ORIGINAL RESEARCH

Performance on the American Board of Family Medicine (ABFM) Certification Examination: Are Superior Test-Taking Skills Alone Sufficient to Pass?

Thomas R. O'Neill, PhD, Kenneth D. Royal, PhD, and James C. Puffer, MD

Introduction: **Certification examinations used by American specialty boards have been the *sine qua non* for demonstrating the knowledge sufficient for attainment of board certification in the United States for more than 75 years. Some people contend that the examination is predominantly a test of superior test-taking skills rather than of family medicine decision-making ability. In an effort to explore the validity of this assertion, we administered the American Board of Family Medicine (ABFM) Certification to examinees who had demonstrated proficiency in taking standardized tests but had limited medical knowledge.**

Methods: **Four nonphysician experts in the field of measurement and testing were administered one version of the 2009 ABFM certification examination. Scaled scores were calculated for each examinee, and psychometric analyses were performed on the examinees responses to examination items and compared with the performance of physicians who took the same examination.**

Results: **The minimum passing threshold for the examination was a scaled score of 390, corresponding to 57.7% to 61.0% of questions answered correctly, depending on the version of the examination. The 4 nonphysician examinees performed poorly, with scaled scores that ranged from 20 to 160 (mean, 87.5; SD, 57.4). The number of questions answered correctly ranged from 24.0% to 35.1% (mean, 29.2%; SD, 0.05%). Rasch analyses of the examination items revealed that the nonphysician examinees were more likely to use guessing strategies in an effort to answer questions correctly. Distracter analysis suggest near-complete randomness in the nonphysician responses.**

Conclusions: **Though all 4 nonphysician examinees performed better than would have been predicted by chance alone, none performed well enough to even fall within 8 SE below the passing thresholds; their performance was far below that of almost all physicians who completed the examination. Given that the nonphysicians relied heavily on the identifying cues in the phrasing of items and the manner in which response options were presented, the results affirm the notion that the ABFM certification examination is not primarily a measure of generic test-taking ability but measures information critical to the estimation of a family physician's knowledge sufficient for certification. Item analysis confirmed that items were well written, provided minimal cueing, and required medical knowledge to answer correctly. (J Am Board Fam Med 2011;24:175–180.)**

Keywords: **Certification, Graduate Education, Statistical and Mixed Methods**

Intensive examinations have been used by specialty boards to certify physicians for more than 75 years and are considered to be the gold standard in the certification process. These high stakes examinations carry "considerable implications for candidates' career progression, future employment, and remuneration."[1] The public is also keenly interested in the certification status of physicians because certification is a measure of a physician's medical knowledge in his or her specialty area.[2] Though considerable evidence exists that correlates physician performance on these examinations with desirable physician behaviors and patient outcomes,[3–11] limited information exists regarding the influence of test-taking skills on the ability to suc-

This article was externally peer reviewed.
Submitted 12 July 2010; revised 25 October 2010; accepted 10 November 2010.
From the American Board of Family Medicine, Lexington, KY.
Funding: none.
Conflict of interest: All authors are employees of the American Board of Family Medicine.
Corresponding author: Thomas O'Neill, PhD, American Board of Family Medicine, 1648 McGrathiana Parkway, Suite 550, Lexington, KY, 40511 (E-mail: toneill@theabfm.org).

cessfully pass the examination and thus become certified. This is largely because of experimental design issues in isolating test-taking skills from ability. Typically, one administers a test to assess someone's ability, but if the test score is assumed to be biased by the effects of test-taking skills, then an independent and highly reliable measure of that ability,[12] free from the effects of test-taking skills, is needed for comparison. If such a measure existed the effects of test-taking skills could be easily separated from the effects of ability, but, in the absence of such a measure, it is difficult to experimentally disentangle these 2 concepts.

Though it is true that some test-taking skill is required to succeed on the American Board of Family Medicine (ABFM) certification examination, the requirement is fairly minimal. Necessary skills include having the visual acuity and adequate language mastery to read the test items. A successful testing outcome also is based on some assumptions: (1) the test taker understands that all questions should be attempted and that unanswered questions are scored as incorrect; and (2) the test taker appreciates that only one answer will be scored as correct and that they should attempt to identify and mark the best response option. These skills and assumptions do not seem to impose an onerous burden on the examinee, nor are they expected to have an appreciable impact on the examinee's performance. However, good test takers are sometimes able to find psychological cues within the wording of an item or its response options. As such, a savvy test taker could be expected to perform fairly well on a test, regardless of the content, should these cues be evident.

