The Evolution of Assessment: Thinking Longitudinally and Developmentally

Eric S. Holmboe, MD, Kenji Yamazaki, PhD, and Stanley J. Hamstra, PhD

Abstract

Becoming a physician or other health care professional is a complex and intensely developmental process occurring over a prolonged period of time. The learning path for each medical student, resident, and fellow varies due to different individual learner abilities and curricular designs, clinical contexts, and assessments used by the training program. The slow and uneven evolution to outcomes-based medical education is partly the result of inadequate approaches to programmatic assessment that do not fully address all essential core competencies needed for practice or account for the developmental nature of training. Too many assessments in medical education still focus on single point-in-time performance or function as indirect proxies for actual performance in clinical care for patients and families.

Milestones are a modest first step of providing predictive, longitudinal data on a national scale. Longitudinal Milestones data can facilitate the continuous improvement efforts of programs in assessment. However, Milestone judgments are only as good as the assessment data and group processes that inform them. Programmatic assessment should be longitudinally focused and provide all learners with comprehensive and actionable data to guide their professional development and support creation of meaningful individualized action plans. Efforts are urgently needed to rebalance programmatic assessment away from an overreliance on assessment proxies toward more effectively using developmentally focused work-based assessments, routinely incorporate clinical performance and patient experience data, and partner with learners through iterative coproduced assessment activities.

The journey to becoming a practicing health care professional is an intensely developmental process that occurs over a substantial period of time. A number of educational models and theories recognize this longitudinal process, such as the Dreyfus and Kegan stage models of professional development, mastery-based learning, acquisition of expertise through deliberate and intentional practice, and learning curves. The iconic Miller’s pyramid is a quasi-developmental assessment model, with the knows, knows how, and shows how levels of assessment still substituting as proxies for the does level in many training program’s assessment system. In the 1990s, the medical education community began to explicitly embrace this developmental nature of training through competency-based medical education (CBME) models and greater attention to educational outcomes.

For example, the Accreditation Council for Graduate Medical Education (ACGME) and American Board of Medical Specialties jointly launched the general competency framework, informed by the work of the Dreyfus brothers, as part of the Outcome Project in 2001. Despite the introduction of the competencies in 2001, most curriculum and assessment approaches continue to rely heavily on a “building block,” or cross-sectional, educational philosophy. For example, most graduate medical education (GME) programs still use a rotation-based, block schedule model with episodic assessments unevenly “sprinkled” into this rotational experience. Regular feedback is uncommon and feed-forward even rarer to guide professional development of learners. Yet, multiple studies have clearly demonstrated variable learning and growth curves among residents and fellows are the norm. For example, the number of procedures necessary for a learner to acquire proficiency can vary widely among individual residents. These observations point to the essential need to use more work-based assessment as part of a continuous, longitudinal process to guide feedback, coaching, and self-regulated learning for successful professional development.

Regrettably, assessment still depends too heavily on “proxies” for judging learner abilities. One example is the continued reliance on multiple-choice tests as a primary “gating mechanism” between undergraduate to GME to practice. Another example is the overreliance on volume as an assessment metric, especially for procedures. Volume is a crude proxy for developmental progression, a fact laid bare by the work in mastery-based learning and learning curves. When work-based assessments do directly target actual performance, past studies of these tools often analyzed validity using time-based stages and not actual abilities. For example, the original mini-clinical evaluation exercise studies used a 9-point scale with anchors ranging from unsatisfactory to superior. Numeric differences between first-, second-, and third-year residents in internal medicine were modest and grouped at the upper end of the scale, regardless of year or stage of training. These nondevelopmental, norm-referenced tools provide limited guidance to learners’ professional development.

Recognizing the shortcomings of a cross-sectional, norm-referenced approach to assessment, medical educators began exploring more developmentally based approaches to assessment. For
Milestones are an early attempt to introduce a longitudinal, developmental, and criterion-referenced assessment rubric for residency and fellowship training. Importantly, the Milestones are designed as part of a formative national assessment system to guide learner development and progression, ideally supported by a program of assessment embedded within a residency or fellowship program. The ACGME planned from the beginning that Milestones would be grounded in continuous quality improvement principles and require ongoing revision and change. Just as biologic, clinical, and systems science evolves, so too must the Milestones also incorporate evolving pedagogical and assessment science.

Programmatic assessment is a core component of CBME. The introduction of clinical competency committees (CCCs) into all specialties along with Milestones implementation has helped to advance programmatic assessment and group judgments around learner professional development. CCCs integrate and synthesize multiple sources of assessment data, both qualitative and quantitative, into a developmental Milestone judgment twice a year over the entire course of training. As a result of these early efforts across the U.S. GME community, we now have a national, longitudinal assessment database that enables early research into learning trajectories across multiple competencies and specialties. Perhaps more importantly, this national system of assessment targets key competencies, or abilities, beyond the traditional focus of medical knowledge and clinical skills.

In 2019, the ACGME began to provide Milestones predictive probability values (PPVs) for all entry-level residencies. The PPVs are grounded in the epidemiological concepts of sensitivity, specificity, and prevalence. Using a national sample and accounting for nesting within programs, these PPV matrices provide probability estimates that residents in their specialty, at each 6-month reporting period, will attain the recommended Milestone level at the time of graduation. The PPVs were generated as population averages, accounting for nesting within programs. This approach means PPVs could be overestimations for one group of programs (i.e., lenient rating programs) and underestimations for another group of programs (i.e., stringent rating programs). The PPVs represent a modest first effort to bring the power of learning analytics and longitudinal Milestone trajectories, or learning curves, into GME programs nationally.

