Examining the Functioning and Reliability of the Family Medicine Milestones

Michael R. Peabody, PhD
Thomas R. O'Neill, PhD
Lars E. Peterson, MD, PhD

ABSTRACT

Background The Family Medicine (FM) Milestones are a framework designed to assess development of residents in key dimensions of physician competency. Residency programs use the milestones in semiannual reviews of resident performance from entry toward graduation.

Objective To examine the functioning and reliability of the FM Milestones and to determine whether they measure the amount of a latent trait (eg, knowledge or ability) possessed by a resident or simply indicate where a resident falls along the training sequence.

Methods This study utilized the Rasch Partial Credit model to examine academic year 2014–2015 ratings for 10 563 residents from 476 residency programs (postgraduate year [PGY] 1 = 3639; PGY-2 = 3562; PGY-3 = 3351; PGY-4 = 11).

Results Reliability was exceptionally high at 0.99. Mean scores were 3.2 (SD = 1.3) for PGY-1; 5.0 (SD = 1.3) for PGY-2; 6.7 (SD = 1.2) for PGY-3; and 7.4 (SD = 1.0) for PGY-4. Keyform analysis showed a rating on 1 item was likely to be similar for all other items.

Conclusions Our findings suggest that FM Milestones seem to largely function as intended. Lack of spread in item difficulty and lack of variation in category probabilities show that FM Milestones do not measure the amount of a latent trait possessed by a resident, but rather describe where a resident falls along the training sequence. High reliability indicates residents are being rated in a stable manner as they progress through residency, and individual residents deviating from this rating structure warrant consideration by program leaders.

Introduction

In 2012, the Accreditation Council for Graduate Medical Education (ACGME) introduced the Next Accreditation System, of which a primary component is the educational milestones.[1] The milestones are organized around the 6 ACGME core competencies and describe attributes that residents are expected to demonstrate as they progress through their program. The Family Medicine (FM) Milestones were implemented in July 2014 and were designed to provide a framework to assess the development of residents in key dimensions of physician competency. Each of the 22 items consist of 6 levels representing the progression from preresidency to master physician, with Level 4 representing the target for independent practice.

There has been little research conducted examining the functioning of the milestones. Swing et al[2] examined the development process and content validity claims for the Emergency Medicine (EM) Milestones, and while they found sufficient evidence for content validity, they called on future research to utilize milestone data to provide further validity evidence. Beeson et al[3] examined the validity and reliability of the EM Milestone ratings to determine the degree to which subcompetencies were interrelated. They found that they were able to discriminate between residency program years and concluded that the EM Milestones "demonstrated validity and reliability as an assessment instrument for competency acquisition."[3]

Our study builds on this body of research by examining the functioning and reliability of the FM Milestones. It is important for both residency program directors and the ACGME to understand how the milestones function, and their capacity for providing useful information about resident and residency performance. In particular, this study seeks to determine whether the FM Milestones are measuring the amount of a latent trait (eg, knowledge or ability) possessed by a resident or simply indicating where along the training sequence a resident falls.

Methods

Participants

This study used end-of-year FM Milestone ratings for all residents in ACGME-accredited FM programs for the 2014–2015 academic year (10 563 residents from a total of 476 programs: postgraduate year [PGY] 1 = 3639; PGY-2 = 3562; PGY-3 = 3351; PGY-4 = 11)

DOI: http://dx.doi.org/10.4300/JGME-D-16-00172.1

provided by the ACGME to the American Board of Family Medicine.

Instrumentation

The FM Milestones encompass 22 items across 6 domains: patient care (5 items), medical knowledge (2 items), systems-based practice (4 items), practice-based learning and improvement (3 items), professionalism (4 items), and interpersonal and communication skills (4 items).[4] For each item, there are 10 rating scale categories representing 6 levels (0 to 5), with 4 categories representing half-point categories (the point at which a resident demonstrates all of the milestones in the lower levels and some of the milestones in the higher levels). The FM Milestones can be found online.[4]

This research was approved without restrictions by the American Academy of Family Physicians Institutional Review Board.

