ORIGINAL RESEARCH

# Impact of One Versus Two Content-Specific Modules on American Board of Family Medicine Certification Examination Scores

*Thomas R. O'Neill, PhD, and Michael R. Peabody, PhD*

***Background:*** **We consider the question of whether requiring diplomates to select only 1 content-specific module, rather than 2, would increase, decrease, or produce no change in scores among the examinee population.**

***Methods:*** **Examinees' scores were computed under 3 different conditions: the examination core plus (1) both modules, (2) the module on which they scored higher, and (3) the module on which they scored lower.**

***Results:*** **Although the differences in scores across the 3 conditions were relatively small, asking examinees to select only a single module would likely benefit more examinees than it would harm by a 4:1 ratio, assuming that the diplomates selected the module on which they scored higher. Only 114 of the 29,088 examinees (0.4%) would have changed from a pass to a fail, whereas 467 (1.6%) would have changed from fail to pass.**

***Conclusion:*** **These results suggest that having examinees select 1 module rather than 2 will likely produce a slight score increase for examinees. Simultaneously, it would improve the standardization of the examination across examinees. (J Am Board Fam Med 2017;30:85–90.)**

***Keywords:*** **Clinical Competence, Educational Measurement**

The American Board of Family Medicine (ABFM) is the second largest medical specialty board in the United States. One component of ABFM's certification program is the periodic demonstration, via a standardized test, that a diplomate has at least a minimum knowledge of medical information and at least minimal clinical decision-making abilities to be considered ABFM certified. Presently, 74% of the ABFM Family Medicine Certification Examination is defined by the core test plan specifications[1]; however, examinees are also required to select 2 content-specific modules from a menu of 8.[2] The 8 modules are Geriatric Medicine, Emergent/Urgent Care, Ambulatory Family Medicine (AFM), Child & Adolescent Care, Women's Health, Maternity Care, Hospital Medicine, and Sports Medicine. The 2 modules selected by an examinee account for the remaining 26% of his or her examination. The initial intent was to make the examination more reflective of an individual physician's practice; however, there is an ongoing tension between making the examination sufficiently relevant to a family physician's practice and standardizing the examination so that it reflects the full spectrum of family medicine that is implied in the ABFM certificate.

ABFM began to examine systematically the impact of module selection on examination performance; using data from 2008,[3] the ABFM found a tendency for examinees with a high ability to use the modules to further inflate their scores and a tendency for examinees with a low ability to be disadvantaged by them. This finding was replicated in studies using data from 2013[4] and 2015.[5] As an extension of this previous work, we consider in this study the question of whether requiring diplomates to select only 1 content-specific module, rather

This article was externally peer reviewed.
*Submitted* 24 May 2016; revised 3 August 2016; accepted 5 August 2016.
*From* the American Board of Family Medicine, Lexington, KY.
*Funding:* none.
*Conflict of interest:* none declared.
*Corresponding author:* Thomas R. O'Neill, PhD, American Board of Family Medicine, 1648 McGrathiana Pkwy, Lexington, KY 40511 (E-mail: toneill@theabfm.org).

than 2, would produce an advantage or a disadvantage for the examinee population.

## Methods

### *Instruments*

The ABFM Family Medicine Certification Examination is a 370-item, fixed-length, computer-administered, nonadaptive certification examination. Of the 370 questions, 20 are unscored pretest items, 260 are based on the core test plan specifications, and 90 are from 2 content-specific modules (45 items each).[6] Examinees are required to select the 2 from among 8 possible modules. All questions, including those in the content-specific modules, are calibrated to a common scale using the dichotomous Rasch measurement model[7] in conjunction with a common-item equating design. Person ability estimates are also computed using this model. The Rasch model's conventional unit of measure is log-odds units (logits); however, these logits are transformed into scaled scores before they are reported to examinees. Reported scores can range from 200 to 800 and increase in units of 10; scores <200 are reported as 200 and scores >800 are reported as 800. The passing standard for the ABFM Family Medicine Certification Examination was 380 in both 2014 and 2015.