For programs to interpret PPVs most effectively, they must understand their own CCC rating tendencies. The explicit goals for these PPVs are to help program directors and CCCs identify areas for improvement among their learners and maximize the probability the resident will achieve proficiency across all competencies by graduation. Ideally, a program director or faculty advisor will use the PPVs to help learners identify opportunities for improvement as part of the feedback conversation that should follow the CCC deliberations. For learners with multiple PPVs of concern, a more formal remediation program may be most helpful.

To realize the full potential of PPVs and the Milestones, residents must be active agents in their own assessment, and programs must provide feedback and coaching using coproduction principles. While PPVs can be helpful by providing a national benchmark, residency programs should also create longitudinal assessment dashboards for their individual residents. Such dashboards can enhance coproduction and enable residents to use real-time assessment data for personal improvement purposes. A local, longitudinal dashboard used in conjunction with national PPV benchmarking data holds substantial promise to advance competency-based assessment.

For programs to interpret PPVs most effectively, they must understand their own CCC rating tendencies. The explicit goals for these PPVs are to help program directors and CCCs identify areas for improvement among their learners and maximize the probability the resident will achieve proficiency across all competencies by graduation. Ideally, a program director or faculty advisor will use the PPVs to help learners identify opportunities for improvement as part of the feedback conversation that should follow the CCC deliberations. For learners with multiple PPVs of concern, a more formal remediation program may be most helpful.

To realize the full potential of PPVs and the Milestones, residents must be active agents in their own assessment, and programs must provide feedback and coaching using coproduction principles. While PPVs can be helpful by providing a national benchmark, residency programs should also create longitudinal assessment dashboards for their individual residents. Such dashboards can enhance coproduction and enable residents to use real-time assessment data for personal improvement purposes. A local, longitudinal dashboard used in conjunction with national PPV benchmarking data holds substantial promise to advance competency-based assessment.

The entire GME community will also need to accelerate the adoption of programmatic assessment ensuring all competencies are effectively taught and assessed. Programmatic assessment must be aligned with the progressive sequencing of the curriculum to support the longitudinal professional development and identity formation of all residents and fellows. This requires programs of assessment to be multifaceted, progressive, and developmental. Unlike traditional cross-sectional assessments, longitudinal assessments may need to identify/ differentiate developmental patterns that may help struggling learners remediate deficiencies and help all residents strive for mastery. As such, assessments that provide feedback differentiating individual learner stages in developmental progress may be important. The implications for developmentally focused assessment start with a recognition we all need a better balance between quantitative- and qualitative-based assessment. Numeric ratings are nothing more than a “code,” much like what clinical researchers use when analyzing large datasets. Codes need to be attached to valid meaning and description, a neglected issue with most scale-based assessment. Without the narrative developmental descriptors provided at each Milestone level, the PPVs would have substantially less utility. However, these developmental descriptors are just a start—ultimate meaning depends on the assessments informing the Milestones and the effectiveness and rigor of the CCC process.

Second, there are still too many gaps in assessments for professionalism, practice-based learning and improvement, and systems-based practice. Valid assessment methods exist for interpersonal skills and communication, such as multisource feedback, patient experience surveys, and informed decision-making rubrics—we simply do not use them enough in GME programs. Third, we need to continuously improve the validity of assessment proxies for performance in clinical care as long as they remain part of programmatic assessment. However, we also need to reframe and advance our concepts of validity and reliability in the growing number of longitudinal assessment models where assessments evolve and change over time. Do the assumptions underlying validity and reliability for cross-sectional, point-in-time tests still hold for assessments repeatedly produced
and refined over 3 to 6 years? Ultimately, GME needs to continue to shift the balance of programmatic assessment that uses valid clinical performance assessment and measures more closely linked to patient outcomes. Currently, clinical performance data to guide professional development in GME remains woefully underused. For example, continuing to correlate multiple-choice examination scores with performance on clinical care measures should lead to the question, how can we better assess clinical performance directly and ensure the validity of those measures?

You may now be asking where do Milestones and EPAs fit in the proxy conversation? Both incorporate elements of assessment proxies. Milestones and EPAs are best viewed as a bridge on medical education’s journey to the next phase of more effective programmatic assessment. Milestones and EPAs must continue to evolve as programs incorporate more and more measurements of actual clinical performance with and from patients and families.

Finally, we all must put greater effort into implementation of developmental and longitudinal approaches to programmatic assessment. Implementation has become somewhat politicized within the GME community. Yet, several decades of important work in the field of implementation science has revealed key elements and processes necessary for implementing change and interventions. We would do well to learn from this science as medical education has benefitted from social and pedagogical sciences.

In conclusion, medical education must continue its journey toward a more robust developmental mindset in curricula and assessment. Milestones with the PPV matrices are a modest first step on this complex journey. Learning analytics, augmented intelligence, machine learning, and greater connection to and use of clinical performance data will also help to accelerate this transformation by bringing increasing amounts of real-time data into longitudinally and developmentally based assessments.

Funding/Support: None reported.

Other disclosures: All authors are employed by the Accreditation Council for Graduate Medical Education. Eric Holmboe receives royalties for a textbook on assessment from Elsevier.

Ethical approval: Reported as not applicable.

E.S. Holmboe is chief, Research, Milestones Development and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois.


S.J. Hamstra is vice president for outcomes research, Research, Milestones Development and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois.

References