ORIGINAL RESEARCH

The American Board of Family Medicine's 8 Years of Experience with Differential Item Functioning

Thomas R. O'Neill, PhD, Ting Wang, PhD, and Warren P. Newton, MD, MPH

Introduction: **Differential item functioning (DIF) procedures flag examination questions in which examinees from different subpopulations who are of equal ability do not have the same probability of answering it correctly. Few medical certification boards employ DIF procedures because they do not collect the needed data on the examinee's race or ethnicity. This article summarizes the American Board of Family Medicine's (ABFM) combined use of DIF procedures and an expert panel to review certification questions for bias.**

Methods: **ABFM certification examination data from 2013 to 2020 were analyzed using a DIF procedure to flag questions with possible ethnic or racial bias. The flagged questions were reviewed by a racially and ethnically diverse panel of content experts. If the panel judged the source of the DIF was not clinically relevant for the practice of family medicine, the question was removed from the examination.**

Results: **Out of the 3487 questions analyzed, 374 unique questions (11%) were flagged by DIF procedures as potentially biased. Of the flagged questions, the review panel felt 4 should be removed for fairness.**

Discussion: **Using DIF procedures and panel review can improve the quality of the board certification questions and demonstrate the organization's commitment to avoid racial or ethnic bias. (J Am Board Fam Med 2022;35:18–25.)**

Keywords: **Bias, Certification, Ethnic Groups, Medical Education, Minority Groups, Family Medicine, Psychometrics, Quality Control, Social Justice**

Introduction

The American Board of Family Medicine (ABFM) has decided to review its assessments of knowledge for bias as part of both the ongoing continuing quality improvement efforts and our health equity[1] initiative, which focuses on issues of fairness across categories related to geography, type of employment, poverty, gender identity, and under-represented minorities in medicine.

Bias in board certification can take many forms. There are selection biases related to who is eligible to be certified. The eligibility requirements such as graduating medical school and residency are dependent on the person's academic performance in high school and college, which are closely associated with parental socioeconomic status and educational background. The discrepancies across racial and ethnic groups in this regard are well documented[2] and clearly function as a selection bias. When these differences do not prevent eligibility, they may still accumulate, making performance in residency somewhat variable and perhaps creating obstacles to board certification. Through much of the educational pipeline, there are large differences in educational attainment across minority groups. Certification by a medical certification board is at the end of this pipeline.

In addition to selection biases related to educational institutions, there are also selection biases within those populations that health care seeks to recruit. Some economics research has suggested that the failure to recruit more Blacks into postbaccalaureate health care programs is because they perceive other alternatives, such as employment in business, as being more lucrative.[3]

Another type of bias can be related directly to the questions asked on a test. For example, if a question contained language or content that was

This article was externally peer reviewed.
Submitted 17 May 2021; revised 28 July 2021; accepted 3 August 2021.
From the American Board of Family Medicine, Lexington, KY.
Funding: The authors received no funding to conduct this research.
Conflict of interest: The authors are employees of the American Board of Family Medicine.
Corresponding author: Thomas R. O'Neill, PhD, 1648 McGrathiana Pkwy, Ste 550, Lexington, KY 40511 (E-mail: toneill@theabfm.org).

differentially difficult for different subgroups of examinees, then it might be considered biased, especially if the offending content was not relevant to what was being measured. Similarly, if the structure or format of the question stem, distractors, or instructions made a question differentially difficult across subpopulations, it could be considered biased. This article describes our review of the potential bias in the multiple-choice questions used for making pass–fail decisions on the ABFM's board certification examination. What follows summarizes our overall approach, the results of our analysis, and our proposed next steps.

Methods—Psychometric Rationale and ABFM Context

In licensure and certification, standardized tests are often used to provide some degree of assurance to the public that an individual has met specific standards related to a profession's scope of practice.[4] These tests should not be influenced by factors irrelevant to the profession. Such construct-irrelevant factors degrade the quality of inferences that can be made based on the examination scores. If the construct-irrelevant factors systematically advantage or disadvantage identifiable subpopulations, then there is a bias in the question.