### *Participants*

The participants in this study were all the examinees (N = 29,088) who took the ABFM Family Medicine Certification Examination during 2014 and 2015. All the examinees were physicians who were testing to earn their initial certification or were already certified and were testing to maintain their certification. A small portion of the examinees sat for more than one of these administrations. The procedures used in this study were reviewed by senior ABFM executive staff to ensure that ABFM privacy policies were not being violated. In addition, the data were deemed exempt by the American Academy of Family Physicians Institutional Review Board.

### *Analysis*

The analyses were based on comparisons of examinees' scaled scores computed under 3 different conditions and the impact that those 3 conditions ultimately had on the population pass rate. The 3 conditions were computing the scaled scores using the examinee's responses from the (1) total examination, which included the examination core plus both modules ("actual"), (2) the examination core plus the module with the higher of the 2 module scores ("better"), and (3) the examination core plus the module with the lower of the 2 module scores ("worse"). Each score was computed using the dichotomous Rasch model, and the difficulty calibrations for all the items were set to the same values that were used in the actual scoring. Tests for statistical significance were conducted using R version 3.0.1 (including plyr, reshape2, and ggplot2; available at http://www.r-project.org/). An 0.05 level of significance was used as the critical value for all statistical tests.

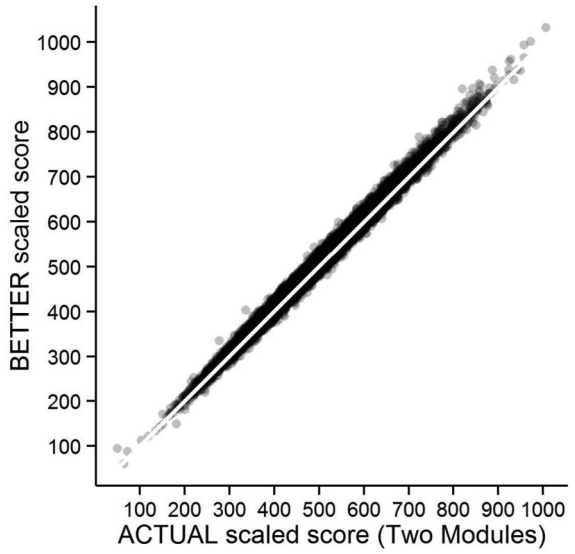
We first created a scatterplot of the examinees' actual scores on the x-axis with their better scores on the y-axis. An identity line was added to provide context. If the scores were exactly equal, then all the points would be plotted along the identity line. Points above the identity line represent examinees who would benefit from including the better of the 2 modules and excluding the worse of the 2, whereas points below the identity line represent examinees who would be disadvantaged by including the better of the 2 modules and excluding the worse of the 2. We also calculated a table of summary statistics comparing actual with better and actual with worse. Then we compared the aggregated actual condition with the better condition with regard to the pass-fail decisions made under each condition to determine how many people would be advantaged and disadvantaged.

Finally, we created an inverse cumulative frequency chart to show the potential impact of the change on the pass rate along the ability spectrum. The comparisons between actual and better assume that if examinees were required to select only 1 module, they would, in fact, pick the module on which they performed the best. Because this is unlikely to be true in an absolute sense and the extent to which it is true is unknown, the worse score was used to represent the lower bound of what the outcome might be. In this way, better and worse may be thought of as producing a confidence interval around actual, representing the likely impact of using a 1-module test format.

## Results

Figure 1 shows a cloud of points rather than all the points falling on the identity line, indicating that a

**Figure 1. Density scatterplot of each examinee's actual score using both modules with their better score.**




change in scores has occurred. More points lie above the identity line than below it, indicating there was an overall score increase using the "better" condition. Points below the line indicate instances in which both of an examinee's module scores were higher than their core score; thus removing the lower of the 2 modules scores caused the "better" condition score to drop to be more aligned with their core score.

