ACGME Milestones in the Real World: A Qualitative Study Exploring Response Process Evidence

Ashley M. Maranich, MD
Paul A. Hemmer, MD, MPH
Sebastian Uijtdehaage, PhD
Alexis Battista, PhD

ABSTRACT

Background Since the Accreditation Council for Graduate Medical Education (ACGME) introduced the Milestones in 2013, the body of validity evidence supporting their use has grown, but there is a gap with regard to response process.

Objective The purpose of this study is to qualitatively explore validity evidence pertaining to the response process of individual Clinical Competency Committee (CCC) members when assigning Milestone ratings to a resident.

Methods Using a constructivist paradigm, we conducted a thematic analysis of semi-structured interviews with 8 Transitional Year (TY) CCC members from 4 programs immediately following a CCC meeting between November and December 2020. Participants were queried about their response process in their application of Milestone assessment. Analysis was iterative, including coding, constant comparison, and theming.

Results Participant interviews identified an absence of formal training and a perception that Milestones are a tool for resident assessment without recognizing their role in program evaluation. In describing their thought process, participants reported comparing averaged assessment data to peers and time in training to generate Milestone ratings. Meaningful narrative comments, when available, differentiated resident performance from peers. When assessment data were absent, participants assumed an average performance.

Conclusions Our study found that the response process used by TY CCC members was not always consistent with the dual purpose of the Milestones to improve educational outcomes at the levels of residents and the program.

Introduction

In 2013, the Accreditation Council for Graduate Medical Education (ACGME) introduced the Milestones as a way for residency programs to "...monitor and iteratively improve educational outcomes...at the level of the individual learner and the program."[1] Milestones serve a dual purpose—provide formative feedback to residents and inform the quality improvement of residency programs. Since the Milestones were implemented, several studies have used Messick's framework to examine "...the degree to which evidence and theory support the interpretation of [Milestone] scores for proposed uses..."[2] These include studies pertaining to the content,[3,4] internal structure,[5,6] correlation,[7] and consequence[8,9] aspects of the validity of the Milestones. However, there remains a dearth of evidence for response process. In the context of Milestones, response process pertains to the extent to which the processes of Milestone raters are consistent with the intended interpretation of Milestone ratings and their dual purpose. Specifically, it involves *how* members of a Clinical Competency Committee (CCC) understand, interpret, and operationalize tasks to "review the completed evaluations to select the Milestone levels that best describe each learner's current performance, abilities, and attributes for each subcompetency."[10] Without response process evidence showing that the "real world" use of Milestones aligns with the ACGME's intent, evidence supporting the validity of Milestones remains insufficient.

Prior studies have hinted at, but not directly assessed, response process evidence for Milestones. For example, in 2016, Dzara et al conducted an interview-based study of program directors across multiple specialties to gather information about Milestone implementation. Among the findings, participants indicated that some programs relied on "a benchmark approach to assigning Milestones levels," assuming time-based achievement in the absence of complete assessment data,[11] signaling a flawed interpretation of, or approach to, Milestone ratings. Furthermore, 3 additional studies, with the

DOI: http://dx.doi.org/10.4300/JGME-D-21-00546.1

Editor's Note: The online version of this article contains the interview guide used in the study and a description of observed Transitional Year Clinical Competency Committee meetings.

aim of examining response process evidence,[12-14] retrospectively analyzed Milestone ratings for groups of residents, using quantitative analyses of large data sets to infer how raters use Milestones in resident assessment. While these studies provide a "forest level" view of Milestone assignments, none directly assessed *how* individual CCC members think through the rating process or any other factors that may lead them to select a specific rating.

Our study sought to fill this gap by using qualitative methods to *directly query* individual residency CCC members about their application of Milestones. The purpose of this study was to examine individual CCC members' response process in applying Milestones for resident assessment, specifically exploring their understanding of the purpose of the Milestones, training for assessing residents using Milestones, and thought process when assigning Milestone ratings to a resident.

Objectives
This study used interviews of Clinical Competency Committee members to qualitatively explore validity evidence pertaining to the response process of Accreditation Council for Graduate Medical Education (ACGME) Milestone implementation.

Findings
Our study found 5 different themes around response process evidence: an absence of formal training, Milestones used primarily for resident assessment, challenges in the translation of data into Milestone values, the utility of meaningful comments, and assignment of Milestone values in the absence of data.

