ORIGINAL RESEARCH

Examining the Construct Stability of the Family Medicine Certification Scale Between One-Day Exam and Longitudinal Assessment

Thomas R. O'Neill, PhD, Keith Stelter, MD, MMM, and Ting Wang, PhD

Purpose: **To determine whether the construct of family medicine clinical decision making ability was invariant across modes of administration, the 1-day examination and the longitudinal assessment. We attempted to identify item characteristics associated with differences in difficulty across modes of administration.**

Methods: **The data were item difficulty calibrations based on examinee responses to the 1-day examination and the longitudinal assessment. A repeated measures design was employed to identify question calibration differences across modes of administration, so that the stability of the question difficulty across modes of administration could be assessed. A qualitative review of the flagged questions was conducted to identify characteristics associated with questions becoming easier or more difficult.**

Results: **The correlation between the pairs of calibrations was moderately positive $r(298) = 0.558$, $P < .001$ suggesting that the questions are functioning somewhat similarly across the different modes of administration; however, the scatterplot demonstrates that many of the questions became easier. Of the 298 repeated measures *t* test, 37% (110) did not show a significant difference, 43% (128) became easier on the longitudinal assessment, and 20% (60) became more difficult.**

Conclusions: **This study suggests that changes in item difficulty do occur when extra time and the use of external resources are permitted. Usually the questions get easier, but in some cases the question becomes more difficult. Possible reasons for this are presented, and a method to adjust the item difficulty in a way to maintain a single construct is presented. (J Am Board Fam Med 2024;00:000–000.)**

Keywords: **Certification, Evaluation Study, Family Medicine, Licensing, Longitudinal Studies, Psychometrics, Research Design, Scales, Statistics**

Introduction

Since 2008, the Family Medicine Certification Scale (FMC-S) has been the basis for the American Board of Family Medicine's (ABFM) certification examination, the Family Medicine Certification Examination (FMCE). The construct that FMC-S is intended to measure an examinee's medical knowledge and clinical decision making ability across the full scope of family medicine. Until 2019, this definition has implied that the medical knowledge was limited to what was available in an examinee's memory because the use of external resources was prohibited during the examination.

With the introduction of the Family Medicine Certification Longitudinal Assessment[1] (FMCLA) in 2019, this notion was implicitly modified by permitting the use of external resources such as books, journals, and intranet searches, but prohibiting collaboration with other people. To permit participants adequate time to look up information in a manner similar to what they might do in practice, the time permitted to answer questions was increased. For example, in each of the 4 sections of the FMCE, examinees must answer 75 questions in 95 minutes which is 1.3 minutes per question. On the FMCLA, examinees have 5 minutes per question. Although the

This article was externally peer reviewed.
Submitted 3 December 2023; revised 3 April 2024; accepted 8 April 2024.
This is the Ahead of Print version of the article.
From the American Board of Family Medicine, Lexington, KY (TRO, KS, TW); University of Minnesota - Mankato Family Medicine Residency Program, Mankato, MN (KS).
Funding: None.
Conflict of interest: None.
Corresponding author: Thomas R. O'Neill, PhD, 1648 McGrathiana Parkway, Suite 550, Lexington, KY 40511 (E-mail: toneill@theabfm.org).

FMCLA is still intended as a measure of an examinee's medical knowledge and clinical decision making ability, it may have subtly incorporated into the construct the ability to research medical topics quickly and accurately. It seems likely that in addition to medical knowledge that a physician possesses which can readily be accessed from memory and one's clinical decision making ability, the FMCLA might also include the skill of "rapid and accurate retrieval of information" as part of what is being measured.

The purpose of the study was to determine whether the construct of family medicine clinical decision making ability as manifested by the hierarchy of item difficulty was invariant across modes of administration. If there were noticeable differences, then we would attempt to identify what item characteristics were associated with the changes in difficulty across modes of administration. In addition, we discuss what were the implications for better standardizing the FMCE and FMCLA. We recognize that some of the concepts used in this article are statistical and psychometric and that these aspects may not be of interest to many of our readers. For readers who would like more background on these concepts, we have included an Appendix with relevant explanations.

Method

Participants

A total of 11,497 family physicians responded to questions on the FMCLA. They were family physicians who passed the FMCE in 2009, 2010, and 2011 and then volunteered to participate in the FMCLA starting in 2019, 2020, or 2021 respectively, as an alternative to taking the FMCE again to maintain their board certification.