Table 1 shows that overall there was a 5.4 mean scaled-score point increase in examinee scores when the better score is used rather than the actual score. Accordingly, the overall pass rate increased by 1.2%. The mean scaled-score increases and the pass rate increases occurred in each of the 4 test administrations to a statistically significant degree (April 2014: actual, $\bar{x}$ = 507.1, standard deviation [SD] = 108.3; better, $\bar{x}$ = 512.6, SD = 109.3 [$t(-56.1)$ = 10,617; $P < .000$]); November 2014: actual, $\bar{x}$ = 471.9, SD = 109.6; better, $\bar{x}$ = 476.8, SD = 110.6 [$t(-33.4)$ = 4,801; $P < .000$]; April 2015, actual $\bar{x}$ = 499.9, SD = 105.8; better $\bar{x}$ = 505.6, SD = 105.9 [$t(-58.6)$ = 9,604; $P < .000$]; November 2015: actual $\bar{x}$ = 454.2, SD = 108.9; better $\bar{x}$ = 459.0, SD = 109.1 [$t(-31.1)$ = 4,602; $P < .000$]).

**Table 2. Comparison of Pass/Fail Results**

| | Better | | |
|---|---|---|---|
| Actual | Pass | Fail | Total |
| Pass | 24,549 | 114 | 24,663 |
| Fail | 467 | 3,958 | 4,425 |
| Total | 25,016 | 4,072 | 29,088 |

Similarly, across all 4 administrations, the worse mean scaled score was lower than the actual mean scaled score, to a statistically significant degree (April 2014: actual, $\bar{x}$ = 507.1, SD = 108.3; worse, $\bar{x}$ = 499.0, SD = 108.1 [$t(88.4)$ = 10,617; $P < .000$]; November 2014: actual, $\bar{x}$ = 471.9, SD = 109.6; worse, $\bar{x}$ = 462.9, SD = 109.1 [$t(67.6)$ = 4,801; $P < .000$]; April 2015: actual, $\bar{x}$ = 499.9, SD = 105.8; worse, $\bar{x}$ = 491.9, SD = 105.4 [$t(85.3)$ = 9,604; $P < .000$]; November 2015: actual, $\bar{x}$ = 454.2, SD = 108.9; worse, $\bar{x}$ = 446.1, SD = 108.2 [$t(54.9)$ = 4,602; $P < .000$]).

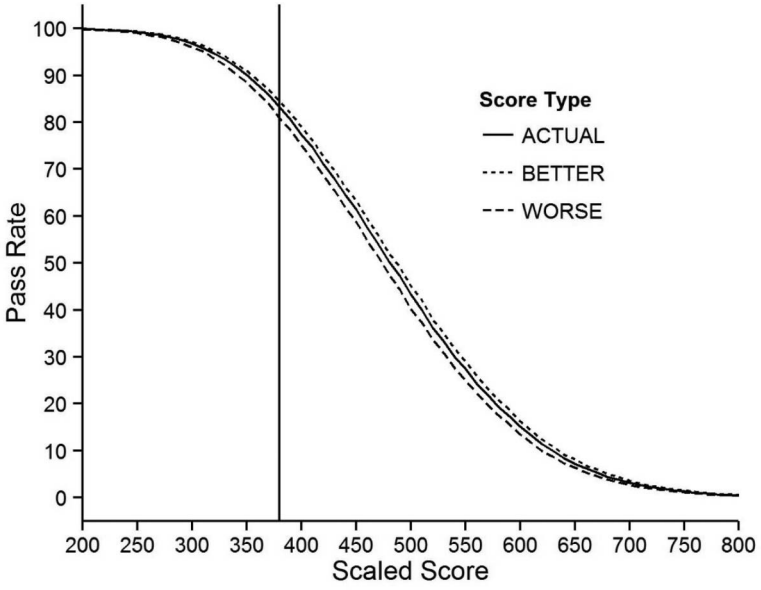
Examining the potential impact on individuals, Table 2 shows that of the 29,088 total examinees, 98% (n = 28,507) would have had no change in their pass-fail status if better was used instead of actual. There were 24,549 examinees whose actual status was pass and whose better status was also pass, whereas there were 3,958 whose actual status was fail and whose better status was also fail.