Limitations
This study is limited by a small sample size with a focus on the Transitional Year internship.

Bottom Line
Our work suggests that Milestones may not be applied in the "real world" as intended by the ACGME. Although more work into response process is indicated, there is a need for more resources and training for those who are implementing Milestone assessment in residency programs.

Methods

We focused on individual CCC members as a starting point for gathering evidence, recognizing that Milestone decisions are ultimately a group decision. We selected a qualitative research approach using cognitive interviewing strategies near the time of a CCC meeting because it required participants to describe their actual thought process and allowed an interviewer to probe for details to gain a fuller understanding of responses.[15] Utilizing thematic analysis with a constructivist paradigm,[16] which acknowledges that research is co-constructed by researchers and participants, our aim was to take a structured approach to investigating this phenomenon that has not previously been well described. We used reflexivity when designing and conducting the study to recognize the interviewer's prior experiences as a Transitional Year (TY) program director and TY CCC member as well as the other research team members' prior knowledge and experiences with residency program direction, assessment design, and validation.

Study Participants

We sought a purposeful sample of CCC members with the goal of maximum variation[17] in program location, affiliation, and size as well as individual CCC member specialty and teaching experience. To achieve this, we selected 1-year TY residency programs in order to take advantage of their resident and faculty diversity, with faculty representing many specialties. We anticipated sampling from a broad range of faculty backgrounds to include service on other specialty CCCs. Additionally, at the time this study was conceived, TY programs had implemented the revised Milestones 2.0,[10] ensuring that the specialty-specific Milestones were the most current and consistent throughout the study.

Initial recruitment communication was directed to a convenience sample of program directors known to the authors to obtain their assent for program inclusion in the study. We then contacted CCC chairs to enlist their participation and asked them to identify one additional CCC member of a different medical specialty from their committee for inclusion. Ultimately, we had 8 participants, 2 from each of 4 TY programs (the CCC chair and one other member).

Interview Questionnaire and Data Collection

We used a semi-structured interview guide (see online supplementary data) developed by the research team and piloted with an experienced CCC member who was not a part of the study team or a participant. The guide was developed drawing on perspectives from Cook et al regarding what can represent response process evidence,[18] including questions about training, assessor reflexivity, and thought process. The interview guide utilized initial and follow-up probing questions designed to investigate participants' response process for 2 Milestones. Here, we drew upon cognitive interviewing strategies to "provide evidence about the extent to which psychological processes and cognitive operations performed by the respondents actually match those delineated in the test specifications."[15]

The Milestones used as the basis of discussion were in the Patient Care and Systems-Based Practice

TABLE 1
Participant Demographics and Program Information

Participant	Gender	Specialty	Years in GME (Mean=8.5)	Years on CCC (Mean=3)	Serve on Another CCC?	Program Size
1A	F	Neurology	6	2	Yes	>15
1B	F	Emergency medicine	0.5	0.5	No	>15
2A	M	Ophthalmology	3.5	1.5	Yes	<11
2B	M	Internal medicine subspecialty	11.5	4.5	Yes	<11
3A	M	Emergency medicine	3.5	1.5	No	11–15
3B	F	Pediatrics	11.5	9.5	Yes	11–15
4A	F	Internal medicine	20.5	4.5	No	11–15
4B	F	Pathology	11.5	1.5	No	11–15

Abbreviations: GME, graduate medical education; CCC, Clinical Competency Committee.

domains. These examples were chosen because we judged them to represent sub-competencies that were either readily directly observed (ie, Patient Care) or better determined through indirect methods (ie, Systems-Based Practice). Lastly, given that we collected data during the COVID-19 pandemic, we added questions related to its perceived impact on resident assessment and CCC deliberations.

We collected data between November and December 2020. One team member (A.M.M.) virtually observed a CCC meeting for each program during which Milestone values were determined. Within 1 week, this was followed with a 1:1 interview via video conferencing with each study participant from the program by the same team member (A.M.M.). The time frame was chosen so that the Milestone assessment process remained recent in participants' memory, although we also incorporated stimulated recall from investigator observations and notes.

Data Analysis

The role of researchers and their perspectives in the research process are a critical component of constructivist research and in identifying and dealing with validity threats. Given the primary investigator's prior TY experience, something that was a strength but also posed threats, we assembled a diverse research team that included a PhD researcher with expertise in qualitative research (A.B.), assessment (S.U.), and a second clinical physician educator with graduate medical education (GME) leadership experience (P.H.). This diversity secondarily allowed us to form an interpretive community to continuously assess and address bias in coding and analysis that may have influenced the study conduct and conclusions. This collaborative process also supported intensive, long-term involvement, especially during analysis and writing, and led to a richer set of insights into response process.

