A Validity Framework for Effective Analysis and Interpretation of Milestones Data

Stanley J. Hamstra, PhD
Kenji Yamazaki, PhD

As part of its transition to competency-based medical education, in 1999 the US graduate medical education (GME) community identified 6 core competency domains, attributes that residents and fellows must develop in order to carry out professional roles. In 2013, as part of the Next Accreditation System (NAS),[1] the Accreditation Council for Graduate Medical Education (ACGME) implemented the Milestones, brief narrative developmentally based statements that describe the skills, knowledge, and behavior for each of the 6 competency domains within a specific medical field. More than 150 specialty and subspecialty Milestone sets have been created. Each program's clinical competency committee (CCC) reviews each GME learner's progress on these Milestones every 6 months and makes recommendations to the program director for the final decision.

Summary measurements of each resident and fellow's competence are sent to the ACGME every 6 months through the Milestones rating system. As a measurement system, Milestones ratings are subject to the same questions of validity as any other. This article examines the reporting of Milestone ratings through the lens of validity theory.

Monitoring of Milestone ratings during GME should allow for enhanced feedback and training to help prepare every resident and fellow for meeting the needs of the patient population in their specialty. While there is increasing evidence that Milestones ratings accurately reflect a learner's competence,[2–4] there are also legitimate concerns about the influence of irrelevant factors on these ratings, such as rater bias, inadequate opportunity to observe performance, and confusion about the Milestones language.[5–7] Fortunately, some of these issues are now being addressed systematically as more evidence accumulates about the influence of irrelevant factors.[8–10] For example, in response to program directors' concerns about the clarity and precision of Milestones language within specialties, there is currently a large effort underway to revise the full set of Milestones—a project known as Milestones 2.0.[11]

DOI: http://dx.doi.org/10.4300/JGME-D-20-01039.1

Despite these efforts, and many individual programs' attempts to generate Milestone ratings with great care, there is still legitimate concern about the validity of decisions made using the Milestone ratings that are received and processed at the ACGME.[12,13] This article attempts to provide an overall framework, using validity theory, to guide further developments along these lines. This framework will provide useful tools for making sense of Milestones data at the national level by separating out "signal" from "noise" in the data.

This article is designed to be of broad interest to those interested in improving the quality of interpretation of Milestone ratings at a national level, including educators, researchers, and institutional GME leaders. Such insights might also provide guidance for program directors, CCC members, and faculty in understanding how to improve the quality of data within their program.

Milestone Ratings as an Assessment Process and the Resulting Validity Imperative

While the concept of validity has a substantial history in psychometric theory, most of the work in this area in medical education has focused on standardized assessment methods such as multiple-choice tests of knowledge or objective structured clinical examinations. It is relatively recent that attempts have been made to apply validity theory in a systematic way to *in situ* workplace-based assessments such as Milestone ratings. As such, it is instructive to highlight a definition of validity that was offered by one of the seminal figures in educational testing, Samuel Messick: "Validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual—as well as potential—consequences of score interpretation and use."[14]

Given this context, it is crucial for the health professions, being largely self-regulated,[15] to devise a system for determining whether the data collected in GME affords us the confidence to make decisions for individual resident progression through their training program to the point of board eligibility. One way of addressing this accountability constraint is to consider the Milestones as a large and complex assessment system. By viewing this problem as a complex

assessment process, the measurement of validity becomes a focus for analysis and a means for improvement (BOX).

BOX Key Points

- Ratings of Milestones attainment can be considered part of a complex assessment system, and therefore, subject to the same expectations for validity as any psychometric assessment process.
- Data should only be reported in the context of interpretive statements and assumptions that are relevant to that particular stakeholder group.
- In the absence of perfect knowledge about a resident's true competence at any point in time, the next best thing is to put a CQI system in place to systematically improve the quality of estimates about that competence.
- Unexpected patterns of data allow for detailed feedback to trainees (for the creation of individualized learning plans) or programs (ie, for curricular quality improvement). Either option improves subsequent validity of the Milestones data and is entirely consistent with a CQI approach.