The ABFM is committed to making the pass–fail decisions used on the Family Medicine Certification Examination (FMCE) be closely tied to the practice of family medicine and be unbiased for identifiable racial and ethnic subsets of the population. For this reason, ABFM conducts differential item functioning[5] (DIF) analysis to flag potentially biased questions and then has them reviewed by a diverse panel of subject matter experts for sources of potential bias.

The subtext of the problem is that even if there is a real and substantive difference in the pass rates across groups, pass rates are silent regarding the causes of the difference. The disparity might be attributable to differences in socioeconomic status, the inequities inherent in the US educational system, biases in the questions that are irrelevant to family medicine, a combination of these factors, or other factors. DIF analysis permits the investigator to disentangle question-level bias from differences in ability among subpopulations.

DIF Procedures

DIF procedures are based on the idea that a test question may be systematically biased if individuals from different subpopulations, who are of equal ability on the characteristic being measured, do not have the same probability of answering it correctly. DIF is a category of analytic techniques that are used to test this proposition. Although these techniques can be used on any identifiable subsets of examinees, they are commonly used in the standardized testing industry as quality-control checks to identify test questions that might advantage or disadvantage a legally protected class of people. Although DIF analyses were developed in the mid-1960s, since then, they have become more commonly used by professional test publishers. Generally, the medical specialty certification community has been reticent to collect race and ethnicity data, which has made it impossible to perform DIF analyses. In 2013, ABFM became the first medical specialty certification board to collect these data and routinely implement these analyses as part of the standard quality-control process.

DIF is generally an undesirable characteristic for an examination because it means that the test is measuring both the intended latent trait and some other irrelevant characteristic that is associated with group classification or membership. For example, if a question that includes the word "spelunker" is harder for Hispanic examinees than White examinees (after adjusting for the examinees' ability level), it might be that the change in difficulty is caused by knowing that spelunker means cave explorer. If this knowledge is not significantly related to the construct being measured, it is degrading the quality of the measure. Conversely, if a sickle-cell anemia question is easier for Black examinees than White examinees, it is likely that Black physicians are more knowledgeable about that condition than White physicians because it is more likely to occur to a member of their family or to 1 of their patients. On an examination of medical knowledge, it makes sense to keep the sickle-cell anemia question because it is congruent with the purpose of the examination.

Selection of a DIF Procedure

DIF procedures began to emerge around 1964[6] and many different procedures have been proposed since then. Despite the variety of proposed procedures, only three seem to have been widely adopted: Holland and Thayer's[7] adaptation of the Mantel–Haenszel[8] procedure, Wright and Panchapakesan's[9] procedure that is based in the Rasch[10] measurement

model, and Angoff's[11] Δ plot method. The Angoff and Wright–Panchapakesan approaches are similar except that the Angoff approach assigns each person's response experimentally to an ability stratum, whereas the Wright–Panchapakesan estimates the difficulty of the question conditioning the question difficulty estimates on the ability of the examinees in the respective groups. The ABFM selected the Wright–Panchapakesan DIF procedure because it permits questions to be flagged based on both the amount of DIF manifested in each item and the probability that the observed degree of DIF that occurred was statistically significant. This model is also a log-odds model that avoids the complications of scale compression at the ends of the restricted range raw score scale that can lead to floor and ceiling effects.[12]

ABFM DIF Process

In the construction of test questions, ABFM is attentive to not only the clarity of the question and the correctness of the answer but also to whether there are elements in the question that are likely to be confusing or misleading. In addition to trying to keep bias out of the questions during the item creation phase, ABFM also uses DIF postexamination flagging procedures to identify questions that should perhaps be screened again by a group of subject matter experts that were not involved with the creation of the questions. The ABFM DIF review process can be viewed as having two stages, the flagging of potentially biased questions and the panel review process. The terms "reference group" and "focal group" are used in DIF for group comparisons and generally refer to the "majority" and the "minority" demographic groupings for the examination population.