All interviews were audio recorded and transcribed, with removal of identifiers and pauses, as well as "ahs" and "ums." Each transcript was read and reread by one team member (A.M.M.) to gain familiarity with the data and begin open coding. Then, the team met to discuss initial codes and selected 2 additional team members (A.B., P.H.) to read selected, but independent, transcripts to support internal credibility and assist with code development. During subsequent coding efforts, we focused on passages where participants discussed their (1) perspectives on the purpose of the Milestones; (2) training relevant to Milestones; and (3) thought process in assigning Milestone values. The research team met biweekly to (1) jointly discuss and refine coding; (2) develop operational definitions to support comparison between participants; and (3) group similar codes to form our themes as part of the interactive and iterative analytic process. Notes of the observed CCC meetings as well as memoranda of reflections and research team discussions were maintained and updated by the interviewer (A.M.M.). NVivo version 12 software (QSR International Inc, Burlington, MA) was used to assist with coding analysis.

This study was determined to be exempt by the Uniformed Services University Institutional Review Board.

Results

Program and Participant Demographics

Participating TY programs included academic medical centers and community-based hospitals of varied size and locations. The 8 participants represented 7 different medical specialties, with half also serving as a member of another specialty program CCC. Participants reported 0.5 to 20.5 years of GME teaching experience and 0.5 to 9.5 years of experience on their TY CCC (TABLE 1).

TABLE 2
Representative Quotations by Theme

Theme	Definition	Participant Response
Absence of formal training	Reference to training provided relevant to role in CCC	"I was given the ACGME, I don't remember what they call it, but there's some overarching document about running a CCC. So, I was given that and tried to read through it. . ." (Participant 1A)
		"I guess my training would have been that [the program director] had me in more of an observational role the first several months." (Participant 3A)
		"I don't know if it's necessarily considered training but the person who was on the CCC last year...he kind of took me through how it works or how he used to do it." (Participant 2B)
		"I did go through and print out everything available from the [ACGME] website." (Participant 4A)
Milestones are primarily a tool for resident assessment	Reference to the purpose of the Milestones in assessment of residents or evaluation of the residency program	"I think it's practical in that. . .it helps us assess and evaluate and then I think there could be some real help when you're giving feedback to the learner and you can say, 'this is where you should be.'" (Participant 1A)
		". . .it's to show that they're [residents] progressing towards becoming independent competent providers." (Participant 2B)
		". . .I think it offers a standardized approach to evaluating them [residents]." (Participant 3A)
		". . .it's all part of a movement in medical education to be a little more precise about what residents are learning, what they should be learning, and trying to measure something that's difficult to measure. . ." (Participant 4B)
Translating data into Milestone values	Reference to how data is used to generate a Milestone value	
▪ A priori assumption of average	Assumption of where a resident *should be*	"The way we have set it up as a group is that they would achieve a 3 across the Milestones, is what we would expect them to have by the time they graduate from their PGY-1 year. And we would expect that after several months of internship they would be at about a 1.5—we would not expect them to be at 3 at this point." (Participant 1A)
		". . .about 2.5 is where we expect people to be at the 6-month mark and that allows room for growth and hopefully by the end of the year they're more like a 3.5 or a 4." (Participant 2A)
▪ Use of mean assessment values	How individual assessment inputs are used	"Whatever they have gotten on their cumulative evaluations to date is going to determine where they fall on this Milestone." (Participant 4B)
		"The associate program directors take the evaluations and look at the scores on each of the individual sections [sub-competencies] for the core competencies and translate them to these Milestones so that it's already summarized. . ." (Participant 3B)
		"We collect data off of [the electronic evaluation system] and get the scatterplots from that, which is a visual image of the actual learner versus their learning community. So, it shows where they are compared to their peers." (Participant 4A)
▪ Comparison to peer assessment data	Use of peer comparison in assessing Milestones	"I usually started people at 1.5 [expected midyear value] and then based on that number that was given in [the electronic evaluation system], I would either keep it there if it seemed to be in line with average or I would move it up and down [if they were up or down from average]." (Participant 1B)
		". . .today at the meeting we talked about peer, below peer, above peer. This is where your peer should be at this. And so as of right now, 2.5 is what we've determined would be where they should be right now." (Participant 2B)
		"And in terms of the core competency skills that we're looking at; where they fall, where it's equivalent from the question on the evaluation as well as where they fall in relation to their peers." (Participant 3B)

