Proceed With Caution: Implementing Competency-Based Graduate Medical Education

M. Douglas Jones Jr, MD
Tai M. Lockspeiser, MD, MHPE

Competency-based medical education offers the possibility of specifying, in advance, the outcomes that medical education is to achieve. Starting with the end in mind directs curriculum and assessment to ensure that goals will be accomplished.[1] Although the concept is attractive, implementation has been challenging. Challenges include the need to create a shared understanding of what is meant by *competence*[2] and generating and accumulating accurate assessment data to support competency decisions. Both depend on a language to express accomplishment, ideally a language shared by learners, assessors, educational leaders, accrediting and certifying institutions, and, ultimately, the public.

Rating scales provide 1 way to create a shared understanding of assessments and decisions regarding competency.[3] At first glance, scales that describe degrees of accomplishment simply, with numbers from 1 to 5 or 10, or with words ranging from *unsatisfactory* to *excellent*, would seem to satisfy the criterion. However, it soon becomes apparent that definitions of numbers and words vary from assessor to assessor and are not necessarily shared by the individual being assessed.[4] Numbers and words beg for scales tied to richer descriptions of performance that might be less arguable. The search for better rating scales linked to descriptions of behaviors is not unique to medical education. It mirrors similar, decades-long conversations about how to improve rating scales for performance assessments in employment settings.[5] Human resources managers have long used so-called behaviorally anchored rating scales (BARS) that employ short narratives to describe accomplishment of important aspects of the job category under consideration.[5,6]

Crossley et al[4] compared 5-point scales rating performance according to "expectations," from well below to well above, with scales anchored by short narratives describing levels of clinical sophistication or short narratives describing the need for clinical supervision. The authors found that the latter "construct aligned" scales reduced assessor disagreement, improved discrimination between high and low performers, and, importantly, reduced the number of assessments required for reliable conclusions. They concluded that aligning points on assessment scales with the construct to be assessed likely would reduce different interpretations by multiple assessors. These results support the notion that criterion-referenced scales will outperform norm-referenced scales. The authors also cautioned that, although narrative anchors were likely to improve performance of any instrument, it is important to collect evidence of validity for any new instrument and/or new use of an existing instrument.

In this issue of the *Journal of Graduate Medical Education*, Reddy and colleagues[7] followed that advice in collecting validity data for use with the Accreditation Council for Graduate Medical Education (ACGME) Milestone Projects as narrative anchors for assessment of chart-stimulated recall video scenarios. They found that the 5-point Milestones-Based Rating Scale outperformed a scale rating performance on a "standard" 5-point scale ranging from *critically deficient* to *aspirational*. Nevertheless, although interrater reliability improved, it remained moderate at best.

What do these results tell us? First, they illustrate the importance of accruing validity evidence for each new instrument and new use of an existing instrument, which means collecting that evidence for each new group of assessors and learners.[4,8–10] In doing so, the findings provide a salient example of another challenge to implementation of competency-based medical education—the administrative burden of a competency-based framework.[11,12] Ideally, each element of an assessment system would have sufficient validity evidence to support defensible assessment decisions for that individual element and for the program as a whole.[13] This also applies to milestones and to entrustable professional activities.[14,15]

Second, although one would hope that criterion-referenced, narrative-based rating scales would facilitate what Reddy et al[7] refer to as *shared mental models* among raters, as seen in the optimistic findings reported by Crossley et al,[4] improvement in interrater reliability was modest in the current study. In that regard, the results of Reddy and

DOI: http://dx.doi.org/10.4300/JGME-D-18-00311.1

colleagues[7] are similar to the experience in employment settings. Although the BARS approach to performance assessments is widely used, improvement in interrater reliability has not been impressive.[5] In addition, BARS have been critiqued for not including important aspects of job performance, a validity threat termed *content underrepresentation*,[8,16] and for being demanding of manager time and attention.[6]

Third, the current study findings remind us that although milestones may be better than minimally anchored rating scales, they must be used with caution. Milestones represent an enormous amount of careful thought by experts,[10] but expert opinion is still just opinion. For summative assessment, milestones remain hypotheses.[10]

Fourth, a milestone-based scale for workplace-based assessment shares the challenges of any rating scale. Accumulating evidence suggests that no rating scale is likely to eliminate assessor disagreement.[17–20] One reason may be that assessors are not all evaluating the same aspect of performance.[20] Different assessors may be focusing on different aspects at different times with different learners. Moreover, the idea that the primary focus of rating scales should be perfect agreement among assessors diverts attention from the important role of subjectivity and individual expert opinion in assessments.[17,18,20–22] Clinical medicine involves working with diverse individuals in a multitude of contexts. It is unlikely that there will be 1 "right" answer as to how a learner is performing. Assessment is more about gathering data in different contexts and understanding differences as well as similarities[22] and about using both appropriately to guide teaching and learning.