Instruments

The FMCE measures physicians' clinical decision making ability as it relates to family medicine. Passing this examination is one of the requirements for ABFM certification. It consists of a common core of 260 multiple-choice questions that contribute to the examinee's score plus 40 pretest questions that are unscored. Before 2020, the FMCE also included 1 or 2 45-question modules that were selected by the examinee. The 260 core questions and the examinee-selected module questions were scored as right or wrong using the dichotomous Rasch[2–4] model, and the resulting ability estimates were converted to scaled scores that range from 200 to 800. In conjunction with a common-item equating design, the Rasch model was also used to equate examinations across test forms and years of administration onto a common scale, the FMC-S. Rasch reliability estimates for the FMCE are typically about 0.94.[5,6] During the time frame from which the data were gathered, the minimum passing score for the FMCE was 380. The content specifications for this examination were developed by Norris et al.,[7] and additional validity studies[8,9] have supported its continued use.

The FMCLA is a multiple-choice-question, longitudinal-assessment that can be taken as an alternative to taking the certification examination[1]. It has the same proportion of questions in each content domain as the FMCE. It delivers up to 25 questions online per quarter, and participants have up to 5 minutes to answer each question. A completed assessment has 300 answered questions over a maximum of 4 years, but it can be completed in 3 years if all the questions are answered each quarter. This allows participants to defer questions (opt out of 4 quarters or answer fewer than 25 items within a quarter), if they so desire. At the end of the administration window, unanswered questions are scored as wrong. FMCLA is not available for initial certification because it takes at least 3 years to complete it. When the participant completes the 300 questions and if a passing score was achieved, it satisfies the examination requirement for maintaining board certification. The FMCLA is calibrated to also be on the FMC-S.

Data

The FMCLA questions used in this analysis were the questions deployed from 2019 to 2021. Of the 300 possible questions, 1 question was deleted without replacement for administrative reasons and a second question did not have a starting calibration to use in the comparison, leading to 298 questions analyzed in total. For each question, there were 3 variables: the FMCLA difficulty, the FMCE difficulty, and the FMCE question's administration category (operational core question, user selected module question, or unscored field test question) when the FMCE difficulty was estimated.

The difficulty calibrations used in this study were different from the operational calibrations used for scoring. To get the most precise and

accurate difficulty calibrations 2 adjustments were made. To maximize the calibrations' precision, we estimated the FMCE calibrations on responses from the entire year, not just the April administration. For the FMCLA calibrations, we used the responses collected after quarter 15 of the 16 possible quarters. To increase the accuracy, all calibrations were adjusted by adding the displacement value to the preassigned calibration. The FMCLA calibrations were adjusted from the preassigned FMCE values by adding the displacement value estimated from the FMCLA administration. The FMCE calibrations were adjusted by adding the displacement value estimated from the FMCE administration in which they were most recently used. Displacement was not added to FMCE pretest calibrations because they had not been used previously.

Design

For this study, a repeated measures design was employed to identify question calibration differences across modes of administration, so that the stability of the question difficulty across modes of administration could be assessed. A qualitative review of the flagged questions was then conducted to identify the characteristics associated with questions becoming easier or more difficult.

Procedure

For each FMCLA question, 2 separate difficulty calibrations were computed using a Rasch model. The first calibration was based on responses from examinees who saw the question when it was most recently administered on the FMCE. The second calibration was based on examinee responses when the question was administered on the FMCLA. The FMCE-based difficulty and the FMCLA-based difficulty of the questions were plotted[10] against each other, a correlation between the 2 was computed, and 298 repeated measures two-tailed t test with a Bonferroni correction were conducted to determine whether there was a significant difference for each pair. A histogram of the difference in difficulty between the 2 scores was also created.

Sub-Analysis

The 298 questions were also classified based on their most recent FMCE calibration as being a pretest item (which would not contribute to examinee scores), an operational item (a scored item from the nonmodule portion of the FMCE), or a module item (an item administered on the FMCE in a user-selected content module). This sub-analysis may shed light on intentionally created platform-related factors that cause differences in difficulty calibrations, such as the ability to look up information and the additional time permitted to answer, as well as, nonintentional factors that might be related to how the examination responses were collected historically, such as the self-selection of the content modules from older FMCE administrations and the difference in statistical power between pretest and operational questions. Across these 3 categories, the FMCE-based difficulty and the FMCLA-based difficulty of the questions were plotted against each other and summarized.

