Are Milestones Really Measuring Development?

Lars E. Peterson, MD, PhD
Wade Rankin, DO

The primary outcome of graduate medical education is to produce a physician with the requisite knowledge, skills, and professional behaviors for unsupervised practice. While residents' progress would be assessed, traditionally the program director's final overall assessment of training often was the stamp of approval at the time of graduation.

As part of its effort to advance competency-based medical education, the Accreditation Council for Graduate Medical Education (ACGME) began to implement the Next Accreditation System in 2013. A key element of this system entails measuring and reporting educational outcomes through the educational milestones.[1] The milestones are built on the 6 competencies of patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice, and by defining specialty-specific narrative descriptions of the trajectory of professional development within each specialty. A goal of the milestone framework is to create more specific, actionable, and developmentally based assessments of residents.

In this issue of the *Journal of Graduate Medical Education*, Barlow et al[2] detail how differences in measures of central tendency of milestone ratings at their family medicine residency produce diverse assessments of resident competence. They used milestone assessment data from a university-based family medicine residency for 2014–2015 and 2015–2016. Their Clinical Competency Committee (CCC) created new assessment instruments based on the milestone framework, and faculty entered rotation evaluations into an institutional data system. Over 2 years, this produced 841 evaluation forms with 6417 unique ratings for 22 subcompetencies. Residents could be rated at discrete intervals from 0 (not applicable) to 5 (mastery of more complex milestones), with half-point ratings between categories. After removing the 0 ratings, the authors compared the mean versus mode ratings for each subcompetency by resident for each postgraduate year (PGY). When data were bimodal, the lower mode score was used, and when multimodal, the rating was excluded. They defined a difference greater than 0.5 as meaningful and termed it an *estimation error*. The authors found 175 (22% of all evaluations) estimation errors with over half of the errors occurring on PGY-1s' ratings, with the frequency of errors decreasing for PGY-2s and PGY-3s.

For each of the 22 subcompetencies, the study found a large variation in the prevalence of estimation errors, with 3 intriguing findings. First, the less frequently a subcompetency was assessed, the higher the estimation error rate. Second, subcompetencies with wording that was more easily recognized as ordinal had fewer estimation errors. Third, use of the mean (versus the modal value) overestimated the ability of PGY-1 and PGY-2 residents, yet it underestimated that of PGY-3 residents. The authors stress that their findings offer support that the milestones are nominal, not ordinal data, and that modes are a more appropriate measure of central tendency to analyze ratings of a resident within a subcompetency, compared to the mean. They memorably state that "computing a mean subcompetency level is equivalent to computing a resident's mean eye color."[2]

We respectfully disagree with the authors' assertion that milestones are *nominal* data. The ACGME interpretation guidelines quoted by the authors clearly state that the milestones are intended to be *ordinal* data.[3] ACGME leaders voice this same interpretation when they describe the milestones as "developmentally based, specialty-specific achievements that residents are expected to demonstrate at established intervals as they progress through training"; argue that the milestones "create a logical trajectory of professional development in essential elements of competency"[1]; and suggest that with further refinement the milestones will meaningfully connect trainee programs from undergraduate to graduate medical education, and forward into practice and maintenance of certification.[4] Prior research supports a developmental framework for the milestones ranging from "beginning resident" to an "aspirational" level, with level 4 denoted as the level at which trainees are "ready for unsupervised practice."[3,5–7] If we were to believe Barlow et al[2] that the milestones were truly nominal and analogous to eye color, then the "level" a resident should achieve

DOI: http://dx.doi.org/10.4300/JGME-D-17-00203.1

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-24 via free access

TABLE
Examples of Milestone Assessment With Different Scales

Ideal Construction of the Milestones as Continuous Categories										
	1		2		3		4		5	
Milestones as **ordinal** data										
	1	2		3	4				5	
	1		2	3				4	5	
Milestones as **nominal** data										
	2	3	5	1	4					
	5	1	2	4	3					
Milestones as **poorly worded ordinal** data that are used as nominal										
	1		3	2		4			5	
	1	2		4	3				5	

Note: In the above examples, spacing and order are both important to demonstrate the concepts. In the ideal case of continuous categories, the distance is constant between each of the levels. In ordinal data, the levels are ordered but the distance between 2 levels can vary.

to be ready for unsupervised practice would be arbitrary.

While Barlow et al[2] do a good job explaining nominal, continuous, discrete, and ordinal data in the article, some readers may need more details to discern the finer points. Thus, before we further explore the issues raised by the authors, a short refresher on scales and data is in order. The TABLE graphically displays the differences between continuous, ordinal, and nominal data. The authors are correct: the milestones are not on a continuous scale. To be continuous, the distance between each point is the same at any point on the scale. To meet this criterion each milestone subcompetency would have to be precisely and evenly measured and scaled with the "distance" of 1 unit being the same at any point. We agree with the ACGME that milestone data are ordinal: the distance between 2 points on the scale is not the same, but the levels are "ranked" such that level 3 is always above level 2, level 4 above level 3, etc. However, the distance between the levels is not consistent. This contrasts with true nominal data where order does not exist and a resident could "progress" through residency from levels 5 to 2 to 1 without raising alarm.

We believe that the evidence supports the Family Medicine Milestones as ordinal data, but with often poorly described anchors between points causing some subcompetencies to be functionally nominal.