TABLE 2
Representative Quotations by Theme (continued)

Theme	Definition	Participant Response
Utility of meaningful narrative comments	Reference to narrative comments in assessment of residents	"...Patient Care and Communication might be easier because my sense is that they get mentioned a lot on the evaluation comments, so we get a lot more information about that...versus something like Practice-Based Learning or Systems-Based that doesn't get mentioned in the comments nearly as much." (Participant 1A)
		"There's always an evaluation comment although it is not always good [useful]." (Participant 1A)
		"...It was just making sure there was nothing necessarily glaring or terrible because most of the evaluations were, 'great,' 'did a great job,' or 'was very professional.' So, there wasn't anything specific to take from that." (Participant 1B)
		"I don't know that I would expect a comment on how they are navigating the health plans of their patients [Systems-Based Practice]." (Participant 2A)
		"...they're just going to always say 'Oh, you're great in [specific specialty], we love you!'" (Participant 4A)
Assigning Milestone values in the absence of data	Reference to Milestone value determination in the absence of data	"...give us a reason to deviate this learner from the average; otherwise, they're going to be on the average." (Participant 1A)
		"We discussed that we expect everybody to be at about a 2.5 for their Milestones, so if we didn't hear anything, we kept it at a 2.5." (Participant 2A)
		"I guess I would say I feel less pressure to get it right early on and that's partially because we have less data, too." (Participant 3A)

Response Process Evidence

Our analysis led to 5 themes surrounding evidence for response process: (1) absence of formal training; (2) Milestones are primarily a tool for resident assessment; (3) translating data to Milestone values; (4) utility of meaningful narrative comments; and (5) assigning Milestone values in the absence of data. Theme definitions and example participant quotations are detailed in TABLE 2.

Theme No. 1—Absence of Formal Training: None of the participants reported having formal training for their role on the CCC. Three of the chairs reported being made aware of the ACGME's guidebook for CCCs by their program director, but none indicated that they had read it in full nor did our analysis suggest that these guidelines informed their decision-making. Some participants referenced informal mentorship by prior committee members or observational activities, although they did not consider this actual training.

Theme No. 2—Milestones Are Primarily a Tool for Resident Assessment: In discussing the purpose of Milestones, participants described them as a useful assessment tool that provides a common frame of reference for faculty and residents. Further, they believed Milestones clarify expectations of residents, provide a common language to guide resident feedback, and define a developmental pathway residents can follow. Participants described Milestone sub-competencies as a blueprint for assessing resident performance and defining standards to be met by the time of graduation.

However, participants' descriptions of the purposes of Milestones centered solely on their being a means for operationalizing competency-based assessment of residents without any reference to the program evaluation and quality improvement functions of Milestones.

Theme No. 3—Translating Data to Milestone Values: When presented with a specific sub-competency (ie, Patient Care, Systems-Based Practice), participants described having a stepwise process in assigning a Milestone value using the available assessment data on a given resident. These 3 steps are outlined as subthemes: (1) a priori assumption of average based on time in training; (2) use of mean assessment values; and (3) comparison to peer assessment data.

A priori assumption of average based on time in training. Before reviewing any individual data, all participants reported holding a preconceived assumption of where a resident *should be* on the Milestone rating scale based on their time in training. For all,

this was determined by viewing the Milestone scale linearly and presuming the scale was designed to depict a typical, or "average," resident to be "around the middle" of the scale by the midpoint of their training. None explicitly referenced the Milestone behavioral anchors as a starting point for this assumed expectation.

Use of mean assessment values. Participants used summarized reports of faculty-completed assessments when determining Milestone values. These reports were automatically collated by a software program that displayed mean values of numerical data and all available narrative comments, which highlights how the TY programs' use of a web-based platform to collect and collate resident data played a role in influencing Milestone assessment choices. Participants described using mean values when considering each Milestone sub-competency, without taking into consideration any descriptive data such as the range and outliers of assessments, number of data points, rotation setting, timing across the 6 months, or whether there was growth, regression, or stagnation. Participants seemed to be passive recipients of information from frontline assessments, viewing these averaged, aggregated data as the current representation of the resident's sub-competency achievement, without further interpretation of the information.