Flagging Procedure

In the flagging procedure, the questions are calibrated independently for the reference group and the focal group, and then the calibration of the focal group is subtracted from the calibration of the reference group to determine the DIF contrast. The statistical significance of observing a contrast of that magnitude is also computed. Items are flagged if they meet or exceed the flagging criteria on any of the designated subpopulations; therefore, it is possible for a single item to have multiple flags. For an item to be flagged, it must first meet the sample size requirements. These requirements are that there are at least 200 responses to the question across both groups and that in any reference–focal group comparison, the smaller of the 2 groups must have at least 50 responses. The next set of criteria is that the *absolute* value of the DIF contrast must be greater than or equal to 0.70 logits (our criteria for a substantively important difference) and the *absolute* value of the t-value must be greater than or equal to 1.96 (α =0.05; 95% certainty). These criteria flag questions without regard to which group found the item to be more difficult.

Panel Review Process

After the Spring FMCE is administered and scored, the DIF flagging procedure is run, and a DIF review panel is convened to review the flagged questions. The panel is intended to represent expertise in family medicine with a diversity of race/ethnicity and gender. Panelists cannot be involved in the item writing or test form creation process. There is also a linguist, and the meeting is moderated by a psychometrician. The psychometrician helps the panel to discuss which incorrect response options seem to be disproportionally more attractive to the disadvantaged group. The panel meeting begins with an explanation of DIF and the purpose of the panel. The rest of the meeting is dedicated to reviewing the questions to determine if there is an identifiable, content-based source of bias. If such a source of bias is identified by the panelists representing the focal group, then the entire panel determines by majority whether the content is an important aspect of family medicine or not. If the content generating the bias is an important aspect of family medicine, then the panel retains the item. If the source of bias is not an important aspect of family medicine, then the item is referred to the examination committee (EC) with the recommendation that the item be reworked or deleted. The EC makes the final decision whether to retain or rework/delete the item.

Methods—Review of Results of DIF

Participants

The participants were those examinees who took the FMCE between 2013 and 2020. However, race/ethnicity data were not collected for candidates seeking initial certification in 2013, but race/ethnicity data were available for all examinees starting in 2014. Because race/ethnicity categories are

dependent variables for the DIF analysis, only examinees for whom we had that data could be included. The race/ethnicity categories were White (reference group), Asian, Black, and Hispanic (focal groups). Other focal groups were not included due to inadequate sample sizes.

Instruments

The data were collected through ABFM's standard process for enrolling into and sitting for the FMCE. The FMCE is a multiple-choice question examination that produces scaled scores that reflect the examinees' medical knowledge and clinical decision-making ability. These scores range from 200 to 800, and the same scale has been used since 2008, but the passing standard was lowered from 390 to 380 in 2014. The core of the examination is 260 questions built to the 2006 test plan specifications.[13] More recent validity studies[4,14] have also supported the continued use of the 2006 test plan specifications. In addition to the core questions, examinees testing between 2007 to 2016 were also required to select 2 content-specific modules[15] from a menu of 8. Examinees testing in 2017, 2018, and 2019 were required to select only 1 module, and the modules were retired entirely starting in 2020. Only questions from the core of the FMCE were considered in this study because modules are no longer included on the FMCE, and we wanted the number of questions reviewed to remain similar across the span of the study. The Rasch reliability of the FMCE has generally been between 0.92 and 0.94.[16,17]

Design

Because this article is largely a review of the results generated by an existing ABFM quality-control process, the results are more descriptive than a test of any hypothesis. Its value is largely in providing a baseline for DIF findings when evaluating a certification examination.