Analysis

We applied a Rasch measurement model[5] to the FM Milestone data. Because each item has its own distinct set of category descriptors, the analysis was conducted using the Rasch Partial Credit model[6,7] available in Winsteps Rasch Measurement Software version 3.81.0 (Winsteps, Beaverton, OR). We examined data model fit using information weighted (INFIT) and unweighted (OUTFIT) mean square values (MNSQ).

These statistics provide an indication of the amount of useful information provided by an item. Although there are no concrete rules about the acceptable thresholds for INFIT MNSQ and OUTFIT MNSQ, values between 0.5 and 1.5 are generally considered acceptable for use.[8,9] In addition, reliability and mean scores by resident program year are provided.[10,11]

Measurement implies a linear, hierarchical construct that is subdivided into equal interval units on which some objects of measurement (eg, individuals) possess more of the construct and others less. Because Rasch models place item difficulty calibrations and person ability estimates on the same scale, we were able to examine this relationship visually using an item-person map. The distribution of people is shown on the left side and the distribution of items is shown on the right side. Ideally, these distributions will be sufficiently well targeted to each other to allow for adequate discrimination.[10] Items located at a similar place along the continuum may be redundant, and large gaps may indicate a place where an additional item is needed.

Increasing amounts of the latent trait correspond to an increasing probability of a person receiving a rating in a higher category, such that, as a person advances along the ability spectrum, each category in turn must be the most probable.[11] For this, category probability curves were created showing person ability relative to item difficulty on the x-axis and the probability of observing each ordered category plotted on the y-axis, such that each category has its own probability distribution. As the person's ability increases from left to right along the x-axis, each category should at some point be the most probable; that is, each category should have its own distinct peak and the entire chart should resemble a mountain range. Categories that are never the most probable, often referred to as "submerged" categories because they are located beneath other categories, contribute little to the rating scale. These category probability curves allow us to visually determine which rating scale categories are providing useful information.

Finally, a Keyform was created to illustrate the relationship between the expected category responses for each item. Person ability estimates were placed on the x-axis, and items were placed along the y-axis with expected rating categories for each item plotted. Keyforms help us to visually predict someone's rating on an unobserved item based on their ratings on observed items or to identify anomalous response patterns.

What was known and gap
The Family Medicine (FM) Milestones were designed to assess resident progression toward unsupervised practice.

What is new
A study of the properties of FM Milestone ratings for 10 563 residents from 476 programs.

Limitations
As residencies get more acquainted with creating ratings over time, scoring patterns may change.

Bottom line
Lack of variability in the FM Milestones suggests that they describe resident progress in the educational program, and deviations from the structure by individual residents warrant consideration by program leadership.

Post Hoc Analysis

Because the 10-point rating scale produced a substantial number of submerged categories, a post hoc analysis was conducted in which half-point categories were collapsed down into the nearest full-level category. This collapsed 6-category rating scale reflected the original theoretical progression levels rather than the extended 10-category rating scale structure.