Applying a Validity Framework to Milestone Ratings

Validity is a key concept in any field that is concerned with measurement. In physics and astronomy, instruments are carefully calibrated prior to taking measurements of particles or objects. This is, in essence, an exercise in validity; without calibration, confidence in one's measurements is reduced. There are many approaches to calibrating an instrument. Reliability is the simplest and most obvious one: Does the instrument produce the same result if the measurement is taken again on exactly the same phenomenon? Other calibration questions relate to validity as well: Does the instrument measure the intended phenomenon? Does it correlate as expected with related variables, or is it affected significantly by irrelevant variables? An example from physics would be the presence of artifacts on a radiographic image, which might lead to erroneous conclusions. In the same way, Milestone ratings are subject to a variety of factors that include both meaningful "signal" (ie relevant to the resident's competence) and "noise."

In theory, the Milestone ratings that are received at the ACGME every 6 months can be formally expressed as being comprised of signal and noise. In addition, the noise element can be separated further into components that can be measured (and hence, reduced or controlled for) and those that remain unknowable (ie, random variation).

$$M = t + \sigma_{CI} + \varepsilon \quad (1)$$

Where:

M – represents a single resident's Milestones achievement score as reported to the ACGME by the CCC,

t – represents the resident's "true" ability (ie, the "signal" or "construct of interest"),

σ_{CI} – represents measurable variance due to factors irrelevant to the primary construct of interest, and

ε – represents any additional residual variance otherwise unaccounted for.

This formal model can be helpful as a statement—it is essentially a concise way of representing all possible factors that go into generating a Milestone rating. By stating the model in this way, we can readily see which components need to be focused on to improve the quality of the data (ie, the validity of Milestone ratings). An obvious way of doing this is to further deconstruct the σ_{CI} and ε components of Equation 1 above (ie, the noise components). In doing so, we can readily identify a number of sources of variance in Milestone ratings that are unrelated to the construct of interest; σ_{CI} reflects systematic irrelevant variance and ε represents "random" (or un-measured) variance in the Milestone rating. For example, a simple summary of such sources of construct-irrelevant variance could be specified as follows.

The observed variance in reported Milestone ratings may reflect:

"t" – The construct of interest:

1. The resident's true competence at the point in time the measurement was taken

"σ_{CI}" – Construct-irrelevant variance (systematic portion, can be measured):

2. Variation in exposure to certain subcompetencies
3. Incomplete specification of the underlying construct in the Milestones language
4. Quality of assessment tools or processes
5. Rater factors
6. Curriculum factors/resources

"ε" – Construct-irrelevant variance (random portion, cannot be measured):

7. Other unwanted sources of error

In terms of validity theory, the construct-irrelevant (noise) variance component in Milestone ratings can

be considered a "threat to validity," in that the extent of this variance can overwhelm the signal we are looking for in the data. In this sense, the construct-irrelevant variance components become targets for intervention in terms of processes for improving the quality of data. This might include, for example, faculty development for reducing rater bias. The following are potential sources of construct-irrelevant variance: (1) faculty rating processes, including opportunity to observe; (2) degree and quality of faculty development regarding education, specific curriculum, and assessment processes; (3) quality of assessment instruments and degree of construct alignment; (4) CCC structure; (5) CCC processes like straight-lining; (6) variations in understanding of the Milestones language for that specialty; (7) institutional and program culture regarding education and accreditation requirements; (8) perceptions about role of assessment in curriculum; and (9) implementation validity (correspondence of NAS vision and program implementation).

Using the Validity Framework to Improve the Quality of Milestone Ratings

Messick built on what has come to be known as the "classical" model of validity, involving 3 major components—content, construct, and criterion-related validity. He proposed a simplification and realigning of these fundamental constructs into a unified framework. In essence, his first achievement was to frame all aspects of validity (including reliability) into the concept of construct validity[16]:

- Integration of evidence that bears on the interpretation/meaning of scores
- Measure is just one of an extensible set of indicators of the construct
- Part of construct validity is construct representation, decomposing the task into requisite component processes and assembling them into a functional model or process theory
    - Where "construct representation" refers to the relative dependence of task responses on the processes, strategies, and knowledge implicated in task performance

In specifying each of these components with greater clarity, Messick arrived at the following 5 essential elements of validity, or what has come to be known as the "modern unified theory of validity":

1. Content (ie, test items are representative of the construct of interest; eg, an expert group writes the content for each test item or Milestone)
2. Response process (ie, evidence of data integrity, including clear test instructions for candidates, rigorous rater training, methods for scoring, and data entry; eg tools constructed with rater in mind, well-accepted, feasible scoring processes)
3. Internal structure (ie, psychometric properties of the examination including score reliability, examination difficulty, and interitem correlations that help to assess factor structure; eg the number of dimensions or subscales that are latent in the construct of interest)
4. Relations with other variables (ie, convergent and discriminant evidence, including correlations to other variables that would be expected based on theory)
5. Consequences (ie, impact on learners, instructors, and the system in which the assessment is made, such as the curriculum or other high-stakes contexts such as certification or accreditation).