**Table 1. Mean Scores and Pass Rates across 3 Conditions and 4 Administrations**

| | | Mean (SD) of the Sample | | | Pass (%) | | |
|---|---|---|---|---|---|---|---|
| Administration | N | Worse | Actual | Better | Worse | Actual | Better |
| April 2014 | 10,618 | 499.0 (108.1)* | 507.1 (108.3) | 512.6 (109.3)† | 87.3 | 89.0 | 89.9 |
| November 2014 | 4,802 | 462.9 (109.1)* | 471.9 (109.6) | 476.8 (110.6)† | 76.7 | 79.5 | 81.1 |
| April 2015 | 9,605 | 491.9 (105.4)* | 499.9 (105.8) | 505.6 (105.9)† | 85.4 | 87.1 | 88.2 |
| November 2015 | 4,063 | 446.1 (108.2)* | 454.2 (108.9) | 459.0 (109.1)† | 71.3 | 74.4 | 76.0 |
| Total | 29,088 | 483.3 (109.2) | 491.5 (109.5) | 496.9 (109.5) | 82.7 | 84.8 | 86.0 |

*Statistically significantly lower than the actual condition ($P < .000$, 2-tailed).
†Statistically significantly higher than the actual condition ( $P < .000$, 2-tailed).
SD, standard deviation.

**Figure 2. Inverse cumulative frequency distributions showing the potential pass rates using examinees' actual 2-module scores, better module scores, and worse module scores. The vertical line represents the current minimum passing standard (380).**




Among the 2% (n = 581) who would have had a change in their pass-fail status, 4 times as many people would have gone from fail to pass than from pass to fail. There were 114 examinees whose actual status was pass and whose better status was fail, whereas there were 467 whose actual status was fail and whose better status was pass.

Figure 2 shows the potential impact of a change in the pass rate along the ability spectrum. The solid line represents the actual examination results, the dotted line represents the best-case scenario if each examinee selected the module in which they performed best, and the dashed line represents the worst-case scenario if each examinee selected the module in which they performed worse. The vertical line represents the minimum passing score of 380. Based on this graph, the actual examination pass rate was 83.2%, the best-case scenario pass rate would have been 84.4%, and the worst-case scenario pass rate would have been 80.9%. The difference between the best-case scenario and worst-case scenario would have been 3.5 percentage points.

To provide context, the number of examinees selecting each of the 8 modules is presented in Table 3, which shows the number of examinees who selected each module across each administration under the "actual" condition. Note that under the "actual" condition, each examinee selects 2 modules, so the percentage is the observed number of examinees who selected the module divided by the total number of examinees in the study (N = 29,088). Thus, the counts sum to 58,176 and the percentages sum to 200%. For the years of this study, the most popular module was AFM: 82% of the examinees selected it as 1 of their 2 choices. The second most popular selection was Geriatrics (36%) followed by Women's Health (20%).

## Discussion

The results indicate that although there is a difference in examinee scores based on each of the 3 different scoring conditions, the differences are relatively small. Asking examinees to select only a single module would benefit more examinees than it would hurt by a 4:1 ratio—assuming that they accurately select the module on which they would perform better. Under these assumptions, only 114 of the 29,088 examinees (0.4%) would have changed from a pass to a fail, whereas 467 (1.6%) would have changed from fail to pass. These results seem congruent with the idea that most family physicians are generalists who practice broad-spectrum family medicine. Why physicians select specific modules and whether they can accurately determine the modules on which they will perform best has not been studied. Some research suggests

