and society, who examines how such linguistic structures operate in real-world contexts. It is the objective of this commentary to explain the panel process itself and to promote the inclusion of a (socio) linguist in similar panels. I argue that the inclusion of such a linguist on a DIF panel can aid in determining where language itself is the source of bias.

The most recent DIF Review Panel was convened in July 2021. This panel was responsible for reviewing both the 2020 and 2021 FMCE flagged questions due to the COVID-19 pandemic, which precluded the convening of this panel in summer

This article was externally peer reviewed.
Submitted 19 August 2021; revised 17 November 2021; accepted 19 November 2021.
From the University of Kentucky, Lexington (YJC).
Funding: None.
Conflict of interests: None.
Corresponding author: Jennifer Cramer, PhD, 1615 Patterson Office Tower, Lexington, KY 40506-0027, (E-mail: jennifer.cramer@uky.edu).

2020. The typical procedure involves convening the panel at the ABFM headquarters in the morning, conducting introductions, completing paperwork related to the panel (eg, photograph release, expense forms, nondisclosure agreements), and introducing the FMCE and the psychometric procedures used to flag the questions. After this brief orientation, panelists begin the process of reviewing the items.

Panelists review an item for a comparison between the reference group (white or male) and 1 of the several focal groups (Asian, Black, Hawaiian/Pacific Islander, Hispanic, Native American, and female). An item may be flagged across multiple group comparisons. For each comparison, the following information is included: the stem of the question, possible response options, the correct answer, the difficulty of the question for the reference group, the difficulty of the question for the focal group, the percentage of people who selected each response option across subgroups, and a critique (including relevant references) explaining why the correct answer is correct. Panelists attempt to identify if a specific source of bias is present. If no identifiable source of bias is found, then flagged items are retained. If an identifiable source of bias is determined, panelists then determine whether it is an important aspect of family medicine. If it is an important aspect, then flagged items are also retained. However, if the bias is determined to be unrelated to an important aspect of family medicine, these questions are referred to the Knowledge Assessment Committee (KAC) with the recommendation that they be reworked or deleted. For example, if a question depended on one's ability to correctly define the word *spelunking* as "cave exploring" to identify the cave and its likely inhabitants as part of the diagnosis (as in rabies or histoplasmosis), such a question would be recommended for revision if Hispanic physicians as second language learners of English were disadvantaged in the DIF analysis. It should be noted that, in this case, the category Hispanic is used as shorthand for a non-native speaker of English but knowing the language ability of the test taker could be valuable information to have going forward. Many physicians who choose Hispanic are likely native speakers of English, and there are certainly people who select any of the categories who may be non-native speakers.

The role of the linguist on the panel is to explore how the linguistic structure of the flagged items may introduce bias. For example, it might be useful for the linguist to suggest that the use of a term like *tennis elbow*, instead of the more technical term *lateral epicondylitis*, is problematic because use of medical jargon, or words specific to the domain of medicine, is more appropriate than use of the lay term for an examination. Beyond words, the issue may have to do with the ways in which the responses have been posed. If the answer to a question says, "not X and Y," it is possible to interpret such a response in 2 ways: "not X and not Y" or "not X and yes Y." If the inclusion of this answer distracts a specific subgroup who chose the erroneous interpretation, the question should be referred to the KAC for rewording.

In addition to these structural considerations, a (socio)linguist can provide valuable insight about the cultural use of language in questions and answers on the FMCE. Some questions provide additional information about the patient (eg, occupation, hobby); if that information causes an examinee to choose an incorrect answer based on cultural assumptions about what such a patient might request or suggest (such as having a patient be described as a medical professional, in which case the practitioner may feel the need to perform extraneous tests and procedures to appease the patient due to the occupational information provided), it would be important to address the concern with the KAC to explore why this information was included in the example. It is possible that providing this information results in unanticipated privilege, and it would be important to recognize if potential biases were present for such items.

While it is possible that other professionals (such as librarians) could be appropriate for this role, a linguist—especially one with expertise in the role of society in language use—provides a specific skill set that others may lack. The linguist sees the questions from a non-specialist perspective, which allows them to suggest that certain linguistic structures are less innocuous than they might think. Overall, linguists look at language data from a scientific perspective, seeing the structures as they exist both with and without context. This viewpoint is different from that provided by practicing medical professionals, and it is the combined knowledge of these groups that can provide the most well-rounded view of the DIF on an examination. Furthermore, if linguists were involved at earlier stages, perhaps in the question creation stage, those items where structural ambiguity or cultural information causes confusion might be avoided. The process itself is important for the medical community, and I hope that other boards will

consider instituting DIF quality control procedures as part of their standard operating procedures.

Special thanks to Tom O'Neill and Ting Wang at the American Board of Family Medicine for their comments on previous drafts of this commentary. Any errors in the description of the processes of the American Board of Family Medicine are my own.

To see this article online, please go to: http://jabfm.org/content/35/2/387.full.

References

1. O'Neill TR, Peabody MR, Puffer JC. The ABFM begins to use differential item functioning. J Am Board Fam Med 2013;26:807–9.
2. O'Neill TR, Wang T, Newton WP. The American Board of Family Medicine's 8 years of experience with differential item functioning. J Am Board Fam Med. In press.
3. Wainer H, Braun HI, eds. *Test validity*. Lawrence Erlbaum; 1988.3.
4. Perrone M. Differential item functioning and item bias: critical considerations in test fairness. *Columbia U. Working Papers in TESOL & App*. Ling 2006; 6:1–3.
5. Jencks C, Phillips M. eds. *The black-white test score gap*. Brookings Institute Press; 1998. Accessed August 11, 2021. https://www.brookings.edu/book/the-black-white-test-score-gap/.
6. Hinton I, Howell J, Merwin E, et al. The educational pipeline for health care professionals: understanding the source of racial differences. J Human Resources 2010;45:116–58.
7. Hauer KE, Jurich D, Vandergrift J, et al. Gender differences in milestone ratings and medical knowledge examination scores among internal medicine residents. Acad Med 2021;96:876–84.