Using this formulation of validity promoted by Messick and now codified in the *Standards for Educational and Psychological Testing*,[17] we can apply this approach to interpret and analyze the strength of validity arguments for decisions made using the Milestones data that are received at the ACGME from residency and fellowship programs (with examples provided for illustration):

1. Content validity:
    a. Review of Milestones language developed by each specialty (eg, current work on Milestones 2.0)
2. Response processes:
    a. Faculty rating process and understanding of the Milestones language
    b. Investigation and mitigation of "straight-lining" phenomenon
    c. Development and refinement of assessment tools
    d. Guidelines for selecting and using assessment tools
3. Internal structure:
    a. Interrater reliability of faculty ratings
    b. Reliability of the CCC judgments
    c. Factor analysis of data from various subcompetencies

4. Relations with other variables:
    a. Correlations with other independently obtained performance measures, such as board scores, United States Medical Licensing Examination, and case logs
    b. Predictive probability for patient outcomes
    c. Population health outcomes
5. Consequences:
    a. Understanding the needs of the various stakeholder groups and the manner in which Milestones data might be interpreted by these different audiences, especially use of Milestone ratings for high-stakes purposes vs formative assessment and feedback

Strategies for Interpretation

One approach to addressing the complexity in the national Milestones dataset, and to make use of these data to improve residency training, involves analyzing the validity of the decisions drawn from the data to allow for more effective tools for program directors, CCCs, designated institutional officials (DIOs), and others to help them improve their training programs and clinical learning environments.

The validity framework described here allows for the systematic analysis of the various factors that might influence the Milestones ratings as submitted to the ACGME. Ongoing research at the ACGME and in collaboration with external co-investigators includes an iterative approach to assessing the validity of decisions made by the ratings, guided by theoretical questions of interest to stakeholders. One component of this process is to determine the various stakeholder groups who will receive the reports of both aggregate and individual Milestones data. For example, DIOs, program directors, residents, policy makers, and the public have different needs and uses for the data (BOX). For some it may be formative, and for others summative (ie, the assurance of competency for unsupervised practice). Each of these different uses infer different aspects of the data that might be useful for different purposes, which has implications for data analysis and interpretation. In all cases, the data should only be reported in the context of interpretive statements and assumptions that are relevant to that particular stakeholder group. Another way of saying this is that the analyst should be fully aware of the consequences of their analysis, and thus provide context and guidance for interpretation.[16,18]

Analysis, Interpretation, and Communication of Milestones Data

The modern approach to validity espoused by Messick offers a framework to guide our strategies for analysis that relate back to the vision for NAS and the needs of various stakeholders.[1] This is because the validity framework itself explicitly recognizes the limit on validity of decisions made using Milestones data and as such, advocates for processes for continuously monitoring the quality of the data received from any assessment system. This is consistent with the spirit of a well-designed continuous quality improvement (CQI) system and amounts to an interpretive approach that goes beyond simple analysis and generation of descriptive statistics. It should be noted that by adopting this approach, we are not necessarily advocating the use of psychometric theory *per se* as a means of analyzing Milestones data, but rather the larger framework of validity theory can be useful as a foundation to build on. It just so happens that psychometrics is the field where validity theory has been most clearly articulated and fully studied.

One way of addressing the public accountability mandate is to consider the entire Milestones dataset as a large and complex system for making judgments of resident performance and progression. In particular, by viewing this problem as a complex system for supporting such judgments, the validity of the supporting data becomes a focus for analysis. By systematically and continuously inspecting the data stream against expected values, we can build a means for improving the quality of the data we receive as well as the quality of the educational programming in which residents and fellows participate.