Results

Balance of Questions Flagged

Table 1 shows the number of questions flagged by group membership, advantage/disadvantage classification, and year. The table only includes the core (nonmodule) questions for the ease of comparison across years. The mean percentage of questions flagged (either as an advantage or disadvantage) for Black was 5.3% (SD 1.3) with the mean net advantage (difference between the percentage of questions advantaging and disadvantaging the group) being 0.1% (SD 1.0). For the Hispanic group, the mean percentage of questions flagged was 5.0% (SD 1.2%) with the average net advantage being 0.2% (SD 0.7%). For the Asian group, the mean percentage of questions flagged was 6.7%, with the mean net advantage being 1.8% (SD 1.4%). For the focal groups, the mean net advantage was always positive but not to a statistically significantly degree.

Panel Review

Of the 3487 core questions, only 374 unique questions (10.7%) were flagged for DIF since 2013. Of these questions, only 4 (0.1%) were referred to the EC with the recommendation that the question be rewritten or removed from the item bank. In addition to the core questions, 5 questions from examinee-selected, content-specific modules were also forwarded to the EC with the same recommendations. The module questions were excluded from this study because the number of responses per question in the modules was often too small to permit analysis, and ABFM no longer offers modules. To date, all of the questions that were referred by the panel to the EC were removed from the item bank. In some cases, the question had already been removed for an additional, content-based reason. Figure 1 shows the percentage of questions flagged for DIF and the percentage referred to the EC.

Discussion

Balance of Questions Flagged for DIF

The data suggest that about 11% of our questions show a degree of differential performance across groups, but overall, there was no significant advantage to one group over another. Furthermore, close review by a special panel of diverse clinicians concluded that only a few of the questions had an identifiable source of bias that was not an important aspect of family medicine. This is further supported by the results in Table 1, which shows not only that the number of questions flagged was small but also the flagged questions were a relatively well-balanced split between questions that favored the focal and that favored the reference group. Within the date range of this study, the mean advantage noted for the Black and Hispanic groups was positive but very near zero; however, for Asians the average

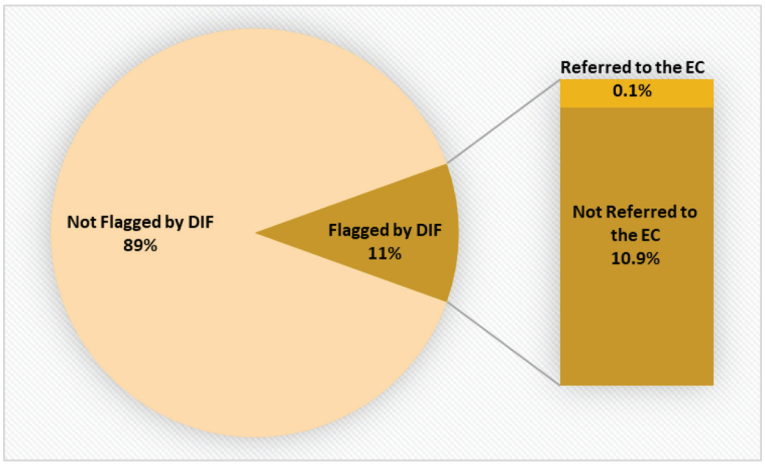
Table 1. Family Medicine Certification Examination Core Content Questions Flagged for DIF by Racial Group 2013-2020