Qualitative Question Review

The questions that were flagged as having a statistically significant difference in difficulty were reviewed by 2 ABFM staff physicians. The task in the analysis was to theorize why the difficulty changed and to identify common themes that emerged for questions that became easier and questions that became more difficult.

IRB Review

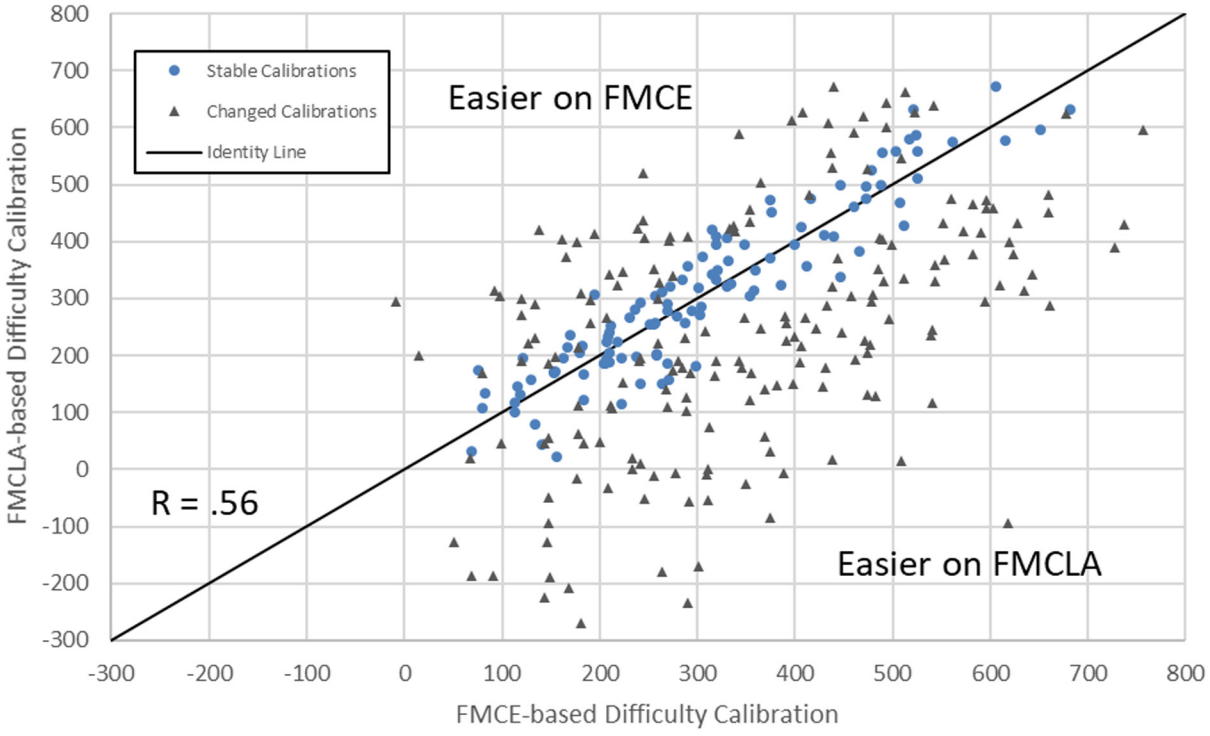
The procedures in this study were reviewed by ABFM executive staff to ensure that ABFM privacy policies were not being violated. In addition, the data were deemed exempt by the American Academy of Family Physicians Institutional Review Board.

Results

Scatterplot and Correlation

Figure 1 illustrates the relationship between the 2 modes of administration, FMCE and FMCLA, for the 298 pairs of difficulty calibrations. The Pearson correlation between the pairs of calibrations was moderately positive $r(298) = 0.558$, $P < .001$ suggesting that the questions are functioning somewhat similarly across the different modes of administration; however, the scatterplot demonstrates that many of the questions became easier. Of the 298 repeated measures t test (Table 1), 37% (110) did not show a significant difference, 43% (128) became easier on FMCLA, and 20% (60) became more difficult.

Figure 1. Comparison of FMCE-based and FMCLA-based difficulty calibrations. ***Abbreviations:*** **FMCE, Force and Motion Conceptual Evaluation; FMCLA, Force and Motion Conceptual Learning Assessment.**



Sub-Analysis

We further subdivided Figure 1 by the origin of the question's starting calibration (Figure 2, Table 1). For questions that became easier to a statistically significant degree, the percentage of questions hovered around 43% across all 3 calibration origin categories.

