

The Promise of Milestones: Are They Living Up to Our Expectations?

Su-Ting T. Li, MD, MPH

Beginning in 2013, the Accreditation Council for Graduate Medical Education (ACGME) required program directors to semiannually report milestone data for their trainees. The milestones are competency-based observable behaviors that mark a trainee's developmental progression toward unsupervised practice by enhancing learner assessment and feedback and program evaluation and improvement.¹ The aim is to help "residencies and fellowships produce highly competent physicians to meet the health and health care needs of the public."²

This raises the question: Are the milestones meeting that goal? Specifically, what is the validity evidence that supports the use of milestones to "produce highly competent physicians"? Validity evidence includes content validity, internal structure, response process, relationship to other variables, and consequences.³ This commentary highlights some of the validity evidence supporting the use of milestones, and outlines areas where further research is needed.

Content Validity: Do the Milestones Encompass All of the Attitudes, Knowledge, Skills, and Behaviors Needed to Be a Competent Physician in a Given Specialty?

Milestones were designed to provide strong evidence of content validity. In concert with the ACGME and the relevant American Board of Medical Specialties (ABMS) specialty board, each specialty convened a Milestone Working Group to develop specialty-specific milestones.⁴ Milestones were developed through working groups' expert consensus, with extensive feedback from stakeholders and subsequent revisions.⁴ Working groups used the 6 ACGME competencies and the Dreyfus model of skill acquisition as theoretical frameworks.⁴⁻⁶ In addition, many used literature reviews to inform the development of milestone sets.⁷⁻¹⁰ However, many subspecialty fellowships did not develop their own milestone sets, and instead share milestone sets. For example, all internal medicine subspecialties share the same milestone sets, as do all pediatrics subspecialties. It is possible that what defines a

competent cardiologist is different than what defines a competent endocrinologist. In addition, many specialties developed more milestone sets than the ACGME requires to be reported. For example, pediatrics developed 51 milestone sets, of which the ACGME chose 21.^{7,11} While judiciously limiting the number of reportable milestones to key measurable outcomes that define a competent physician in a given specialty is important to making assessment and reporting feasible, it will be equally important to ensure that the limited milestone sets encompass all critical aspects of a competent physician in the specialty.

Internal Structure: Are the Milestones Measuring What It Means to Be a Competent Physician?

A key aim of the milestones is to ensure competent physicians. Studies of the psychometric properties of the milestone sets (internal structure) may help specialties pare down their more comprehensive list of milestone sets. In this issue of the *Journal of Graduate Medical Education*, Peabody and colleagues¹² examine the psychometric properties of the Family Medicine (FM) Milestones. They argue that the FM Milestone scores are similar for different subcompetencies, for the same resident, and that all items describe a single construct of a competent FM physician. Therefore, it could be argued that not all 22 FM Milestone sets are needed to evaluate whether a resident is a competent FM physician.

In contrast to the finding of a single FM competence construct, emergency medicine identified 3 constructs,¹³ and internal medicine and obstetrics-gynecology identified 6 constructs aligning with the 6 ACGME competencies.¹⁴ Several specialties found that milestone ratings differed by subcompetency,¹⁵⁻¹⁷ with pediatrics and internal medicine finding that resident professionalism and interpersonal and communication skills were rated highest.^{15,16}

Response Process: Are Milestone Scores Reliable?

In order to trust the validity of milestones for making high-stakes decisions for trainees or programs, it is

DOI: <http://dx.doi.org/10.4300/JGME-D-16-00694.1>

important to ensure that scores are reliable, both within and across programs. If milestones are intended to allow for a shared mental model, would Clinical Competency Committee (CCC) members agree on a resident's milestone score? If a resident transferred programs, would he or she receive the same milestone scores? Before the ABMS utilizes milestone scores to compare residents across programs or the ACGME uses aggregate milestone data to compare programs, it is important to ensure that programs rate trainee performance consistently within their own program and across programs. Faculty development can reduce rater variability of milestone ratings,¹⁸ and standard-setting videos may be 1 tool to ensure consistent milestone ratings both within and across residency programs.¹⁹ Recently, the ACGME began releasing end-of-residency milestone scores to fellowship directors for matriculating fellows.²⁰ To date, no studies have demonstrated similarity of milestone ratings among programs of the same specialty. Without evidence of reliability, there may be unintended consequences to fellowship directors' interpretation of the milestone scores their new fellows received through this educational handoff.