**Table 3. Counts and Percentages of Examinees Selecting Each Module**

| Module | April 2014 (n = 10,618) | November 2014 (n = 4,802) | April 2015 (n = 9,605) | November 2015 (n = 4,063) | Total (N = 29,088) |
|---|---|---|---|---|---|
| Geriatrics | 3,814 (36%) | 1,835 (38%) | 3,258 (34%) | 1,529 (38%) | 10,436 (36%) |
| Emergent/urgent care | 1,757 (17%) | 1,065 (22%) | 1,590 (17%) | 899 (22%) | 5,311 (18%) |
| Ambulatory family medicine | 8,794 (83%) | 3,876 (81%) | 7,834 (82%) | 3,325 (82%) | 23,829 (82%) |
| Child and adolescent care | 768 (7%) | 406 (8%) | 621 (6%) | 347 (9%) | 2,142 (7%) |
| Women's health | 2,055 (19%) | 977 (20%) | 1,948 (20%) | 914 (22%) | 5,894 (20%) |
| Maternity Care | 2,022 (19%) | 612 (13%) | 2,056 (21%) | 439 (11%) | 5,129 (18%) |
| Hospital medicine | 1,286 (12%) | 465 (10%) | 1,278 (13%) | 374 (9%) | 3,403 (12%) |
| Sports medicine | 740 (7%) | 368 (8%) | 625 (7%) | 299 (7%) | 2,032 (7%) |
| Total | 21,236 (200%) | 9,604 (200%) | 19,210 (200%) | 8,126 (200%) | 58,176 (200%) |

The percentages are the observed counts of examinees who selected the module divided by the total number of examinees in the study (N = 29,088). Because examinees select 2 modules, the counts sum to 58,176 and the percentages sum to 200%.

that physicians are not good at self-assessment,[8] and asking them to make this determination may be creating problems for examinees.

In this study, >80% of the examinees selected AFM as 1 of their 2 modules (Table 3). The selection of the AFM module is understandable for 2 reasons. First, this module closely aligns with the specifications of the core of the examination, so examination preparation for the core is likely to be applicable for the module, too. Second, it represents what most family physicians do in practice. For those physicians who do subspecialize, it seems that most subspecialize in only 1 area. Allowing these physicians to select only a single module will likely create a greater sense of fidelity to practice by allowing them to either select the AFM module or select a different module that more accurately reflects their practice.

When looking at Figure 2, the actual score line is not directly in the middle of the better and worse lines, meaning the negative impact of the worse module is greater than the positive impact of the better module. So, the score increase in the "better" condition is less of a product of letting examinees select a topic in which they specialize and more a product of not requiring diplomates to select an area of specialization when they do not have one.

### *Limitations*

The primary limitation of this study is that the "better" condition operates under the assumption that physicians can accurately predict the module on which they will perform better. The "worse" condition attempts to mitigate this by showing what would happen if every physician chose the module in which they performed worse. In reality, the answer lies somewhere in between; this is why Figure 2 was designed to illustrate the possible best-case and worst-case scenarios.

This study also only infers the tendency of physicians to specialize and to select modules that reflect their practice. We do not know why physicians choose the modules they do, nor whether those module selections represent alignment with their practice. It is entirely possible that physicians select modules not based on whether they mirror their particular practice; rather, they might choose based on what they enjoy doing, regardless of whether they actually do it on a day-to-day basis.

## Conclusions

This study shows that permitting candidates to select the content category for portions of their examination has a tendency to bias their scores in a systematic way. From a psychometric perspective, this is undesirable: it makes the scale less stable and makes the meaning of the scores dependent on the particular modules selected. From a policy perspective, the desirability of permitting this choice is less clear. Policymakers want the measurement system to be as stable as possible, but they also want to have the largest possible number of candidates agree that the examination was relevant to their practice as physicians. These results suggest that removing 1 module would likely increase both the psychometric stability of the examination and more closely align the content to the practices of more family physicians.

*To see this article online, please go to: http://jabfm.org/content/30/1/85.full.*

## References


1. 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.
2. ABFM certification/recertification examination content. Lexington, KY: American Board of Family Medicine; 2010.
3. O'Neill TR, Shirley K. Psychometric department report to the ABFM Board of Directors. Lexington, KY: American Board of Family Medicine; 2009.
4. O'Neill TR, Peabody MR. Memo on the effect of modules on MC-FP examination results. Lexington, KY: American Board of Family Medicine; 2013.
5. O'Neill TR. Psychometrics Department Update. Presentation to the ABFM examination committee at the October board of directors' meeting. October 5, 2015. Lexington, KY: American Board of Family Medicine.
6. O'Neill TR, Li Z, Peabody MR, Lybarger M, Royal KD, Puffer JC. The predictive validity of ABFM's in-training examination. Fam Med 2015;47:340–56.
7. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research; 1960.
8. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA 2006;296:1094–102.