For questions that became more difficult, there was a substantial difference in the percentage of questions between the *Operational* category and both the *Pretest* and the *Module* categories. A higher percentage became more difficult in the *Operational* category than the other 2 categories. For questions with no change in difficulty, the reverse of this was true. A lower percentage had no change in the *Operational* category than the other 2 categories.

The magnitude of the questions' changes in difficulty between FMCE and FMCLA administrations is displayed in Figure 3. Almost half of the questions demonstrated a changed in difficulty within ±100 scaled score points, but overall, more questions' calibrations were harder in FMCE.

Qualitative Question Review

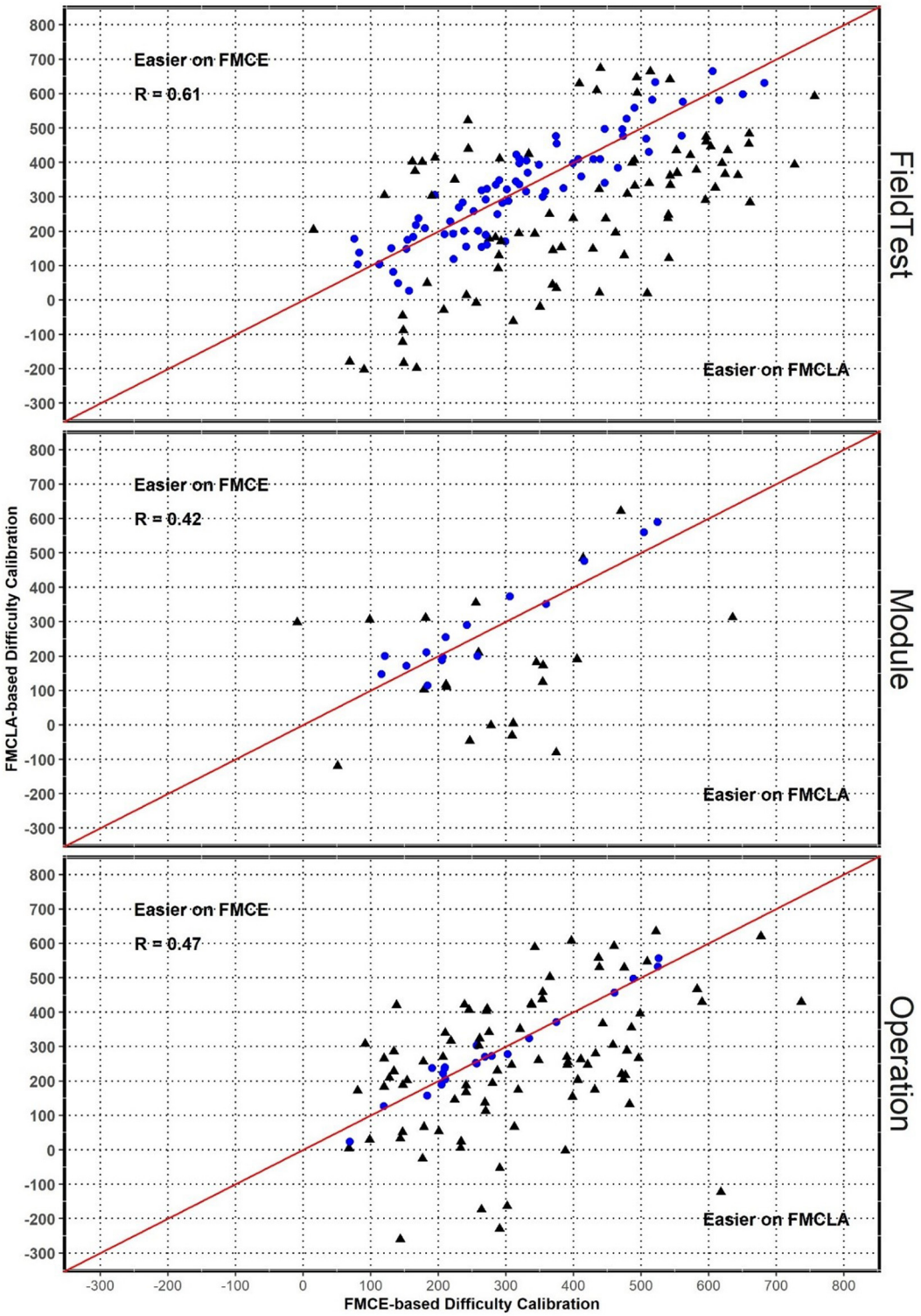
Of the 298 examination questions studied (Table 1), 188 questions manifested a statistically significant degree of change with 128 becoming easier