Relationship to Other Variables: How Do Results From Milestone Scores Relate to Other Assessments of the Learner?

Ultimately, there should be evidence that graduates with higher milestone scores are better physicians, or, alternatively, graduates with low milestone scores are more likely to have patients who experience complications, be sued, and lose their medical license (predictive validity). We would like evidence that residents receive higher milestone scores as they progress through training and that faculty regarded as "experts" in a subcompetency area receive higher milestone ratings than an intern (concurrent validity). Based on their findings of limited variability in milestone scores for residents in the same training year, Peabody and colleagues¹² contend that FM Milestones do not measure the amount of inherent ability possessed by a resident, but instead identify where residents are in their progression through residency, and identify residents with lower milestone scores than peers for possible remediation. This study adds to the growing body of literature that provides concurrent validity evidence that residents with higher levels of training have higher milestone scores,^{12,14-16,21,22} and lower milestone scores within a postgraduate year level may identify struggling learners.¹⁴

Consequences: What Is the Impact of the Interpretation of Milestone Scores?

From the interpretation of milestone scores, and decisions based on these scores, what is the potential impact on trainees, residency programs, and society? At the individual resident level, milestones offer the opportunity for formative feedback and summative assessment to help program directors make advancement and remediation decisions. Theoretically, milestones allow learners and educators to have a shared mental model of expectations of a competent physician in that specialty, and a roadmap to get there. This should improve feedback given to trainees.²³

In a study of internal medicine residents, half found milestone-based feedback helped identify their strengths, weaknesses, specific areas for improvement, and educational progress, and felt that milestone-based feedback was more helpful than previous forms of feedback.²⁴ Specialty-specific milestones could help medical students plan their final year's medical school curriculum to prepare them for entering residency.²⁵ Similarly, fellowship-specific milestones could help residents shape their elective experience to prepare them for entering fellowship. More research needs to be done on how to make the milestones more useful to the learner.

Using milestone scores for higher-stakes decisions, such as graduation, eligibility for board certification examination, or licensure, would require the determination of a threshold milestone score. Trainees who receive ratings above the threshold milestone score would be deemed satisfactory and be able to advance; trainees who receive ratings below the threshold score would be identified for remediation. We would like to know that graduates who achieve threshold milestone scores are ready to practice without supervision, and that additional progression along the path to expertise can be accomplished postgraduation without detriment to the patient.

Mapping milestones to entrustable professional activities (EPAs) may allow us to simultaneously establish a milestone threshold that corresponds to a given EPA threshold (entrustment to perform an activity without supervision), and decrease the assessment burden by allowing assessment of multiple milestone sets at a time in a way that may be more understandable to both evaluators and trainees.²⁶⁻²⁸ EPAs could be mapped to milestones and, along with research, determine whether they were mapped correctly. Each rotation could then assess a limited number of EPAs along supervisory lines, as suggested by Rekman and colleagues' Ottawa Clinic Assessment Tool.²⁹ Evaluators then could determine if the trainee

was trusted to observe only (“I had to do it”); trusted to perform with direct observation (“I had to talk them through”); trusted to perform with indirect observation and key findings repeated (“I had to direct them from time to time”); trusted to perform with indirect observation (“I needed to be available just in case”); trusted to perform independently with no supervision (“I did not need to be there”); or trusted to supervise others.^{29,30} EPA descriptors for each level of supervision could be described to standardize entrustment decisions.²⁷ Milestones could be helpful to drill down where trainees are struggling to facilitate appropriate remediation.

While some specialties have explicitly defined Level 4 as the target score for graduation and “ready for unsupervised practice,” in other specialties, it is not clear what milestone scores should lead to remediation.³¹ In pediatrics, only 21% of end-of-year graduating residents received a 4 or higher on all subcompetencies, with most receiving a 3 or higher on all subcompetencies.¹⁵ In 2015, Pediatrics Milestones were revised to establish milestone Level 3 as the graduation target.¹¹ Should the milestone threshold score be the same for all subcompetencies in a given specialty? In addition, it is unclear whether a threshold score needs to be established for all subcompetencies. The danger is that, if we set the threshold score too high, residents who would have been competent physicians may not graduate. If we set it too low, residents may graduate whose lack of competence may harm patients.