How a Validity Framework Contributes to the CQI Process

Analyzing Milestones data from a validity framework aligns with the CQI approach of the NAS. By recognizing that we will never have perfect estimates of any resident's true ability at any point in time, the next best thing is to put into place a CQI system for monitoring the quality of data with a feedback loop for continuous improvement in the quality of the data. This approach—borrowed from the field of systems science—should be familiar to any health care professional who participates in clinical quality assurance programs to enhance the quality and efficiency of health care delivery; here this approach is applied to medical education (BOX).

The key in any CQI feedback framework is to regularly communicate back to programs in the field regarding their performance so that they can make adjustments to their training programs, the resources

available for education, or the clinical mix to which residents are exposed, to help achieve the overall goal of improving the quality of training and assessing residents so that they can graduate with the competencies necessary for independent practice. One recent example of this is the "learning analytics" work on preparing predictive probability values for use by program directors.[19] This effort involves calculating the probability, based on national data, that residents who obtain a certain Milestone rating at any time point within their training would achieve the recommended graduation target of Level 4 at time of graduation.

Discussion

The observed variations in patterns of data at the national level (ie, within and across specialties) provide telltale signs of how the Milestones themselves were constructed and the degree to which they represent the underlying spirit of the Milestones developed for each specialty (BOX). In addition, they afford the opportunity to examine in detail any discrepancies from patterns that might be expected when designing and implementing curricula within a specialty. The validity framework helps us make sense of these discrepancies by highlighting areas for potential concern. For example, in terms of content validity, data for the medical knowledge competency may be found to underrepresent the construct of interest if not correlated with independent tests of medical knowledge, such as in-training examinations or board scores. At the same time, while the in-training examination may be a good proxy for the ultimate board certification examination in that specialty, it may be more valid to consider other aspects of the medical knowledge construct when teaching and assessing residents in the clinical environment, to allow them to both develop and display competence of the *application* of medical knowledge to solve clinical problems. By analyzing data within the validity framework, it allows us to systematically make such inferences, whether for formative or summative purposes.

Conclusions and Next Steps

Ongoing work on Milestones 2.0 represents an effort to enhance the quality of data by revising the Milestones language for all specialties and streamlining the Milestones reporting forms across all 150 specialties to assist program directors and CCCs in generating valid and defensible ratings. As such, this effort addresses content validity by working toward an explicit shared mental model about how these specific competencies might be described and implemented across programs within a specialty. This also has implications for efficiencies in faculty development for assessment, which might ultimately lead to more valid judgments of performance. To enhance response process validity, courses and web-based resources have been developed to further assist faculty and program directors in Milestones implementation and interpretation, including FAQs, guidebooks, and webinars posted on the ACGME website. To investigate internal structure validity, research is currently being conducted using exploratory and confirmatory factor analysis approaches to uncover latent structure in how groups of subcompetencies are aligned. Finally, more recent work on the potential of Milestone ratings to predict patient outcomes following graduation are underway, representing one of the most important aspects of validity—relations with other variables.

In conclusion, the Milestones data are complex—both in their structure and the processes and context in which they are collected—and caution is still necessary in how these results are interpreted and communicated to various stakeholders. The validity of the data is only beginning to emerge. As such, there are potentially serious implications for misinterpretation, especially if high-stakes decisions are made without regard to construct-irrelevant variance that currently exists in the data. A validity framework can guide us in the process of CQI and help to realize the vision of the NAS as articulated by Nasca and colleagues in 2012.[1]

References

1. Nasca TJ, Philibert I, Brigham T, Flynn TC. The Next GME Accreditation System—rationale and benefits. *N Engl J Med*. 2012;366(11):1051–1056. doi:10.1056/NEJMsr1200117.
2. Hauer KE, Vandergrift J, Lipner RS, Holmboe ES, Hood S, McDonald FS. National internal medicine milestone ratings: validity evidence from longitudinal three-year follow-up. *Acad Med*. 2018;93(8):1189–1204. doi:10.1097/ACM.0000000000002234.
3. 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. doi:10.1001/jama.2016.17357.
4. Francisco GE, Yamazaki K, Raddatz M, et al. Do milestone ratings predict physical medicine and rehabilitation (PM&R) board certification examination