Table 1. Summary of the Change in Difficulty When Questions Were Administered in FMCLA

	Operational		Pretest		Module		Total	
Significant Change	Count	Percent	Count	Percent	Count	Percent	Count	Percent
No Change	21	19.8%	74	47.4%	15	41.7%	110	36.9%
Harder on FMCLA	35	33.0%	19	12.2%	6	16.7%	60	20.1%
Easier on FMCLA	50	47.2%	63	40.4%	15	41.7%	128	43.0%
Total	106	100.0%	156	100.0%	36	100.0%	298	100.0%

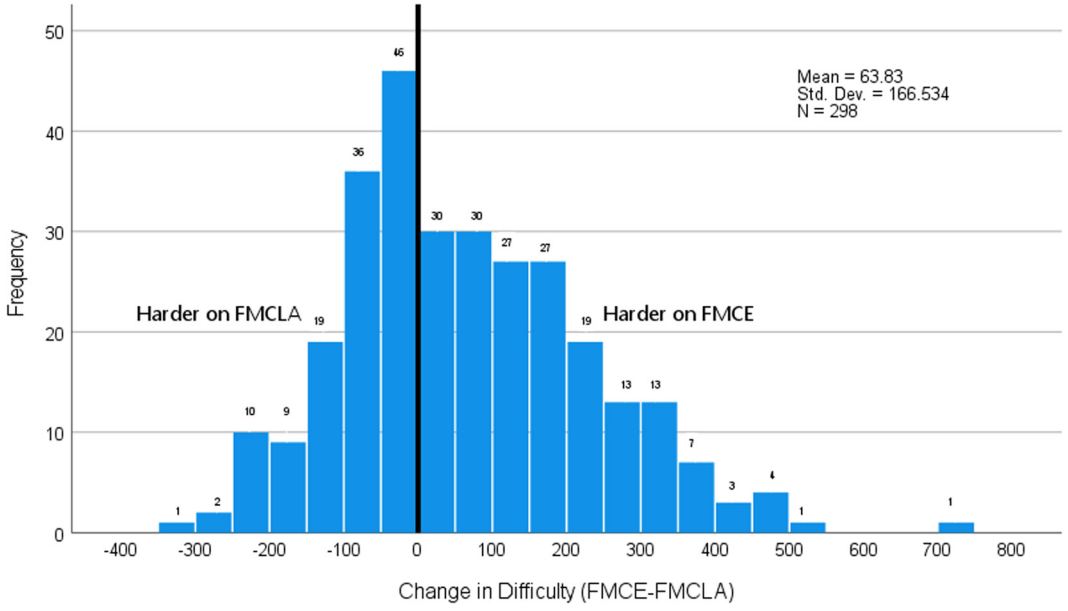
Abbreviation: FMCLA, Force and Motion Conceptual Learning Assessment.
Notes: Significance tests were conducted using a z-score with Bonferroni correction.

Figure 2. Scatter plot of item difficulty calibrations: FMCE-based vs FMCLA-based. ***Abbreviations:*** **FMCE, Force and Motion Conceptual Evaluation; FMCLA, Force and Motion Conceptual Learning Assessment.**



Note: • represents stable calibrations ▼ represents unstable calibrations

Figure 3. Histogram of item difficulty calibration differences across modes of administration (FMCE vs FMCLA). ***Abbreviations:*** **FMCE, Force and Motion Conceptual Evaluation; FMCLA, Force and Motion Conceptual Learning Assessment.**



to answer and 60 becoming more difficult. These questions were reviewed by 2 ABFM staff physicians. Below are some possible reasons for the change.

Of the questions that became easier on FMCLA, almost all could be looked up within 2 to 3 minutes using an accessible internet search engine such as Google or Bing or perhaps doing a PubMed search with key words from the question. Attempting to access an online medical textbook likely would be slower because more reading would be required as they are generally not indexed as well for specific key words. In almost all cases the answers either could be answered or narrowed significantly with quick internet searches.