Given the development process used to create the test, it is clear that the intent of the ABFM certification examination is to measure physician ability within the family physician's scope of practice.[13] Verification that the test is functioning as expected is provided by the psychometric processes used to score the examination. Psychometric validation demonstrates that the construct implied by the questions is stable. However, it is possible that many graduates from family medicine residency programs also possess good test-taking skills. To attempt to separate these 2 concepts—physician ability within the scope of family medicine and generic test-taking ability—we examined a group of highly educated nonphysicians to answer the question, How well can equivalently, highly educated people (ie, those who hold terminal degrees in fields other than medicine), who are experts in testing and historically good at test taking, expect to perform on the ABFM board certification examination?

Methods

Design

Four nonphysicians who are considered experts in the field of certification and licensure testing completed the summer 2009 ABFM board certification examination. Each participant was directed to do his very best to pass the examination. Scaled scores were calculated for each examinee, and psychometric analyses were performed on the examinees' responses to examination items and compared with the performance of physicians who took the same examination.

Participants

The 4 nonphysicians are employed by the ABFM (this requirement was imposed because, for security reasons, there is a very limited number of people permitted to see the test questions outside of those who are taking the test to earn or maintain certification). All 4 participants have doctoral degrees and varying amounts of experience working with certification tests. All 4 participants were appropriately motivated by the possibility of passing a certification examination, especially an examination from a field other than their respective areas of expertise. As an added incentive to perform well, all participants consented to release their individual score results, regardless of the outcome, for possible publication in a conference paper or journal article.

Of the 4 participants, 3 are psychometricians and one is an examination administration and credentials professional. Subject 1 has a background in higher education and quantitative methods and has less than 1 year of experience in the testing/licensure industry. Subject 2 was formally trained and has a doctoral degree in education and 25 years' experience working at the ABFM. Subject 3 has a background in educational psychology and clinical psychology and has worked in the testing/licensure industry for 20 years. Subject 4 is a psychologist with 7 years' experience in the testing/licensure industry and 8 years' experience conducting research in the psychological assessment arena.

Instrument

The 2009 ABFM board certification examination was used,[14] which was administered in the summer of 2009. The examination includes 350 questions that are administered via computer in a random order. As part of the certification examination, all candidates must self-select 2 topic-specific modules on which they wish to be examined. Each of the 4 participants strategically selected modules that they believed would give them the greatest advantage. Of the 350 questions, each module consists of 45 items, leaving 260 items on the core portion of the examination.

Data Analysis

Using Winsteps (2009) measurement software,[15] Rasch analyses were conducted to investigate how people and items were interacting. Rasch analysis operates under the assumption, in the mathematical sense, that the more able a person is, the higher probability he or she has of getting an item correct. Ability, then, is defined as having a higher score. Likewise, more difficult items have a lower probability of being answered correctly. Items were reported in order of the most difficult to the least difficult, and a *P* value that indicated the extent to which each item was answered correctly was reported for each item. Using a distractor analysis, a common psychometric technique that identifies the extent to which each answer was selected (both counts and percents) for each item, we were able to investigate the extent to which nonphysicians correctly guessed at each item, thus providing evidence of false-positives and score inflation caused by a lack of stable content knowledge.

Results

The minimum passing threshold for the 2009 certification examination was a scaled score of 390, corresponding to 57.7% to 61.0% of questions answered correctly, depending on the version of the examination. The 4 nonphysician examinees performed poorly, with scaled scores that ranged from 20 to 160 (mean score, 87. 5; SD, 57.4; see Table 1). The number of questions answered correctly ranged from 24.0% to 35.1% (mean, 29.2%; SD, 0.05%).

Better Than Chance

Most of the questions on the examination have 5 options whereas a few have only 4. The average number of response options on the examination was 4.64, which means that just by responding in a random fashion one could expect to get 21.5% of the questions correct. The percent correct for our participants ranged from 24% to 35%, indicating that each participant scored at a level that was above chance alone. In addition, we investigated what would happen if a test taker simply selected A, B, C or D for all items. A row of data containing the same response (A through D) was included in the data file for each nonphysician. Results indicate that the average score one would obtain from selecting all A, B, C, or D responses would be −20, with a possible range of −120 to 50. The observed variability would be because of differences in the frequency of correct responses for A, B, C, and D and the difficulty of the items associated with each form of the examination.