TABLE
Item Measures, Standard Errors, and Fit Statistics

Domain	Item	Measure	SE	INFIT MNSQ	OUTFIT MNSQ
Patient care	PC_Q1	−0.2232	0.0123	0.79	0.80
Patient care	PC_Q2	−0.1602	0.0124	0.66	0.66
Patient care	PC_Q3	0.0700	0.0129	0.70	0.71
Patient care	PC_Q4	0.0212	0.0123	0.74	0.74
Patient care	PC_Q5	0.3089	0.0123	1.14	1.16
Medical knowledge	MK_Q1	0.2570	0.0128	1.30	1.32
Medical knowledge	MK_Q2	0.0553	0.0126	0.67	0.67
Systems-based practice	SBP_Q1	−0.0862	0.0125	0.83	0.83
Systems-based practice	SBP_Q2	0.3857	0.0127	0.96	0.93
Systems-based practice	SBP_Q3	0.3532	0.0129	1.41	1.40
Systems-based practice	SBP_Q4	−0.3836	0.0124	0.77	0.78
Practice-based learning and improvement	PBLI_Q1	0.7242	0.0121	1.25	1.27
Practice-based learning and improvement	PBLI_Q2	0.0978	0.0127	0.85	0.84
Practice-based learning and improvement	PBLI_Q3	0.9066	0.0126	1.41	1.41
Professionalism	PROF_Q1	−0.1400	0.0122	1.09	1.10
Professionalism	PROF_Q2	0.2411	0.0121	1.49	1.48
Professionalism	PROF_Q3	−0.3493	0.0127	1.01	1.03
Professionalism	PROF_Q4	−0.2519	0.0126	1.12	1.11
Interpersonal and communication skills	C_Q1	−1.0602	0.0127	1.02	1.03
Interpersonal and communication skills	C_Q2	−0.2796	0.0125	0.79	0.79
Interpersonal and communication skills	C_Q3	−0.3440	0.0125	0.85	0.85
Interpersonal and communication skills	C_Q4	−0.1425	0.0122	1.21	1.20

Abbreviations: INFIT, information weighted; MNSQ, mean square values; OUTFIT, information unweighted.
Note: INFIT and OUTFIT MNSQ are chi-square statistics divided by their degrees of freedom and reported as ratios with an expected value of 1 and a range of 0 to infinity.

Results

Individual fit statistics for each item are shown in the TABLE. There were no items for which INFIT or OUTFIT values were lower than 0.5 or higher than 1.5, indicating acceptable data model fit (TABLE). The mean scores were 3.2 (SD = 1.3) for PGY-1, 5.0 (SD = 1.3) for PGY-2, 6.7 (SD = 1.2) for PGY-3, and 7.4 (SD = 1.0) for PGY-4. Reliability, an index of internal consistency similar to a Cronbach's alpha or KR-20, was 0.99.

The item-person map (FIGURE 1) illustrates the construct of the FM Milestones. Person ability estimates ranged from –10 to 10 logits with a mean of 0.77 logits and standard deviation of 2.73 logits. The item difficulty calibrations ranged from –1 to 1 with a mean of 0.00 logits (as imposed by the model) and a standard deviation of 0.41 logits. FIGURE 1 shows very little spread in the item difficulty, meaning that all items are of similar difficulty.

The Keyform (FIGURE 2) illustrates the relationship between the expected rating categories for each item. Keyforms typically have a step-like structure because the difficulty of the items usually varies to a noticeable extent; however, when items do not vary in difficulty, the categories look more like columns than steps. This Keyform is more column-like than step-like, indicating that the items and rating scales are functioning in a near identical manner.

FIGURE 3 provides an illustration of the category probability curves for examining the assumption that, at some point along the ability spectrum, each category will be the most probable. Only 3 items met this assumption: 11 items had 1 category, 6 items had 2 categories, and 2 items had 3 categories that were never the most probable. Of the 29 instances of these nonprobable (submerged) categories, 28 (97%) were half-point categories.

Post Hoc Analysis

Because 19 of the 22 items had at least 1 submerged category and 97% (28 of 29) of the submerged categories were half-point categories, we conducted a post hoc analysis in which the half-point categories were collapsed such that a 1.5 became a 1, 2.5

