Using the Family Medicine Certification Longitudinal Assessment to Make Summative Decisions

Thomas R. O'Neill, PhD, Warren P. Newton, MD, MPH, John E. Brady, MD, and Daniel Spogen, MD

(J Am Board Fam Med 2019;32:951–953.)

On January 4, 2019, the American Board of Family Medicine (ABFM) launched the Family Medicine Certification Longitudinal Assessment[1] (FMCLA) pilot. Our hope is that FMCLA will provide both summative feedback—assessing whether a candidate has the cognitive expertise to be a board-certified family physician—as well as formative feedback—to help diplomates know more accurately what they do not know and, thus, focus their learning. With respect to the formative component, early reports are very positive. Of the eligible diplomates, 71% took advantage of the pilot. The technology platform is functioning well. Very few diplomates have withdrawn and many report that the tool is helping them learn. Evaluation from this quarter and the next will begin to give us a better understanding of how FMCLA fits into the other ways diplomates learn, and we will explore new formats of reports to support diplomates' learning efforts.

What about the summative assessment component? This editorial summarizes our strategy for assessing the validity of FMCLA as a summative assessment. A little background may be helpful. Since 1970, ABFM has used a standardized examination, the Family Medicine Certification Examination (FMCE), to periodically assess whether the candidate has the breadth and depth of cognitive expertise to be board certified. We use Rasch[2,3] item response models to create and maintain a single scale with which to measure this cognitive expertise in family medicine. This scale is used with the FMCE,[4] the In-Training Examination,[4] and the Continuous Knowledge Self-Assessment.[1] This single scale used in conjunction with the minimum passing standard defines the minimally acceptable ability level for a diplomate.

FMCLA uses questions from the same question bank as the FMCE, but the method of administration is very different, given that candidates have 5 minutes per question and an "open book" as opposed to 75 seconds per question and no references. Because the standard needs to be the same for all family physicians, it is very important for ABFM to demonstrate that what is being measured is comparable and that the comparable pass-fail decisions are being made. To do this, we will examine whether the questions function comparably and whether the scores and pass-fail decisions are comparable at both the individual and population levels.

Question comparability

The commonness of the common ABFM scale arises from all the questions being identified as belonging to one of the FMCE's blueprint[5] categories and then being placed within a hierarchical framework by the question's difficulty. This question difficulty is how difficult a question is relative to all the other questions within the framework. Although examinee responses are used to compute this difficulty, the difficulty is relative to the other questions, not relative to the people. This requires estimating the difficulty of each question in a way that experimentally removes the ability level of the particular person answering it. For this reason, the dichotomous Rasch[2] model, a logistic model of probabilities that is sample-free[6,7] when estimating

Conflict of interest: The authors are employees of the ABFM.

question difficulties and item-free when estimating person abilities, is used. The hierarchy of questions, therefore, describes what is very easy, what is very difficult, and what lies between those 2 extremes. These sample-free estimates of item difficulty are referred to as calibrations.

An examinee's scaled score, which is based on the examinee's number of correct answers and the difficulty of those questions, is the examinee's position within this hierarchy. Placing both questions and people within the same hierarchy permits an understanding of the question difficulty-level that a person with a particular scaled score is likely to answer correctly.

In comparing FMCE and FMCLA, we seek to answer: does the difficulty of the question depend on the mode of administration? It would not be surprising if candidates were more likely to get a specific question correct in FMCLA, in which they have more time and access to references. The important issue, however, is the question difficulty hierarchy. If the hierarchy is preserved despite access to references, then what we are measuring is the same. The good news so far is that, as of this writing, this seems to be the case. In coming quarters, we will continue to assess this comparability of items. We will also attempt to describe the characteristics of items that got easier and items that became more difficult. If we find that certain types of questions are easier when there is more time to answer, ABFM will have to identify why and address what that means for the type of questions that are asked on the examination. If there are questions that are easily answered by looking them up, ABFM may need to adjust the process for developing examination questions.

Population score comparability

The comparability of scores for individuals can be hard to evaluate when 1 measure is at a single point in time and the other measure is accrued over years. Looking at score distributions for populations can provide another perspective. How does the examination perform given the entire population of candidates? Does the test given under different modes of administration perform similarly—passing and failing a similar proportion of people?

To address this question, we will compare cohorts of people who took the 2019 FMCE and FMCLA. We will not include initial certifiers, diplomates who were testing before the year in which their certification would expire, and those who were testing to regain their certification. This cohort's scores will also be compared with the FMCE scores of cohorts from 2015 to 2018 that meet similar selection criteria. This will include a comparison of the means for these groups and of the proportions of people above and below the passing standard.