Furthermore, a distracter analysis was performed to reveal the specifics of how each nonphysician responded to each item. This process involved visually inspecting all 260 of the common core items and investigating each participant's response to each item. As noted previously, more able persons have a higher probability of marking correct answers. The distracter analysis found near-complete randomness among the nonphysician study participants, indicating that the most able person in the sample did not regularly outperform his colleagues on an item-by-item basis. In fact, there were numerous items for which the 4 nonphysicians each selected a different response option. This lack of consistency is a clear indicator

Table 1. Performance of Nonphysician Participants

Name	Scaled Score	Correct (%)	Ranking
Participant 1	20	24.0	Outscored 0 physicians
Participant 2	80	28.9	Outscored 0 physicians
Participant 3	90	28.9	Outscored 0 physicians
Participant 4*	160	35.1	Outscored 4 physicians†

*Out of more than 10,000 physicians tested, only one participant was able to outscore even the lowest-performing physicians.
†Of the 4 physicians who were outperformed, 2 were international medical graduates (one with a history of multiple failures and one a first-time candidate) and 2 were US medical graduates who left many questions unanswered (one left 33 unanswered and one left 79 unanswered) on their examination.

that the participants in this sample relied heavily, if not almost entirely, on some form of guessing.

Distribution of Summer 2009 Candidates

A total of 10,818 candidates completed the ABFM's board certification examination in the summer of 2009. Approximately 86% passed the examination. The mean score was 500 (SD, 108). It should be noted that scores are reported on a scale ranging from 200 to 800. Scores lower than 200 are reported as 200 and scores greater than 800 are reported as 800.

Only 8 physicians scored below a 200 on the Summer 2009 examination. Of the 8 physicians, 3 left multiple questions on the examination unanswered (25, 33, and 79 unanswered items), which resulted in incorrect answers. By sheer probability, one might expect these physicians to surpass a score of 200 had they actually answered all of the questions. One might argue that, although only 8 of the 10,818 physicians scored below 200, it is likely that only 5 physicians truly would have scored below 200 had all physicians completed every item on the examination. To put these results in perspective, about 0.0004% to 0.0007% of the examinee population failed to reach a score of 200.

Distribution of Nonphysician Candidates

None of the participants were able to score within the reportable range of the scale. Compared with the cohort of more than 10,000 physicians, 3 of the 4 participants scored lower than all the physicians. It is worth noting that some of the physicians who outscored the participants actually left large numbers of questions unanswered, which were then scored as wrong. Only one participant outscored physicians, actually 4 physicians, and 2 of the physicians he outscored had left many questions blank.

Discussion

Nonphysician scores were so low that none of the participants scored high enough to reach the minimum reported measure on the scale (200). These results tend to support the idea that the examination assesses family medicine content knowledge and that one cannot pass the examination by making a series of random or educated guesses.

Experts in tests and measurement have identified 3 primary types of guessing: random guessing, cued guessing, and informed guessing.[16] Random guessing occurs when examinees respond blindly to a test item. Cued guessing occurs when examinees respond based on a stimulus in a test item (ie, wording cues and cues associated with the nature of the distracters), and informed guessing (also called an "educated guess") takes place when examinees respond based on partial knowledge or misinformation. As Downing[17] points out, medical examinees rarely rely on random guessing; rather, they make decisions based on informed guesses. Being able to remove responses that are unlikely to be correct dramatically increases one's chances of answering an item correctly. However, if one is able to eliminate all but 2 response options, one still has, at best, a 50% chance of answering an item correctly. For the 4 examinees in this study, all of whom had no a background in family medicine, instances of "informed guessing" were greatly reduced. The data clearly demonstrate more random guesses, thus increasing the probability of marking an item incorrectly. According to Downing,[17] the odds of passing the examination based on random guessing alone are somewhat comparable to the odds of winning the lottery.

When investigating the performance on the certification examination of the nonphysicians, it was clear that guessing was rampant because the pattern of items answered correctly spanned both ends of the difficulty continuum (ie, some items that were very difficult for physicians were guessed correctly by the nonphysicians, and some items that were very easy for physicians were answered incorrectly by the nonphysicians). In other words, the difficulty of the items had little, if anything, to do with each nonphysician's probability of getting the item correct because the nonphysicians generally selected a response based on some form of guessing.

To qualitatively tease out some explanations for the variance in scores among the 4 nonphysicians, it is likely that subject 4's superior performance over his colleagues could be a result of having a PhD in clinical psychology; he has received extensive training about issues of emotional health, mental disorders, and their psychological treatment. Empirical results indicate that this is indeed the case. Items relating to psychogenics comprise 7% of the core portion of the examination, and subject 4 correctly answered 11 of 19 items in this clinical category. Subject 3, whose doctorate is in the field of educational psychology and whose master's degree is in clinical psychology, answered 8 of 19 items cor-

rectly in this clinical category. Subject 2 and subject 1 correctly answered 5 of 19 and 3 of 19 items, respectively, in this category. It is clear that the 2 persons with a background in this area are the 2 who performed "reasonably well." A similar trend was present in the mental health clinical category. It seems that, even among nonphysicians, having some content knowledge in a particular domain will somewhat improve scores in relevant areas.