Focal Group	Category	2013	2014	2015	2016	2017	2018	2019	2020	Count Mean (SD)	Percentage Mean (SD)
Black	Advantage	11 (2.6%)	10 (2.4%)	10 (2.3%)	11 (2.6%)	11 (2.6%)	10 (2.3%)	10 (2.3%)	18 (3.7%)	11.4 (2.9)	2.6% (0.5)
	Disadvantage	19 (4.4%)	9 (2.1%)	8 (1.9%)	7 (1.6%)	13 (3.0%)	11 (2.6%)	10 (2.3%)	11 (2.3%)	10.7 (3.9)	2.5% (0.9)
	Total	30 (7.0%)	19 (4.5%)	18 (4.2%)	18 (4.2%)	24 (5.6%)	21 (4.9%)	20 (4.6%)	29 (7.6%)	22.1 (5.1)	5.3% (1.3)
	Net advantage	−8 (−1.9%)	1 (0.2%)	2 (0.5%)	4 (0.9%)	−2 (−0.5%)	−1 (−0.2%)	0 (0%)	7 (1.4%)	0.7 (4.7)	0.1% (1.0)
Hispanic	Advantage	6 (3.3%)	5 (3.6%)	6 (5.4%)	3 (4.0%)	7 (5.8%)	2 (4.9%)	5 (3.7%)	7 (1.4%)	4.9 (1.8)	4.0% (1.4)
	Disadvantage	4 (0.9%)	4 (1.0%)	2 (0.5%)	7 (1.6%)	2 (0.5%)	5 (1.2%)	3 (0.7%)	7 (1.4%)	4.6 (1.9)	1.0% (0.4)
	Total	10 (4.2%)	9 (4.6%)	8 (5.9%)	10 (5.6%)	9 (6.3%)	7 (6.1%)	8 (4.4%)	14 (2.8%)	9.4 (2.3)	5.0% (1.2)
	Net advantage	2 (0.5%)	1 (0.2%)	4 (0.9%)	−4 (−0.9%)	5 (1.2%)	−3 (−0.7%)	2 (0.5%)	0 (0%)	0.3 (2.9)	0.2% (0.7)
Asian	Advantage	14 (3.3%)	15 (3.6%)	23 (5.4%)	17 (4.0%)	25 (5.8%)	21 (4.9%)	16 (3.7%)	19 (3.9%)	17.9 (3.3)	4.3% (0.9)
	Disadvantage	19 (4.4%)	10 (2.4%)	14 (3.3%)	7 (1.6%)	10 (2.3%)	10 (2.3%)	8 (1.9%)	8 (1.7%)	10.9 (4.3)	2.5% (0.9)
	Total	33 (7.7%)	25 (6.0%)	37 (8.7%)	24 (5.6%)	35 (8.1%)	31 (7.2%)	24 (4.6%)	27 (5.6%)	28.7 (5.1)	6.7% (1.4)
	Net advantage	−5 (−1.2%)	5 (1.2%)	9 (2.1%)	10 (2.3%)	15 (3.5%)	11 (2.6%)	8 (1.9%)	11 (2.3%)	7.0 (5.7)	1.8% (1.4)
Total number of flags		73	53	63	52	68	59	52	70	61.3 (8.5)	
Number of core questions		430	430	426	426	429	430	432	484	435.9 (19.6)	

Abbreviation: SD, standard deviation.

Note: The thresholds used to flag questions using the Rasch differential item functioning procedure were an absolute value of the logit contrast ≥ 0.70 and $P < .05$. The reference group was White. Advantage means that the question was easier for the focal group. Disadvantage means that the question was more difficult for the focal group. Net advantage = Advantage – Disadvantage. Flagging questions was not attempted if there were fewer than 200 responses overall or if the smaller of the 2 groups had fewer than 50 responses. Some questions received multiple flags.

Figure 1. Percentage of questions reviewed by DIF review process stage.



advantage was almost 2% of the test. Nevertheless, when the number of questions flagged advantaging or disadvantaging a group are roughly equal, it suggests that it may just be the effect of random variation.

DIF Criteria

The ABFM flagging criteria used in this study is primarily interested in "sensitivity"—detecting as many biased questions as possible but also keeping the "specificity" to a level that will not overwhelm the review panel with an excessive number of questions to review. Because the detection of problematic questions (sensitivity) is more important than shortening the list of questions being reviewed (specificity), perhaps the flagging criteria could be adjusted to have a smaller contrast value or higher α value.

Initiating a DIF Program

When considering whether to conduct DIF analyses, understand that any well-trained psychometrician can do it, although the initial setup will require time and effort. Rasch-based DIF is powerful and is easy to implement if the testing program is already using the Rasch model to score the examinations. Running DIF as part of the regular quality-control process is an opportunity to improve the quality of the questions, to demonstrate the rigor of the quality-control processes, and to demonstrate the organization's commitment to fairness.