Person score stability

FMCLA tests cognitive expertise over 4 years. Does cognitive expertise of a candidate vary significantly over time? If so, this might call into question the comparability of scores between the FMCE and the FMCLA. To assess the stability of cognitive expertise over time, we have recruited a cohort of volunteers who took and passed the 1-day examination in November 2018 and are not required to take the examination again for another 10 years. These volunteers will take the FMCLA, although without being subject to failing. Using a repeated measures design, we will compare scores of diplomates taking the test under both conditions. Although we will conduct the analysis at the end of the first year, we envision this facet of the assessment of the FMCLA pilot program to continue for a second year. We will need to be aware that some participants might not take their FMCLA participation seriously or look up every question they receive, which could noticeably increase their score from their 2018 FMCE score. Although participants get immediate feedback on each question regarding whether they answered it correctly or not, their performance is not transformed into a scaled score until the beginning of their second year. Not having this information might influence how they respond to the questions. For this reason, second-year responses might better represent how participants will respond over a longer period than the first-year responses. Teasing these aspects out of the data may be difficult.

The assessment of the comparability of FMCLA and FMCE is not exact or based on a statistical test but rather is a judgment based on a number of lines of evidence. We anticipate having an initial judgment in the first half of 2021; we will report it here and in the peer-reviewed literature.

It is important to underscore that assessment of cognitive expertise is only 1 part of board certification by the ABFM. In addition to cognitive expertise, as measured by FMCE or FMCLA, we

expect successful candidates to meet our standards for professionalism (a full, active and unrestricted license), to demonstrate commitment to lifelong learning and self-assessment through ongoing continuing medical education and Knowledge Self-Assessment activities, and achieve at least 1 performance improvement activity every 3 years if clinically active. In a given year, only a small percentage of diplomates lose their certificates but as the result of each part of the portfolio. For example, in 2018, 0.09% lost their certificate due to a lapse of professionalism, 0.7% because of Knowledge Self-Assessment, 6% either did not take or did not pass the examination, and 1.0% did not do the performance improvement activity. So, the examination is an important hurdle but not the only hurdle.

Board certification continues to play an important role in protecting the public. Although the vast majority of physicians provide quality care and stay up to date, there is a small percentage of physicians who are responsible for a disproportionately large number of problems. Recent studies have reported that 1% of US physicians are responsible for one-third of paid medical malpractice claims,[8] that 3% of the Australian medical workforce accounted for 49% of complaints,[9] and that 10% of Canadian physicians accounted for 20% of licensure actions.[10] Medical certification provides a standardized, unbiased, third-party attestation that a physician is meaningfully and successfully engaged in all the activities that the profession considers crucial for meeting professional standards.

The ABFM has a proud 50-year tradition of thoughtfully considering the rationale for changes in our certification program before implementing them. Our overall and ongoing goal is to demonstrate the value and validity of our certification program to the public, to our diplomates, and to the family medicine community. In the case of FMCLA, our first concern must be the comparability with FMCE to assess cognitive expertise and, thereby, to protect the public. We look forward to reporting back to you what we find. What we have described here is just the beginning.

To see this article online, please go to: http://jabfm.org/content/32/6/951.full.

References

1. Newton WP, Rode K, O'Neill T, Fain R, Baxley E. Longitudinal assessment: where we are and why it is important. J Am Board Fam Med 2019;32:448–50.
2. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research; 1960.
3. Wright BD, Stone MH. Best test design. Chicago: MESA Press; 1979.
4. 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;47:349–56.
5. 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.
6. Wright BD, Panchapakesan N. A procedure for sample-free item analysis. Educ Psychol Meas 1969;29:23–48.
7. Wright BD. Fundamental measurement for outcome evaluation. Phys Med Rehabil State Art Rev 1997;11:261–88.
8. Studdert DM, Bismark MM, Mello MM, Singh H, Spittal MJ. Prevalence and characteristics of physicians prone to malpractice claims. N Engl J Med 2016;374:354–62.
9. Bismark MM, Spittal MJ, Gurrin LC, Ward M, Studdert DM. Identification of doctors at risk of recurrent complaints: a national study of healthcare complaints in Australia. BMJ Qual Saf 2013;22:532–40.
10. Alam A, Klemensberg J, Griesman J, Bell CM. The characteristics of physicians disciplined by professional colleges in Canada. Open Med 2011;5:e166-e172.