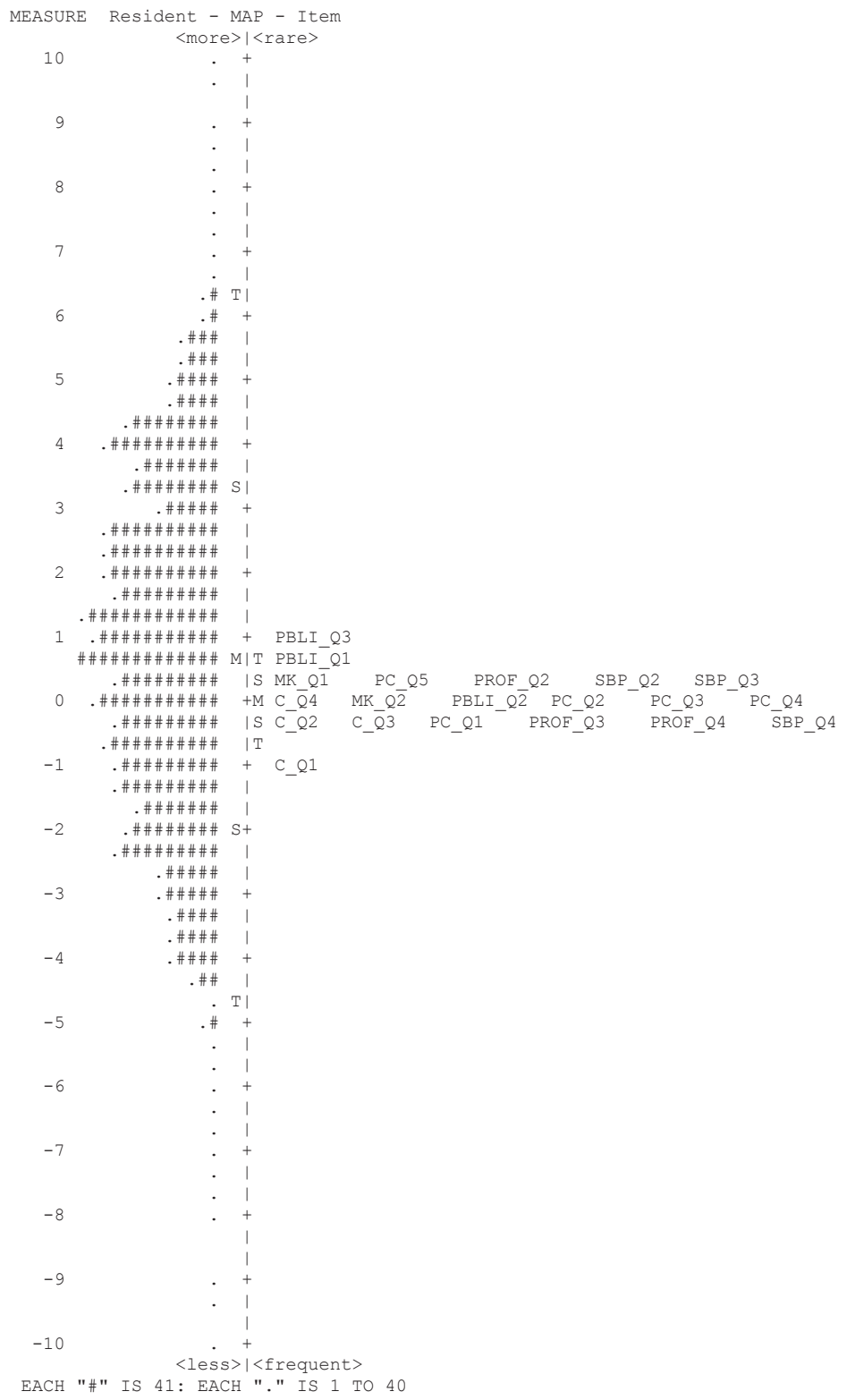

FIGURE 1
Item-Person Map Illustrating Relationship Between Item Difficulty and Person Ability Estimates