Aggregate milestone scores could help programs identify subcompetencies where their trainees perform less well compared to other trainees in the program and to national program aggregate scores. These could be areas in which the program could develop additional curricula. At the program level, before accreditation can be based on milestone scores, evidence that milestone scores are reliable between programs will be needed. This kind of research needs to assess whether the variation in milestone scores among programs is based on differences in residents’ actual performance or on how programs evaluate their learners. Alternatively, minimal variation of milestone scores among programs may indicate that residencies produce comparably competent graduates or suggest that CCCs are concerned that assigning residents a lower-than-threshold milestone score may be a red flag to the ACGME.³² Programs in the same specialty also may have different program aims and purposely train physicians to serve different population needs. A program that seeks to produce family physicians to serve rural populations may need different skill sets in its graduates than a program that produces family physicians to serve urban,

underserved populations or a program that educates the next cohort of academic family physicians. The different skill sets required may result in the need for graduates of a given program to attain a Level 5 for some subcompetencies, and a Level 3 for others.

Validity evidence for the use of milestones to assure the public that programs are producing highly competent physicians is growing. Content validity evidence is strong, and some psychometric evidence supports the internal structure of some milestones. Currently, there is validity evidence to support the use of milestones to provide formative feedback to trainees and programs. However, before milestones are used to make advancement or remediation decisions for trainees, or accreditation decisions for programs, more validity evidence is needed. Local and national faculty development is needed to ensure reliable milestone-based assessments within and across programs in a given specialty. National data are needed to determine appropriate milestone thresholds for entrustment decisions. We need more evidence to determine whether single milestone thresholds are appropriate, or whether thresholds should be tailored to individual resident and program goals. Finally, studies of the predictive ability of milestone scores to produce the next generation of competent physicians and information on the consequences of using different threshold scores to make these decisions are needed. Milestones hold the promise of being able to help produce highly competent physicians—we have our work cut out for us to prove whether they do.

References

1. Carraccio C, Iobst WF, Philibert I. Milestones: not millstones but stepping stones. *J Grad Med Educ.* 2014;6(3):589–590.
2. Holmboe ES, Edgar L, Hamstra S. The Milestones Guidebook. 2016. <http://www.acgme.org/Portals/0/MilestonesGuidebook.pdf>. Accessed November 7, 2016.
3. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830–837.
4. 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.
5. Holmboe ES, Call S, Ficalora RD. Milestones and competency-based medical education in internal medicine. *JAMA Intern Med.* 2016;176(11):1601–1602.
6. Hicks PJ, Schumacher DJ, Benson BJ, et al. The pediatrics milestones: conceptual framework, guiding principles, and approach to development. *J Grad Med Educ.* 2010;2(3):410–418.