scores? *Am J Phys Med Rehabil*. 2021;100(2S suppl 1):34–39. doi:10.1097/PHM.0000000000001613.
5. Santen SA, Yamazaki K, Holmboe ES, Yarris LM, Hamstra SJ. Comparison of male and female resident milestone assessments during emergency medicine residency training: a national study. *Acad Med*. 2020;95(2):263–268. doi:10.1097/ACM.0000000000002988.
6. Holmboe ES, Yamazaki K, Hamstra SJ. The evolution of assessment: thinking longitudinally and developmentally. *Acad Med*. 2020;95(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 59th Annual Research in Medical Education Presentations):7–9. doi:10.1097/ACM.0000000000003649.
7. Edgar L, Roberts S, Yaghmour NA, et al. Competency crosswalk: a multispecialty review of the Accreditation Council for Graduate Medical Education milestones across four competency domains. *Acad Med*. 2018;93(7):1035–1041. doi:10.1097/ACM.0000000000002059.
8. Hauer KE, Cate OT, Boscardin CK, et al. Ensuring resident competence: a narrative review of the literature on group decision making to inform the work of clinical competency committees. *J Grad Med Educ*. 2016;8(2):156–164. doi:10.4300/JGME-D-15-00144.1.
9. Ekpenyong A, Baker E, Harris I, et al. How do clinical competency committees use different sources of data to assess residents' performance on the internal medicine milestones? A mixed methods pilot study. *Med Teach*. 2017;39(10):1074–1083. doi:10.1080/0142159X.2017.1353070.
10. Conforti LN, Yaghmour NA, Hamstra SJ, et al. The effect and use of milestones in the assessment of neurological surgery residents and residency programs. *J Surg Educ*. 2018;75(1):147–155. doi:10.1016/j.jsurg.2017.06.001.
11. Edgar L, Roberts S, Holmboe E. Milestones 2.0: a step forward. *J Grad Med Educ*. 2018;10(3):367–369. doi:10.4300/JGME-D-18-00372.1.
12. Hamstra SJ, Yamazaki K, Barton MA, Santen SA, Beeson MS, Holmboe ES. A national study of longitudinal consistency in ACGME milestone ratings by clinical competency committees: exploring an aspect of validity in the assessment of residents' competence. *Acad Med*. 2019;94(10):1522–1531. doi:10.1097/ACM.0000000000002820.
13. Beeson MS, Hamstra SJ, Barton MA, et al. Straight line scoring by clinical competency committees using emergency medicine milestones. *J Grad Med Educ*. 2017;9(6):716–720. doi:10.4300/JGME-D-17-00304.1.
14. Messick S. Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol*. 1995;50(9):741–749. doi:10.1037/0003-066X.50.9.741.
15. Nasca TJ. Professionalism and its implications for governance and accountability of graduate medical education in the United States. *JAMA*. 2015;313(18):1801–1802. doi:10.1001/jama.2015.3738.
16. Messick, S. Validity. In: R. L. Linn, ed. *Educational Measurement*. 3rd ed. New York, NY: Macmillan; 1989:13–103.
17. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 2014.
18. Hubley AM, Zumbo BD. Validity and the consequences of test interpretation and use. *Soc Ind Res*. 2012;103(2):219–230.
19. Holmboe ES, Yamazaki K, Nasca TJ, Hamstra SJ. Using longitudinal milestones data and learning analytics to facilitate the professional development of residents: early lessons from three specialties. *Acad Med*. 2020;95(1):97–103. doi:10.1097/ACM.0000000000002899.

At the time of research, **Stanley J. Hamstra, PhD,** was Vice President, Milestones Research and Evaluation, Accreditation Council for Graduate Medical Education (ACGME), and is now Professor, Department of Surgery, University of Toronto, Adjunct Professor, Department of Medical Education, Feinberg School of Medicine, Northwestern University, and Research Consultant, ACGME; and **Kenji Yamazaki, PhD,** is Senior Analyst, Milestones Research and Evaluation, ACGME.

The authors would like to thank the program directors who attended the feedback sessions at the ACGME Annual Educational Conferences from 2015 to 2020 for their valuable insights, as well as the various stakeholder groups to whom we have presented preliminary versions of these results.

Corresponding author: Stanley J. Hamstra, PhD, University of Toronto, Toronto, Ontario, Canada, stan.hamstra@utoronto.ca, Twitter @stanhamstra