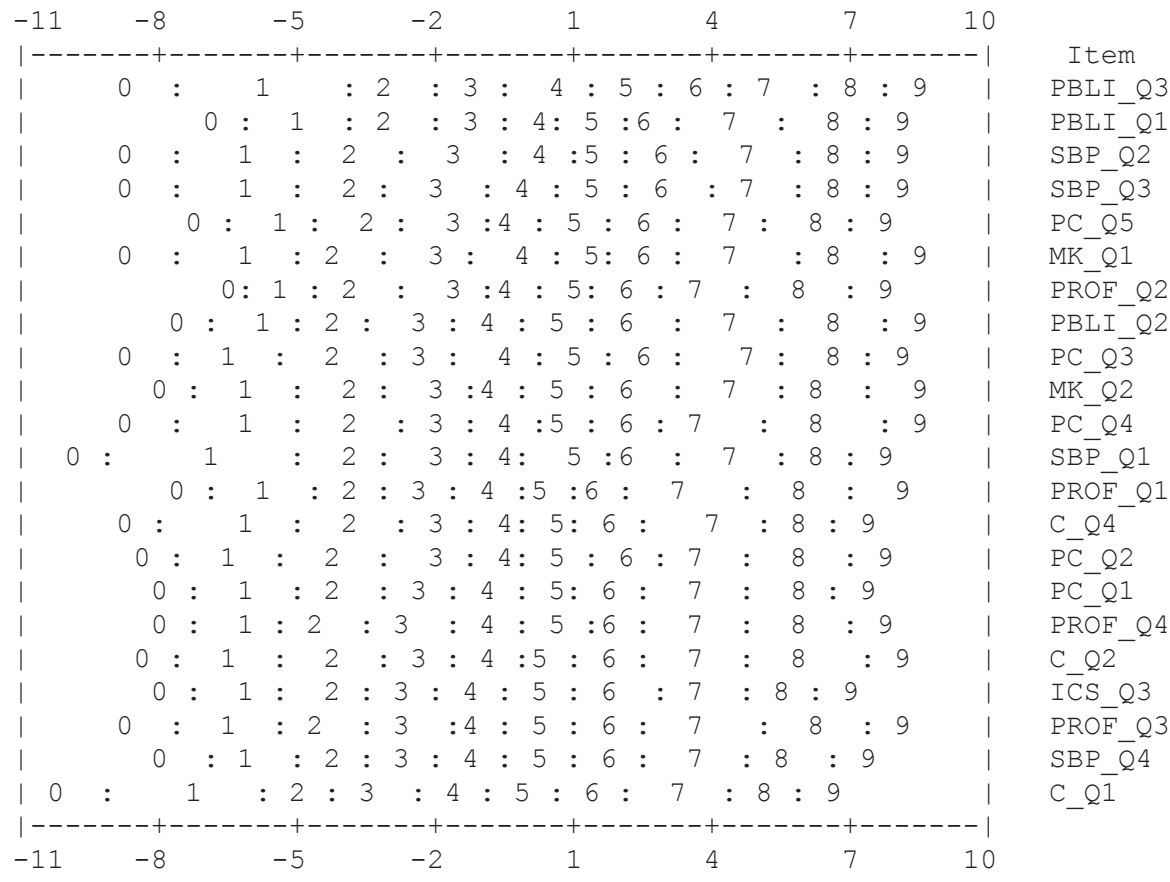

FIGURE 2
Keyform Illustrating Relationship Between Expected Response Categories for Each Item

became a 2, and so on. The resulting analysis provided a reliability of 0.98 and new category probability curves, as shown in FIGURE 4. The collapsed categories produced category probability curves with no submerged categories.

Discussion

The FM Milestones were designed "to create a logical trajectory of professional development in essential elements of competency."[1] Our findings suggest that they seem to function as the designers intended. The lack of spread in item difficulty (FIGURE 1) and the near-deterministic usage of the rating scale by raters (FIGURE 2) indicate that the FM Milestones are not measuring the amount of a latent trait (eg, knowledge or ability) possessed by a resident, but rather indicate where along the training sequence a resident falls. The extraordinarily high reliability shows that residents have no individual differences other than their year in residency and that residents whose ratings deviate from their associated year of residency warrant additional consideration by program leaders.