Assuming all else being equal, having some content knowledge in this domain likely contributed to subject 4's higher score. Likewise, subject 1's lower score in relation to his colleagues is likely because of a lack of experience with and knowledge about health-related issues. Furthermore, subject 1's lower score is indicative of someone who truly guessed at random, whereas subject 4's higher score can be partly attributed to more informed guesses and more stable knowledge in some areas. Subject 2 and subject 3's midrange scores, relative to the nonphysician cohort, are likely because of experience with regard to age, number of years working in health-related industries, and knowledge acquired from having families (each are married with children). However, it is reasonable to assume their performance was not up to par with subject 4's because they also lack the amount of direct exposure to some medical knowledge (ie, psychogenics).

Limitations

The study is obviously constrained by its small sample size. However, given their backgrounds in the area of test design and measurement, these 4 individuals had a greater chance of passing the certification examination than any other non–medically trained staff within the organization except the content development staff, who are constantly exposed to examination content.

Motivation of the nonphysician participants also may be considered a limitation. ABFM diplomates have much at stake when they take the certification examination, and they often spend months preparing. The nonphysicians in this sample did not have the same performance pressures on them as did the physicians who were actually taking the examination. The nonphysicians also had much less invested in their performance. Although the extrinsic motivators were not evident for the nonphysicians, each took the examination very seriously and performed as well as he possibly could.

Conclusions

Although all the participants seem to have performed better than chance would predict, it was not by much. This study affirms the notion that the ABFM board examination is not predominantly a measure of generic test-taking ability and clearly requires medical training to pass.

The results confirm that the existing test development processes do a good job of eliminating cues to the correct answer. The 4 nonphysicians relied heavily on identifying cues in the phrasing of items and the manner in which response options were presented. When cues were present, the nonphysician participants' chances of getting an item correct increased. However, without any cues the nonphysician participants relied solely on guessing, which is why each failed miserably in those areas where they possessed no content knowledge. The evidence demonstrates that the ABFM board certification examination is not a measure of generic test-taking ability, and people without appropriate medical training are extremely unlikely to pass the examination.

References

1. Hutchinson L, Aitken P, Hayes T. Are medical postgraduate certification processes valid? A systematic review of the published evidence. Med Educ 2002; 36:73–91.
2. American Board of Medical Specialties. Annual report and reference handbook. Evanston, IL: American Board of Medical Specialties; 2000:69.
3. Sharp LK, Bashook PG, Lipsky MS, Horowitz SD, Miller SH. Specialty board certification and clinical outcomes: the missing link. Acad Med 2002;77:534–42.
4. Silber JH, Kennedy SK, Even-Shoshan O, et al. Anesthesiologist board certification and patient outcomes. Anesthesiology 2002;96:1044–52.
5. Prystowsky JB. Patient outcomes for segmental colon resection according to surgeon's training, certification, and experience. Surgery 2002;132:663–70.
6. Pham HH, Schrag D, Hargraves JL, Bach PB. Delivery of preventive services to older adults by primary care physicians. JAMA 2005;294:473–81.
7. Norcini JJ, Kimball HR, Lipner RS. Certification and specialization: do they matter in the outcome of acute myocardial infarction? Acad Med 2000;75: 1193–8.
8. Chen J, Rathore SS, Wang Y, Radford MJ, Krumholz HM. Physician board certification and the care and outcomes of elderly patients with acute myocardial infarction. J Gen Intern Med 2006;21:238–44.
9. Brennan TA, Horwitz RI, Duffy FD, Cassel CK, Goode LD, Lipner RS. The role of physician spe-

cialty board certification status in the quality movement. JAMA 2004;292:1038–43.
10. Turchin A, Shubina M, Chodos AH, Einbinder JS, Pendergrass ML. Effect of board certification on anti-hypertensive treatment intensification in patients with diabetes. Circulation 2008;117:623–8.
11. Holmboe ES, Lipner R, Greiner A. Assessing quality of care: knowledge matters. JAMA 2008;299:338–40.
12. Meade AM, Tonidandel S. Not seeing clearly with Cleary: what test bias analyses do and do not tell us. Industrial and Organizational Psychology: Perspectives on Science and Practice 2010:3:192–205.
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 Med 2005;18:546–54.
14. American Board of Family Medicine. Candidate information booklet. Lexington, KY: American Board of Family Medicine; 2010.
15. Winsteps. Rasch measurement computer program, version 3.68.0. Available from: http://www.winsteps.com/index.htm. Accessed January 17, 2011.
16. Rogers HJ. Guessing in multiple-choice tests. In: Masters GN, Keeves JP, eds. Advances in Measurement in Educational Research and Assessment. Oxford, UK: Pergamon; 1999:235–43.
17. Downing SM. Guessing on selected-response examinations. Med Educ 2003;37:670–1.