What have we learned? The foundation of efforts to explore bias in certification is to begin to collect data on race and ethnicity from all Diplomates. Since 2008, as part of a major strategic initiative to develop research capacity to drive the evolution of board certification, the ABFM had begun to expand data collected routinely. Race and ethnicity were included in 2013. Surprisingly, only a handful of Diplomates and residents expressed concern; however, we did find that some participants (approximately 3.6% in 2016 and 5.7% in 2017) selected "other" for their race and ethnicity and subsequently described it as Indian, Middle Eastern, European, etc. For DIF, ABFM points out that we are detecting questions that perform differently across groups, and it does not matter which group was advantaged. If an organization is already collecting race/ethnicity data and fails to test for biases, the organization may seem indifferent to issues experienced by minority physicians. For this reason, it is prudent to have a rationale for collecting these data, especially if the reason is not for DIF analysis.

Weaknesses of This Study

The generalizability of this study is limited to the data on which it was based. It does not necessarily follow that similar results will be found with different medical specialties or even with future ABFM examinations. Changes in the examinee population, in the ABFM content development process, or in the selection and training of raters could conceivably cause changes in the number of questions that are flagged or removed. For this reason, it is important to continue conducting DIF analysis and discussing flagged questions with a review panel. Another issue is detecting

bias in small sample groups. The ABFM DIF flagging criteria requires that there be at least 200 responses to the question across the 2 groups being compared and that the smaller of the 2 groups contribute at least 50 responses. Given the number of Diplomates available for analysis, ABFM can only address the major groups of public interest at this time.

Future Research

DIF Criteria

As the goal of our DIF process is to identify truly biased questions, it would be useful to find a way to replicate DIF in a question so that there could be a greater certainty that it was not merely a statistical artifact. It would be useful to identify a set of questions that have been used across several administrations and then run the DIF analysis varying the criteria to see which criteria most faithfully replicate the bias flags.

Item Writing

From the perspective of writing test questions, it would be very useful to understand the mechanisms that underlie examinee biases in responding. This would allow ABFM to train content developers and reviewers to keep those elements out of test questions before they are ever administered to examinees. If those elements are important aspects of medicine, then those issues could be handled with education or by counterbalancing instances in which it provides advantages and disadvantages. For example, if 1 group is more inclined to select answers related to having a healthy diet and engaging in appropriate exercise and another group is more inclined to select options related to medical procedures or pharmacology, then a test form could be biased against 1 of the groups depending on what the correct answer is. Counterbalancing the number of questions on each test form to adjust for such a bias could help to generally improve examination fairness, but the particular bias would have to be identified.

Conducting DIF with Other Variables

ABFM has previously published a DIF study comparing the performance of 2008 and 2009 In-Training Examination (ITE) questions that were translated into Spanish and administered to a Spanish speaking cohort of residents in Quito, Ecuador, with the performance of residents from Accreditation Council for Graduate Medical Education residencies on the standard (English) version.[18] ABFM has also conducted DIF analysis on gender since 2013 but has not yet published the results. For future studies, we are considering applying these techniques to rurality and social class of origin.

Educational Opportunities

It would be helpful to see if there are examination performance differences by race/ethnicity and by year of residency using a repeated measures design. This could help to disentangle educational efficacy from baseline advantages some groups may have on entering residency. This design would use each resident at postgraduate year 1 as their own baseline or control. This line of reasoning could be extended to the post-initial-certification educational environments as well. It may be that minority Diplomates have different issues regarding getting time and support for quality educational activities, and this may also vary by the stage of the physician's career. The 2 studies described above are now in progress.