Rating Scale Functioning

Our findings suggest that the half-point categories provide little additional information and should be either eliminated or given richer descriptors in order for raters to effectively discriminate between categories. Although no items in the original analysis exhibited misfit to a degree that they should be removed, there were 3 items that caused some concern: SBP_Q1 (Provides cost-conscious medical care), PBLI_Q3 (Improves systems in which the physician provides care), and PROF_Q2 (Demonstrates professional conduct and accountability). When the categories were collapsed, the OUTFIT MNSQ for each of these items improved, indicating that the ratings became more in line with expected responses.

Structure of the FM Milestones

An item-person map (FIGURE 1) places items and people with common residency progression estimates at the same point on the continuum. Typically, items spread along the distribution of the people in order to articulate the range of the construct and to accurately measure the entire continuum. However, the 22 FM

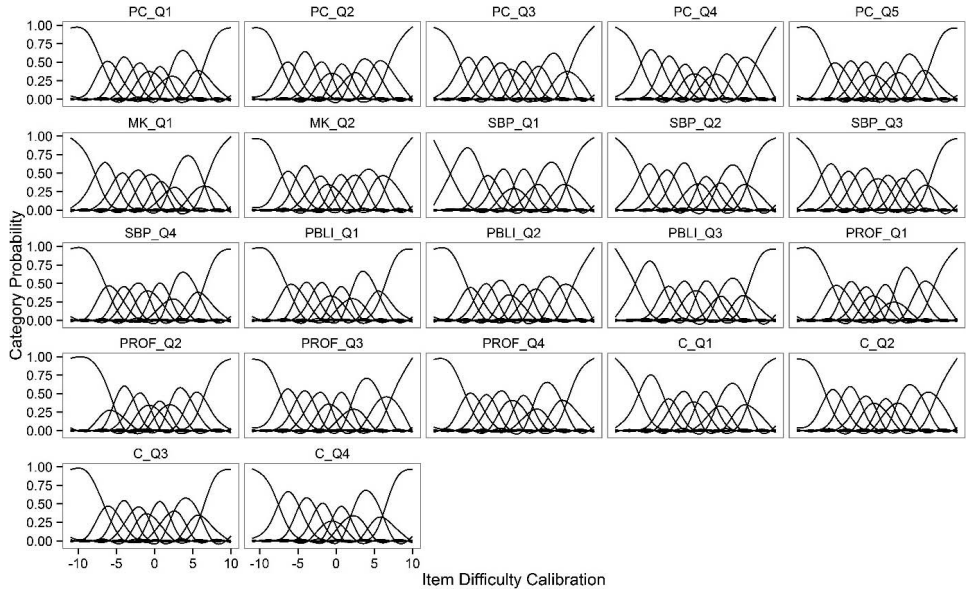

FIGURE 3
Item Characteristic Curves for Family Medicine Milestones Using Rasch Partial Credit Model on Original Rating Scale Categories

Milestone items were designed to represent different aspects of progression through residency into practice, such that the items are of similar difficulty and nearly all of the variation in difficulty is driven into the rating scale categories. This function of the FM Milestones can be seen in the Keyform (FIGURE 2), which shows the relationship between the expected response categories for each item. By drawing a vertical line from the item-response category in question through the other categories, one can see that the rating a resident receives for any item can be expected to be the same for all other items. For example, a resident who received a rating of 2 in PBLI_Q3 would be expected to receive a 2 in PBLI_Q1, a 2 in SBP_Q2, and so on down the list. This suggests that residents are not being rated on each item individually, but rather on a single global trait.