Flexner[23] famously wrote, "Though medicine can be learned, it cannot be taught." Taken as such, out of context, the sentence overstates. The craft of medicine—how to take a history and perform a physical examination or how to apply basic principles of systems physiology to health and disease—can and must be taught. However, Flexner[23] was correct in noting that responsibility for learning to put items of the craft together to care for a patient falls almost entirely to the learner. Understanding the utility of milestones as instruments of assessment is a long, complex, and unfinished work in progress.[10] Regardless of how useful milestones turn out to be for summative assessment, they provide a consensus blueprint for learning and improvement, in other words, for formative assessment.[10,11] If, as suggested by Norman et al,[11] milestones do no more than encourage teachers to pay closer attention to learners and to what they are teaching and coaching and provide learners with a clearer path toward clinical competence, they will have made invaluable contributions[10] to teaching and learning in medicine.

References

1. Touchie C, ten Cate O. The promise, perils, problems and progress of competency-based medical education. *Med Educ*. 2016;50(1):93–100.
2. Hodges BD. The shifting discourses of competence. In: Hodges BD, Lingard L, eds. *The Question of Competence: Reconsidering Medical Education in the Twenty-First Century*. Ithaca, NY: ILR Press; 2012:14–41.
3. Schoenherr JR, Hamstra SJ. Psychometrics and its discontents: an historical perspective on the discourse of the measurement tradition. *Adv Health Sci Educ Theory Pract*. 2016;21(3):719–729.
4. Crossley J, Johnson G, Booth J, et al. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Med Educ*. 2011;45(6):560–569.
5. Schwab DP, Heneman HG III, DeCotiis TA. Behaviorally anchored rating scales—a review of the literature. *Per Psychol*. 1975;28(4):549–562.
6. Lloyd K. Behind BARS: evaluating employees with behaviorally anchored rating scales. http://www.dummies.com/business/human-resources/employee-relations/behind-bars-evaluating-employees-with-behaviorally-anchored-rating-scales. Accessed April 20, 2018.
7. Reddy ST, Tekian A, Durning SJ, et al. Preliminary validity evidence for a milestones-based rating scale for chart-stimulated recall. *J Grad Med Educ*. 2018;10(3):269–275.
8. Messick S. The interplay of evidence and consequences in the validation of performance assessments. *Educ Res*. 1994;23(2):13–23.
9. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e167–e116.
10. Holmboe ES, Yamazaki K, Edgar L, et al. Reflections on the first 2 years of milestone implementation. *J Grad Med Educ*. 2015;7(3):506–511.
11. Norman G, Norcini J, Bordage G. Competency-based education: milestones or millstones? *J Grad Med Educ*. 2014;6(1):1–6.
12. Klamen DL, Williams RG, Roberts N, et al. Competencies, milestones, and EPAs—are those who ignore the past condemned to repeat it? *Med Teach*. 2016;38(9):904–910.
13. Hauer KE, O'Sullivan PS, Fitzhenry K, et al. Translating theory into practice: implementing a program of assessment. *Acad Med*. 2018;93(3):444–450.

14. Parker TA, Guiton G, Jones MD Jr. Choosing entrustable professional activities for neonatology: a Delphi study. *J Perinatol*. 2017;37(12):1335–1340.
15. ten Cate O, Hart D, Ankel F, et al; International Competency-Based Education Collaborators. Entrustment decision making in clinical training. *Acad Med*. 2016;91(2):191–198.
16. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e7–e16.
17. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564–568.
18. Gingerich A, Kogan J, Yeates P, et al. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–1068.
19. Govaerts M. Workplace-based assessment and assessment for learning: threats to validity. *J Grad Med Educ*. 2015;7(2):265–267.
20. Gingerich A, Ramlo SE, van der Vleuten CPM, et al. Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Adv Health Sci Educ Theory Pract*. 2017;22(4):819–838.
21. Govaerts MJ. Competence in assessment: beyond cognition. *Med Educ*. 2016;50(5):502–504.
22. Govaerts M, van der Vleuten CP. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47(12):1164–1174.
23. Flexner A. *Medical Education: A Comparative Study*. New York, NY: Macmillan Company; 1925.

Both authors are with the Department of Pediatrics, University of Colorado School of Medicine. **M. Douglas Jones Jr, MD,** is a Professor; and **Tai M. Lockspeiser, MD, MHPE,** is an Associate Professor.

Corresponding author: M. Douglas Jones Jr, MD, Department of Pediatrics, University of Colorado School of Medicine, C290 Building 500, Room E1305, 13001 E 17th Place, Aurora, CO 80045, 303.829.9621, doug.jones@ucdenver.edu