Some participants indicated that, at this stage, Milestone assessment was a challenge because they had little or no interaction with the residents in patient care settings. Additionally, most participants lacked knowledge of the specifics of the frontline assessments (ie, what the checklists entailed or the scales that were used) and assumed the faculty assessments utilized Milestone language directly.

Comparison to peer assessment data. As the final step in assigning a specific Milestone rating, participants compared individual resident "performance," (the mean score on completed assessments) to the collective peer mean on the same item and integrated this information with their expectation of performance at that particular time in training. In other words, residents performing at the level of their peers were assigned the expected Milestone values at the time of year of the CCC meeting. If the resident performed above or below peer average on a sub-competency, participants adjusted the Milestone value up or down, respectively.

Theme No. 4—Utility of Meaningful Narrative Comments: Participants described that they incorporated *meaningful* narrative comments gathered from assessment forms and informal communications, when they were available, in their determination of Milestone ratings. Several participants described they used written comments to help them corroborate the numerical evaluations and to further adjust assigned values for a specific sub-competency. They found comments describing specific behaviors useful, but also wished for *more* descriptive comments and less comments focused on resident personality and likability. Participants reported that some competencies were more likely to draw helpful narratives (eg, Patient Care, Interpersonal and Communication Skills, and Professionalism), while others (eg, Systems-Based Practice and Practice-Based Learning and Improvement) commonly lacked informative comments. This resulted in robust narratives for a narrow spectrum of sub-competencies and likely also contributed to participants' reliance on numerical assessment data for most sub-competencies.

Theme No. 5—Assigning Milestone Values in the Absence of Data: Participants often described having limited or no assessment data for some residents for one or more Milestone sub-competencies, especially in the category of Systems-Based Practice. In the absence of data, participants indicated they assumed that a resident was performing at the level to be expected ("average") and assigned the Milestone value that corresponded with their a priori expectations. In these cases, none of those interviewed reported using the available options of "Not Yet Completed Level 1" or "Not Yet Assessable."

COVID-19 Impact

After the start of the COVID-19 pandemic, participants noted alterations in patient care activities affecting the number, type, and diversity of resident experiences as well as a decreased number of "face-to-face" interactions between residents and faculty. They also described changes to their CCC meetings with a transition to virtual platforms leading to decreased meeting attendance, less engagement by CCC members, and less group interaction and discussion during the meeting. Although some acknowledged increased personal stress from COVID-19, none felt that the pandemic affected their personal approach to assigning Milestone values to residents.

Discussion

To our knowledge, this is the first qualitative study using cognitive interviews to investigate response process validity evidence supporting the use of Milestones as a means to improve educational outcomes at the resident and program levels. Our

work found problematic evidence that CCC members' thought process and approach to Milestone scoring were not always aligned, and sometimes in conflict with, the intended purpose of this assessment.

Our findings suggest that the participants in our study lacked adequate preparation for their role on the CCC, likely contributing to misconceptions about Milestones and a failure to truly use them for competency-based assessment. Regular training of CCC members could better assist their understanding of Milestones, their intended purpose, and the role faculty play as both individual and group members of a CCC in reaching judgments about trainees. The ACGME CCC guidebook emphasizes the need for "deliberate, ongoing faculty development for those who serve on the CCC"[19] and provides resources for this training, including quizzes and case studies. Additionally, given the recent widespread use of virtual learning, both in synchronous and asynchronous interactions, a new opportunity may be present to close the training gap for CCC members.

Frame of reference training of CCC members promotes a shared understanding and consistent use of assessment standards,[20] and may mitigate the inappropriate reliance on time-based achievement expectations and peer comparisons seen with our participants. Ultimately, the goal is to promote the ACGME's intent that "faculty members should be trained to compare each resident's/fellow's performance to the Milestones as a whole, not just to the performance of other or 'typical' residents/fellows in the program."[19] Our findings reinforce the assertion of Peabody and colleagues that "[Family Medicine] Milestones do not measure the amount of a latent trait possessed by a resident, but rather describe where a resident falls along the training sequence."[6] If Milestones are to truly support competency-based education and promote public accountability by ensuring the competency of residency graduates, programs must adequately prepare those who carry out this process and be adequately supported by GME leaders in these efforts.