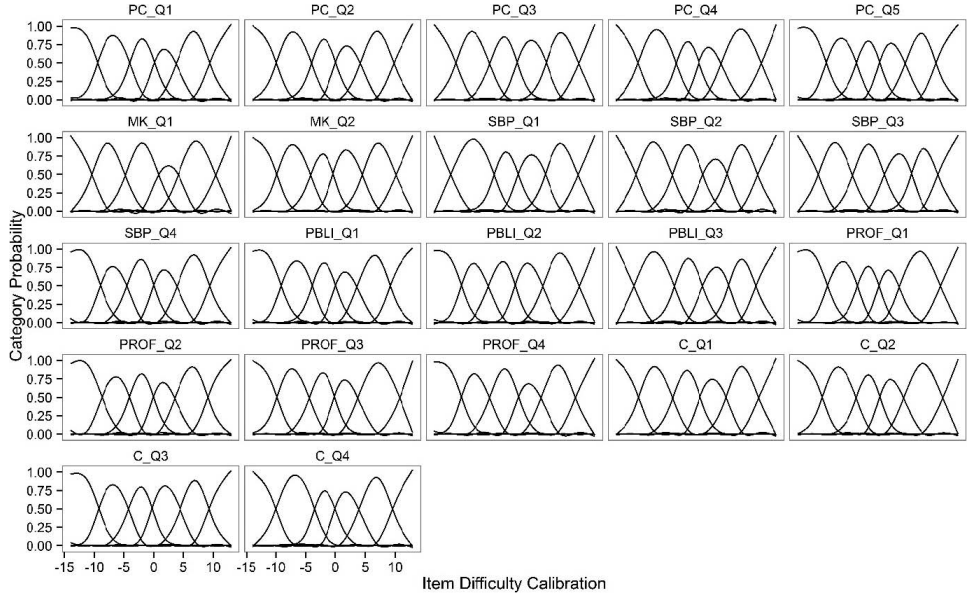

FIGURE 4
Item Characteristic Curves for Family Medicine Milestones Using Rasch Partial Credit Model on Collapsed Rating Scale Categories

Dependencies

Even after reducing the number of rating scale categories, reliability remained absurdly high at 0.98. This is likely due to internal dependencies built into the FM Milestones. For example, to achieve Level 4 on medical knowledge question 1 (MK_Q1), a resident needs to successfully complete the American Board of Family Medicine requirements for certification, and this certification is only open to PGY-3s; thus a PGY-2 can never receive this score. Dependencies like these yield a high level of reproducibility (reliability) in the data because the answers to the questions are driven by a single deterministic process and are really collecting the same piece of information by asking the same question in cosmetically different ways. Since the FM Milestones were designed as a framework to inform and guide curriculum development,[12] these dependencies are not a flaw in, but rather a feature of, their design. However, using the FM Milestone scores as a representation of knowledge or ability in any subsequent analysis would prove problematic, since the variation in scores seems to occur due to progression in residency rather than other characteristics of the resident or residency.

These dependencies, and the lack of stochasticity they cause, make any use of the FM Milestone scores as measurement in the strict sense problematic, but these scores can be useful for identifying residents who deviate from the expected progression. The FM Milestones have an average standard deviation of 1.3, so a PGY-1 would typically receive a rating of 2, 3, or 4 on most items. A PGY-2 would largely receive a 4, 5, or 6, and a PGY-3 would receive a 6, 7, or 8. In this sense, a PGY-3 who received a 4 on any item would probably be in need of remediation. Some have noted that program directors and members of the Clinical Competency Committees often have little direct observation of residents on which to base their ratings.[13,14] The exceptionally high reliability may support the claim that residents are being rated solely on their year of residency.

Sklar[15] commented that residents may be rated a little above or below their training year, but his statement had the subtext that they were largely rated by their year in residency, and our findings are largely congruent. In examining the EM Milestones, Beeson et al[3] claim that their analysis "demonstrates a practice of rating residents across a broad range of the scale, independent of the year of training." We interpret their results somewhat differently, that for nearly all residents, mean milestone scores are indeed equivalent to their training year. A visual inspection of the EM Milestones shows that the questions are written with dependencies similar to those in the FM Milestones, so we have little reason to believe the EM Milestones should function substantially differently.