For the questions that became more difficult, it was challenging to establish a single unifying reason. The reviewers were able to internet search many of these and find some of these answers and could narrow down the possible answers; however, sometimes this required a more complicated search strategy with multiple layered searches. Most of these questions were more complex clinical presentations which required more reading and more data synthesis to consider in creating a search strategy that would be difficult to complete within the allocated 5-minute time frame.

Discussion

In this study, we found a majority (63%) of the questions' difficulty changed by a statistically significant degree across modes of administration. The correlation between the calibrations on FMCLA and FMCE was moderate (R = 0.56). Although it was not surprising that the correlation was not negative or near zero, it was also noticeably lower than what is considered a high correlation, such as 0.8 or 0.9. This suggests that the item hierarchy changed somewhat. We strongly suspect that these detectable differences in the item hierarchy are caused by getting extra time and using external resources. When extra time and external resources are not permitted, then answering the question correctly could be more difficult. In these cases, a correct answer would imply the examinee has more ability. So how can these 2 modes of administration even when using the same questions be made to be comparable?

The construct being measured in both administration platforms is intended to be *clinical decision making in family medicine*. Although the text of the question can be identical across platforms, the process of answering it can be very different. In cases where the question functions differently across platforms, it is treated as if it is a new question, and it is given a new calibration based on the current data

set. When the new difficulty calibration is used in scoring, the change in the test form's difficulty is adjusted to accommodate the change.The same passing standard still applies to both the FMCE and the FMCLA and neither offers a scoring advantage.

This raises the following question: Which of these 2 examination conditions (1 with extra time and searchable resources or 1 without) is preferable for making pass-fail decisions on a standardized test of family medicine? If the skill of "rapid and accurate retrieval of medical information" is important, then searching capabilities should also be included on the FMCE. If it is not, then it makes sense to prohibit using external research tools on the FMCLA. Given that most physicians use of point of care searches in clinical practice, it could be advantageous to permit search capabilities as it would likely better reflect the process of delivering care in family medicine. On the other hand, an examination that requires memory retrieval rather than the use of searchable resources does identify those people with serious medical knowledge deficits. Yet another possibility is to write questions that tap into aspects of family medicine that are less affected by the use of external resources. There could also be other approaches as well.

Item Writing

Regarding the item writing style, a content review of the flagged questions revealed those that became easier often included diagnostic keywords that would considerably facilitate a computer-based search. Ideally, answering the questions using a search should require some physician-level family medicine knowledge to discern the correct answer. In other words, someone with only technical computer skills, but no medical training should not be able to answer the questions correctly. The newly emerging artificial intelligence models like ChatGPT and others make this issue more salient than ever before.

Limitations

There are at least 4 limitations to this study. First, half of the study's data collection window was during the COVID-19 pandemic. Second, it is based only on questions available during the first 3 years of FMCLA. These first 2 limitations do not seem terribly impactful because we have seen similar results with subsequent, post-COVID cohorts. Third, the results of this study are only generalizable to longitudinal assessments that are similar in design to FMCLA. Finally, the sub analysis demonstrated that there was an impact of how many questions were flagged for differences probably due to the differences in sample size and statistical power associated operational, pretest, or module status. The ratio between becoming easier to harder on FMCLA was 2:1. However, one of the findings in the sub-analysis was that the percentage of questions that became easier, harder, or stayed the same was different depending on the source of the anchor item's calibration. A higher percentage of questions anchored to the operational section of the FMCE displayed significant changes (both easier and harder) as compared with the questions calibrated on the module and pretest sections of the FMCE. It is probable that these observed differences are the result of differences in the precision of the anchor calibrations across these different sections of the FMCE. The anchor calibrations from the FMCE's pretest and module sections are based on fewer responses than the calibrations from the operational section. The operational section calibrations are typically based on 2,500 to 15,000 responses. In contrast, the pretest calibrations and the modular calibrations are typically based on 160 to 960 and 50 to 7,800 responses, respectively. The increased precision and larger sample size of the operational questions' calibrations may have provided additional statistical power to detect differences. Although this limits the generalizability of results from the full set of 298 questions, it does seem point out that calibration differences associated with content modules and pretest items tends to inflate the number of questions in which there was no detectable change probably due to the smaller sample sizes for those calibrations.