It is not surprising that programs and CCC members are seeking efficiencies, even if these efficiencies distort the original intent of Milestones. When assessments are automatically compiled and averaged by electronic residency management programs for CCC members to use, meaning can be lost. Multiple faculty perspectives on a single resident adds richness to assessment data. However, when data are reduced to a single number, the diverse range of viewpoints, clinical context, and demonstration of growth over time is obscured. The ACGME intends for Milestones to be "narratives, not numbers"[21]; meaningful narrative comments are the key to moving past average numerical assessment data. Although there are limitations to narrative comments, including that they may not cover the full breadth of subcompetencies[22] or may lack sufficient detail,[23] our participants noted that they can be helpful in corroborating or adjusting numerical scores.

We found that CCC members relied on assumptions about where an individual trainee *should be* based on time in the training program, rather than determining where they *are* performing. This practice is inconsistent with the intent of the Milestone framework and suggests that the "benchmarking" approach found by Dzara et al persists,[11] even after years of experience with the Milestones.

We found it striking that participants described assigning ratings for a resident *in the absence of assessment data*. It remains unclear from our analysis what led participants in our study take these steps; however, based on our team members' professional experiences and informal discussions at professional meetings, one explanation may be that programs view Milestones as high stakes for residents and programs despite the ACGME's intent that they be a formative tool to guide growth and development.[1] Nonetheless, what is the message sent to trainees about the importance of the assessment process when educational program leaders will assign a rating for competency performance when there is no data to inform such a decision?

Although not articulated by our participants, another purpose of the Milestones is to evaluate a program's curriculum and assessment methods and inform program quality improvement efforts. In the places where CCC members are filling in the gaps of assessment data with assumptions, opportunities are being missed (or ignored) to pass along information to Program Evaluation Committees that can be used to improve resident education and assessment of programs. Guiding CCC members to identify, recognize, accept, and then act to close these gaps, rather than making inferences that avoid uncomfortable truths, can improve educational outcomes and should be encouraged whenever possible.

Our study is limited by a small sample size which may have led to some potential missed themes. However, there were no discrepant cases for those themes we did identify. We purposefully selected TY residency programs and recognize they may not fully reflect the response process of members of categorical residency programs; yet half of our participants were also members of another specialty CCC. Furthermore, our study overlapped the COVID-19 pandemic, with interviews delayed from the summer to winter of 2020. This limitation became an opportunity, allowing for direct, albeit virtual,

observation of CCC meetings in advance of interviews, providing a useful context for stimulating recall about the most recent meeting. While all participants articulated an impact from COVID-19 on both the assessment of their residents and format of their CCC meetings, the pandemic did not appear to change their approach to Milestone assessment, and thus we infer that under normal circumstances, we would have identified similar evidence jeopardizing the validity of Milestone ratings. Finally, our study intentionally focused on individual response process as a starting point for gathering evidence. We acknowledge that group decision-making processes are also a factor in response process; future work with larger samples, varied residency specialties, and inclusion of group factors will provide additional insight into this topic.

Conclusions

This qualitative study of response process evidence for the use of the Milestones by individual TY CCC members found evidence that this assessment may not always be applied in the "real world" as originally intended.