7. Accreditation Council for Graduate Medical Education; American Board of Pediatrics. The Pediatrics Milestone Project. January 2012. <https://www.abp.org/sites/abp/files/pdf/milestones.pdf>. Accessed October 20, 2016.
8. Green ML, Aagaard EM, Caverzagie KJ, et al. Charting the road to competence: developmental milestones for internal medicine residency training. *J Grad Med Educ*. 2009;1(1):5–20.
9. Allen S. Development of the family medicine milestones. *J Grad Med Educ*. 2014;6(1 suppl 1):71–73.
10. Cogbill TH, Swing SR. Development of the educational milestones for surgery. *J Grad Med Educ*. 2014;6(1 suppl 1):317–319.
11. Accreditation Council for Graduate Medical Education; American Board of Pediatrics. The Pediatrics Milestone Project. July 2015. <http://www.acgme.org/Portals/0/PDFs/Milestones/PediatricsMilestones.pdf>. Accessed November 7, 2016.
12. Peabody MR, O'Neill TR, Peterson LE. Examining the functioning and reliability of the Family Medicine Milestones. *J Grad Med Educ*. 2017;9(1):46–53.
13. Beeson MS, Holmboe ES, Korte RC, et al. Initial validity analysis of the emergency medicine milestones. *Acad Emerg Med*. 2015;22(7):838–844.
14. Park YS, Zar FA, Norcini JJ, et al. Competency evaluations in the next accreditation system: contributing to guidelines and implications. *Teach Learn Med*. 2016;28(2):135–145.
15. Li ST, Tancredi DJ, Schwartz A, et al. Competent for unsupervised practice: use of pediatric residency training milestones to assess readiness. *Acad Med*. 2016 Jul 26. Epub ahead of print.
16. Warm EJ, Held JD, Hellmann M, et al. Entrusting observable practice activities and milestones over the 36 months of an internal medicine residency. *Acad Med*. 2016;91(10):1398–1405.
17. Bradley KE, Andolsek KM. A pilot study of orthopaedic resident self-assessment using a milestones' survey just prior to milestones implementation. *Int J Med Educ*. 2016;7:11–18.
18. Raj JM, Thorn PM. A faculty development program to reduce rater error on milestone-based assessments. *J Grad Med Educ*. 2014;6(4):680–685.
19. Calaman S, Hepps JH, Bismilla Z, et al. The creation of standard-setting videos to support faculty observations of learner performance and entrustment decisions. *Acad Med*. 2015;91(2):204–209.
20. Edgar L, Holmboe E. Educational Handoff Letter. 2016. <http://www.acgme.org/Portals/0/PDFs/Milestones/EducationalHandoffLetter.pdf>. Accessed November 7, 2016.
21. Ross FJ, Metro DG, Beaman ST, et al. A first look at the Accreditation Council for Graduate Medical Education anesthesiology milestones: implementation of self-evaluation in a large residency program. *J Clin Anesth*. 2016;32:17–24.
22. Goldman RH, Tuomala RE, Bengtson JM, et al. How effective are new milestones assessments at demonstrating resident growth? 1 year of data. *J Surg Educ*. 2017;74(1):68–73.
23. Schumacher DJ, Lewis KO, Burke AE, et al. The pediatrics milestones: initial evidence for their use as learning road maps for residents. *Acad Pediatr*. 2013;13(1):40–47.
24. Angus S, Moriarty J, Nardino RJ, et al. Internal medicine residents' perspectives on receiving feedback in milestone format. *J Grad Med Educ*. 2015;7(2):220–224.
25. Lamba S, Wilson B, Natal B, et al. A suggested emergency medicine boot camp curriculum for medical students based on the mapping of core entrustable professional activities to emergency medicine level 1 milestones. *Adv Med Educ Pract*. 2016;7:115–124.
26. Choe JH, Knight CL, Stiling R, et al. Shortening the miles to the milestones: connecting epa-based evaluations to acgme milestone reports for internal medicine residency programs. *Acad Med*. 2016;91(7):943–950.
27. Carraccio C, Englander R, Gilhooly J, et al. Building a framework of entrustable professional activities, supported by competencies and milestones, to bridge the educational continuum. *Acad Med*. 2016 Mar 8. Epub ahead of print.
28. Carraccio C, Englander R, Holmboe ES, et al. Driving care quality: aligning trainee assessment and supervision through practical application of entrustable professional activities, competencies, and milestones. *Acad Med*. 2016;91(2):199–203.
29. Rekman J, Hamstra SJ, Dudek N, et al. A new instrument for assessing resident competence in surgical clinic: the Ottawa Clinic Assessment Tool. *J Surg Educ*. 2016;73(4):575–582.
30. Rekman J, Gofton W, Dudek N, et al. Entrustability scales: outlining their usefulness for competency-based clinical assessment. *Acad Med*. 2016;91(2):186–190.
31. Accreditation Council for Graduate Medical Education; American Board of Internal Medicine. The Internal Medicine Milestone Project. July 2015. www.acgme.org/portals/0/pdfs/milestones/internalmedicinemilestones.pdf. Accessed October 14, 2016.
32. Witteles RM, Verghese A. Accreditation Council for Graduate Medical Education (ACGME) Milestones—time for a revolt? *JAMA Intern Med*. 2016;176(11):1599–1600.



Su-Ting T. Li, MD, MPH, is Associate Professor, Vice Chair of Education, and Residency Program Director, Department of Pediatrics, University of California, Davis.

Corresponding author: Su-Ting T. Li, MD, MPH, University of California, Davis, Room 220, 2516 Stockton Boulevard, Sacramento, CA 95817, 916.734.2428, fax 916.734.0342, sutli@ucdavis.edu