Conclusion

In conclusion, this study found that the recalibration of test questions when used across test platforms is necessary to make the associated examinee scores comparable across platforms. It also suggests that the best anchor items to use for equating test forms are those that have a large number of responses on the previous administration. It also suggests that questions that can be quickly looked up using key words from the question in a common

search engine (Google, Bing, Pubmed, etc.) tend to get easier on longitudinal assessment.

To see this article online, please go to: http://jabfm.org/content/00/00/000.full.

References

1. O'Neill TR, Newton WP, Brady JE, Spogen D. Using the Family Medicine Certification Longitudinal Assessment to make summative decisions. J Am Board Fam Med 2019;32:951–3.
2. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark.: Danish Institute for Educational Research; 1960.
3. Linacre JM. Winsteps Rasch Measurement computer program [Internet]. Beaverton, OR.: Winsteps.com; 2022. Available at: https://www.winsteps.com/index.htm.
4. Wright BD, Douglas GA. Best Test Design and Self-Tailored Testing. Memo No 19 MESA Psychom Lab Univ Chic. 1975.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. Luppescu S. DIF: graphical diagnosis. Rasch Meas Trans 1991;5:136.

Appendix

Measurement Concepts

In this article, item difficulty refers to a difficulty calibration from a dichotomous Rasch model, rather than the percentage of a reference group that answered the question correctly. Rasch models describe the difficulty of a question relative to the difficulty of the other questions on the scale. Although the unit of measure in Rasch models is the logit, the item calibrations have been converted to the FMC-S to make the difficulty more understandable. Higher scaled score calibrations indicate more difficult questions. To accommodate this journal's diverse readership, this appendix has been included to provide explanations of Rasch models and equating. If the reader is not concerned about these concepts, the appendix can be safely ignored.

Rasch Models

Rasch models are measurement models for unidimensional latent traits. They convert ordinal observations, such as right-wrong answers or rating scale responses into interval scale measures of the latent trait for people and interval scale difficulty calibrations for the questions.[1–3] This results in a hierarchy of questions that range from easy to difficult with each question having its own difficulty calibration and an error term representing the precision of that calibration. Most importantly, both the person ability estimates, and the item difficulty calibrations are computed in such a way that the distribution of person ability in the sample taking the test does not change the difficulty of the questions,[4] and the distribution of item difficulty on the test does not change the ability estimates for the people testing, so long as the responses fit the model's expectations. This separation of parameters is necessary for measurement.[5–8]

Equating

To allow for the exclusion of outdated questions, the inclusion of new questions, and for general test security reasons, multiple forms of an examination are often required. To ensure that examinees are neither advantaged nor disadvantaged by any test form, statistical procedures are conducted to place the scores from all the forms on a common scale. These processes are collectively referred to as equating procedures. Although test forms usually refer to different subsets of items from the item bank, they could also refer to a difference in the mode of administration. For this study, equating refers to both the different sets of questions being administered and the mode of administration.

Typically, certification boards use a common item equating design that employs "anchor items," questions with known difficulties from previous examination administrations, to connect the different examination forms. The anchor items' difficulties are then preassigned for use in the current administration, not estimated from the current data set. These anchor items are expected to have very similar difficulties across the forms or modes of administration. If the anchor items perform drastically differently, then they are not used to connect the 2 forms and the difficulty of those questions are freely estimated using the data from the current administration. The statistic used to assess the similarity of the difficulty is called displacement. Displacement is the difference between the difficulty calibration that would have been estimated based on the current data and the preassigned difficulty calibration.

References

1. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research; 1960.
2. Wright BD, Douglas GA. Best test design and self-tailored testing. Memo No 19 MESA Psychometric Laboratory University of Chicago. Published online 1975.
3. Linacre JM. *Many-Facet Rasch Measurement*. MESA; 1989.
4. O'Neill TR, Wang T. Item calibration invariance across samples with extreme ability differences. RMT 2022;35:1883–5.
5. Luce RD, Tukey JW. Simultaneous conjoint measurement: A new type of fundamental measurement. Journal of mathematical psychology 1964;1:1–27.
6. Wright BD. Sample-free test calibration and person measurement. paper presented at the National Seminar on Adult Education Research (Chicago, February 11-13, 1968). Published online 1967.
7. Bond TG, Fox CM. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Lawrence Erlbaum; 2001.
8. Andrich D. Controversy and the Rasch model: a characteristic of incompatible paradigms? Med Care 2004; 42:I7–16.