References

1. Accreditation Council for Graduate Medical Education. Edgar L, McLean S, Hogan S, Hamstra SJ, Holmboe E. The Milestones Guidebook. Accessed March 3, 2021. https://www.acgme.org/globalassets/milestonesguidebook.pdf
2. Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, and the National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 2014.
3. Aagaard E, Kane GC, Conforti L, et al. Early feedback on the use of the internal medicine reporting milestones in assessment of resident performance. *J Grad Med Educ*. 2013;5(3):433–438. doi:10.4300/JGME-D-13-00001.1
4. 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
5. 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
6. 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. doi:10.4300/JGME-D-16-00172.1
7. Cassaro S, Jarman BT, Joshi ART, et al. Mid-year medical knowledge Milestones and ABSITE scores in first-year surgery residents. *J Surg Educ*. 2020;77(2):273–280. doi:10.1016/j.jsurg.2019.09.012
8. Klein R, Ufere NN, Rao SR, et al. Association of gender with learner assessment in graduate medical education. *JAMA Netw Open*. 2020;3(7):e2010888. doi:10.1001/jamanetworkopen.2020.10888
9. Li ST, Schwartz A, Burke AE, et al. Pediatric program director minimum Milestone expectations before allowing supervision of others and unsupervised practice. *Acad Pediatr*. 2020;20(8):1063–1065. doi:10.1016/j.acap.2020.05.011
10. Accreditation Council for Graduate Medical Education. Transitional Year Milestones. Accessed March 3, 2021. https://www.acgme.org/Portals/0/PDFs/Milestones/TransitionalYearMilestones.pdf
11. Dzara K, Huth K, Kesselheim JC, Schumacher DJ. Rising to the challenge: residency programs' experience with implementing Milestones-based assessment. *J Grad Med Educ*. 2019;11(4):439–446. doi:10.4300/JGME-D-18-00717.1
12. Hauer KE, Clauser J, Lipner RS, et al. The internal medicine reporting Milestones: cross-sectional description of initial implementation in U.S. residency programs. *Ann Intern Med*. 2016;165(5):356–362. doi:10.7326/M15-2411
13. Heath JK, Dine CJ. ACGME Milestones within subspecialty training programs: one institution's experience. *J Grad Med Educ*. 2019;11(1):53–59. doi:10.4300/JGME-D-18-00308.1
14. Tanaka P, Park YS, Roby J, et al. Milestone learning trajectories of residents at five anesthesiology residency programs. *Teach Learn Med*. 2020;33(3):304–313. doi:10.1080/10401334.2020.1842210
15. Padilla JL, Benitez I. Validity evidence based on response processes. *Psicothema*. 2014;26(1):136–144. doi:10.7334/psicothema2013.259
16. Kiger M, Varpio L. Thematic analysis of qualitative data: AMEE Guide No. 131. *Med Teach*. 2020;42(8):846–854. doi:10.1080/0142159x.2020.1755030.
17. Patton MQ. *How to Use Qualitative Methods in Evaluation*. 4th ed. Los Angeles, CA: Sage Publications; 1987.
18. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med*. 2016;91(10):1359–1369. doi:10.1097/ACM.0000000000001175

19. Accreditation Council for Graduate Medical Education. Andolsek K, Padmore J, Hauer KE, Ekpenyong A, Edgar L, Holmboe E. Clinical Competency Committees: A Guidebook for Programs. 3rd ed. Accessed March 3, 2021. https://www.acgme.org/Portals/0/ACGMEClinical CompetencyCommitteeGuidebook.pdf
20. Hemmer PA, Dadekian GA, Terndrup C, et al. Regular formal evaluation sessions are effective as frame-of-reference training for faculty evaluators of clerkship medical students. *J Gen Intern Med*. 2015;30(9):1313–1318. doi:10.1007/s11606-015-3294-6.
21. Edgar L. TY Milestones 2.0—Program Director Panel Discussion. Oral presentation at: Association for Hospital Medical Education Institute; May 2021, virtual meeting.
22. Diller D, Cooper S, Jain A, et al. Which emergency medicine Milestone sub-competencies are identified through narrative assessments? *West J Emerg Med*. 2020;21(1):173–179. doi: 10.5811/westjem.2019.12.44468
23. Raaum SE, Lappe K, Colbert-Getz JM, et al. Milestone implementation's impact of narrative comments and perception of feedback for internal medicine residents: a mixed methods study. *J Gen Intern Med*. 2019;34(6):929–935. doi:10.1007/s11606-019-04946-3

All authors are with the Uniformed Services University. **Ashley M. Maranich, MD,** is Assistant Dean for Clinical Sciences and Associate Professor of Pediatrics; **Paul A. Hemmer, MD, MPH,** is Professor of Medicine and Professor of Health Professions Education; **Sebastian Uijtdehaage, PhD,** is Professor of Medicine and Professor of Health Professions Education, and Associate Editor, *Journal of Graduate Medical Education*; and **Alexis Battista, PhD,** is Assistant Professor of Medicine.

Funding: The authors report no external funding source for this study.

Conflict of interest: Dr Maranich is a member of the ACGME Transitional Year Review Committee. This was disclosed to all participants; no information from this study will have an impact on program accreditation decisions.

Disclaimer: The opinions and assertions expressed herein are those of the authors and do not necessarily reflect the official policy or position of the Uniformed Services University or the Department of Defense.

Corresponding author: Ashley M. Maranich, MD, Uniformed Services University, ashley.maranich@usuhs.edu

Received May 22, 2021; revisions received October 2, 2021, and December 16, 2021; accepted January 5, 2022.