Others have noted problems with specific milestones that lack a clear set of advancing and dependent observable characteristics, making it challenging to determine progression or provide actionable feedback to trainees.[7] A recent psychometric analysis of all Family Medicine Milestones from the 2014–2015 academic year supports the idea that some of the descriptors are poor and may cause confusion with scoring.[8] Specifically, the half-point rating categories were found to provide little additional information, and the authors advocated either eliminating them or providing richer descriptors for the categories to allow better discrimination between the levels.[8] The ACGME is currently working with stakeholders both to improve the milestones and to address issues of clarity and poor wording, a process referred to as "Milestones 2.0."[3]

Many of the assumptions that the authors make, and their supporting documents, endorse the idea that the milestones are ordinal data. Their decision to take the "lower" of 2 values when a distribution was bimodal assumes one level is "above" another. Thus, one level demonstrates more competency than another, not that one has blue eyes and another brown eyes, in which neither is inherently "better" than another. Also, their choice of using a 0.5 "unit" disagreement between mean and mode assumes that this distance is meaningful and somewhat consistent on the underlying scale. In this case, removing the 0 ratings can only make sense if the evaluation forms clearly state this is "not applicable" or "not able to be evaluated," rather than the original milestone intention of "has not achieved."

While Barlow et al[2] describe the issues of assigning residents' milestone ratings from multiple observations, others have investigated the validity of the milestones from a large sample of residencies in a specialty, adding empirical evidence that milestones are ordinal data. A study of the Internal Medicine Milestone ratings from 2013–2014 found little variation by resident class, but did find increasing scores by year of residency.[5] A related study found that compared to the American Board of Internal Medicine residency annual evaluation summary, the Internal Medicine Milestones provided more diverse

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-24 via free access

ratings for junior residents.[6] An analysis of all Family Medicine Milestone data from 2014–2015 also found little variation within each resident class and little variation by subcompetency within each competency.[8] This suggests that milestones are tracking the normal expected progression through training. Taken together, these studies support the notion that a major function of the milestones is to flag problems and delays in expected development, and to identify residents who deviate from the normal developmental path and deserve extra attention or remediation. In this way, the milestones are analogous to a child's development and acquisition of motor and social skills. The analogy could be taken a step further with a population health perspective that a high frequency of deviation in a residency suggests the need for curricular improvements and residency changes. These actions are only possible if the milestone levels are ordered; otherwise, there would be no cause for concern if the resident was moving from brown eyes to blue eyes.

The decision by Barlow et al[2] not to consider another measure of central tendency that is preferred for ordinal data, the median, surprised us. The median is defined as the 50th percentile of observations, with half of the observations below and half above. Combined with the 25th and 75th percentiles (the interquartile range), the median provides an assessment of the most likely range of resident ability in a subcompetency.

The data suggest that refinements are needed for the milestones to make them more explicitly ordinal and progressive within each subcompetency. The ACGME has already begun this work, which will result in Milestones 2.0. A majority of CCCs and researchers view the milestone levels as a developmental trajectory, rather than unconnected levels with no inherent order, like eye color. However, poor wording of some milestones may cause confusion. As our measurement and assessment tools improve, our ability to track our learners' progress will improve as well.

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. Barlow PB, DuChene Thoma K, Ferguson KJ. The impact of using mean versus mode when assessing resident competency. *J Grad Med Educ*. 2017;9(3):302–309.
3. Accreditation Council for Graduate Medical Education. Milestones Annual Report 2016. www.acgme.org/Portals/0/PDFs/Milestones/MilestonesAnnualReport2016.pdf. Accessed March 16, 2017.
4. Allen S. Development of the family medicine milestones. *J Grad Med Educ*. 2014;6(1 suppl 1):71–73.
5. Hauer KE, Clauser J, Lipner RS, et al. The internal medicine reporting milestones: cross-sectional description of initial implementation in US residency programs. *Ann Intern Med*. 2016;165(5):356–362.
6. Hauer KE, Vandergrift J, Hess B, et al. Correlations between ratings on the resident annual evaluation summary and the internal medicine milestones and association with ABIM certification examination scores among US internal medicine residents, 2013–2014. *JAMA*. 2016;316(21):2253–2262.
7. Philibert I, Brigham T, Edgar L, et al. Organization of the educational milestones for use in the assessment of educational outcomes. *J Grad Med Educ*. 2014;6(1):177–182.
8. Peabody M, O'Neill T, Peterson L. Examining the functioning and reliability of the family medicine milestones. *J Grad Med Educ*. 2017;9(1):46–53.

Lars E. Peterson, MD, PhD, is Research Director, American Board of Family Medicine, and Associate Professor, Department of Family and Community Medicine, University of Kentucky; and **Wade Rankin, DO,** is Assistant Professor, Department of Family and Community Medicine, University of Kentucky.

Conflict of interest: Dr Peterson is employed by the American Board of Family Medicine (ABFM). The opinions expressed in this article are his own, and do not necessarily reflect those of the ABFM.

Corresponding author: Lars E. Peterson, MD, PhD, American Board of Family Medicine, 1648 McGrathiana Parkway, Suite 550, Lexington, KY 40511, 859.269.5626, lpeterson@theabfm.org

Downloaded from https://prime-pdf-watermark.prime-prod.pubfactory.com/ at 2025-10-24 via free access