Conclusion

The results show that only a small percentage of the questions were flagged for potential bias, and among those questions, the number advantaging the focal group and the number disadvantaging the focal group were nearly perfectly balanced. This is strong evidence that the FMCE is not disadvantaging people of equal ability based on race or ethnicity. Although the implementation of DIF does help to prevent qualified examinees from being failed, it does nothing to increase the number of under-represented minorities that are eligible to take the examination. Simply put, the population of US physicians does not mirror the US population.[19] An informal comparison of the percentages of under-represented minorities taking the Medical College Admissions Test (MCAT) and the percentages of those same groups taking the FMCE 6 years later (3 years of medical school plus 3 years of residency) were within a few percentages points of each other. Although DIF does nothing to address educational inequities, it is a great tool to identify problematic questions.

To see this article online, please go to: http://jabfm.org/content/33/2/18.full.

References

1. Newton WP, Baxley E, Peterson LE, et al. How the ABFM will address health equity. Ann Fam Med 2020;18:468–70.
2. Jencks C, Phillips M, eds. [Internet]. The black-white test score gap. Brookings Institute Press; 1998 [cited 2021 March 26]. Available from: https://www.brookings.edu/book/the-black-white-test-score-gap/2.
3. 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.
4. O'Neill TR, Peabody MR, Stelter KL, Puffer JC, Brady JE. Validating the test plan specifications for the American Board of Family Medicine's certification examination. J Am Board Fam Med 2019;32:876–82.
5. O'Neill TR, Peabody MR, Puffer JC. The ABFM begins to use differential item functioning. J Am Board Fam Med 2013;26:807–9.
6. Cardall C, Coffman WE. A method for comparing the performance of different groups on the same items of a test. Ewing (NJ): Educational Testing Service; 1964. p. 64–5.
7. Holland PW, Thayer DT. Differential item performance and the Mantel–Haenszel procedure. In: Wainer H, Braun H, eds. Test validity. Mahwah (NJ): Lawrence Erlbaum; 1988. p. 129–45.
8. Mantel N, Haenszel W. Statistical aspects of the analysis of the data from retrospective studies of disease. J Natl Cancer Inst 1959;22:719–48.
9. Wright BD, Panchapakesan N. A procedure for sample-free item analysis. Educ Psychol Meas 1969;29:23–48.
10. Rasch G. Probabilistic models for some intelligence and attainment tests. Danish Institute for Educational Research; 1960.
11. Angoff WH, Ford SF. Item–race interaction on a test of scholastic aptitude. J Educational Measurement 1973;10:95–106.
12. Bond TG, Fox CM. Applying the Rasch model: fundamental measurement in the human sciences. Mahwah (NJ): Lawrence Erlbaum; 2001.
13. Norris TE, Rovinelli RJ, Puffer JC, Rinaldo J, Price DW. From specialty-based to practice-based: a new blueprint for the American Board of Family Medicine cognitive examination. J Am Board Fam Pract 2005;18:546–54.
14. Peabody MR, O'Neill TR, Stelter KL, Puffer JC. Frequency and criticality of diagnoses in family medicine practices: from the National Ambulatory Medical Care Survey (NAMCS). J Am Board Fam Med 2018;31:126–38.
15. O'Neill TR, Peabody MR. Impact of one versus two content-specific modules on American Board of Family Medicine certification examination scores. J Am Board Fam Med 2017;30:85–90.
16. O'Neill TR, Li Z, Peabody MR, Lybarger M, Royal K, Puffer JC. The predictive validity of ABFM's In-Training Examination. Fam Med 2015;47:349–56.
17. O'Neill TR, Peabody MR, Song H. The predictive validity of the National Board of Osteopathic Medical Examiners' COMLEX-USA examinations with regard to outcomes on American Board of Family Medicine examinations. Acad Med 2016;91:1568–75.
18. O'Neill TR, Raddatz MM, Royal KD. Demonstrating the construct stability of a translated exam for family medicine residents. Int J Educ Psychol Assess 2011;6:31–41.
19. Peabody MR, Eden AR, Douglas M, Phillips RL. Board certified family physician workforce: progress in racial and ethnic diversity. J Am Board Fam Med 2018;31:842–3.