Our study is subject to limitations. First, we used the first set of national FM Milestone data, and as residencies get more acquainted with creating ratings over time, scoring patterns may change. Second, each rating is determined by the institution's Clinical Competency Committee and a deeper understanding of the variation in ratings could be gained by having all scores factor in to the final score. However, these data are not available.

Conclusion

In a national study of all family medicine residents in ACGME-accredited programs, we found that lack of spread in item difficulty and lack of variation in category probabilities form the basis of a framework to inform progression through residency; however, the FM Milestones in their current form are not suitable for measuring residents or programs due to the lack of independence in the ratings. If year of residency is indeed the primary factor in assigning ratings, then the utility of the FM Milestones seems to be that of an educational framework to identify residents for remediation.

References

1. Nasca TJ, Philibert I, Brigham T, et al. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366(11):1051–1056.
2. Swing SR, Beeson MS, Carraccio C, et al. Educational milestone development in the first 7 specialties to enter the next accreditation system. *J Grad Med Educ*. 2013;5(1):98–106.
3. Beeson MS, Holmboe ES, Korte RC, et al. Initial validity analysis of the emergency medicine milestones. *Acad Emerg Med*. 2015;22(7):838–844.
4. Accreditation Council for Graduate Medical Education, American Board of Family Medicine. The Family Medicine Milestone Project. https://www.acgme.org/Portals/0/PDFs/Milestones/FamilyMedicineMilestones.pdf. Accessed November 7, 2016.
5. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Educational Research; 1960.
6. Masters GN. A Rasch model for partial credit scoring. *Psychometrika*. 1982;47(2):149–174.
7. Wright BD, Masters GN. *Rating Scale Analysis*. Chicago, IL: MESA Press; 1982.
8. Linacre JM. Optimizing rating scale category effectiveness. *J Appl Meas*. 2002;3(1):85–106.
9. Wright BD, Linacre JM. Reasonable mean square fit values. *Rasch Meas Trans*. 1994;8(3):370.

10. Smith EV. Evidence for the reliability of measures and validity of measure interpretation: a Rasch measurement perspective. In: Smith EV, Smith RM, eds. *Introduction to Rasch Measurement*. Maple Grove, MN: JAM Press; 2004:93–122.
11. Andrich D. Measurement criteria for choosing among models for graded responses. In: von Eye A, Clogg CC, eds. *Analysis of Categorical Variables in Developmental Research*. Orlando, FL: Academic Press; 1996:3–35.
12. Holmboe ES. Realizing the promise of competency-based medical education. *Acad Med*. 2015;90(4):411–413.
13. Chisholm CD, Whenmouth LF, Daly EA, et al. An evaluation of emergency medicine resident interaction time with faculty in different teaching venues. *Acad Emerg Med*. 2004;11(2):149–155.
14. Williams RG, Dunnington GL, Mellinger JD, et al. Placing constraints on the use of the ACGME milestones: a commentary on the limitations of global performance ratings. *Acad Med*. 2015;90(4):404–407.
15. Sklar DP. Competencies, milestones, and entrustable professional activities: what they are, what they could be. *Acad Med*. 2015;90(4):395–397.

All authors are with the American Board of Family Medicine. **Michael R. Peabody, PhD,** is Psychometrician; **Thomas R. O'Neill, PhD,** is Vice President, Psychometric Services; and **Lars E. Peterson, MD, PhD,** is Research Director, and Associate Professor, Department of Family and Community Medicine, University of Kentucky College of Medicine.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

A portion of this study was presented at the Accreditation Council for Graduate Medical Education Annual Educational Conference, National Harbor, Maryland, February 28, 2016.

Corresponding author: Michael R. Peabody, PhD, American Board of Family Medicine, 1648 McGrathiana Parkway, Suite 550, Lexington, KY 40511, 859.269.5626, ext 1226, fax 859.335.7501, mpeabody@theabfm.org

Received March 16, 2016; revision received July 22, 2016; accepted September 14, 2016.