# ACGME Milestones in the Real World:

# A Qualitative Study Exploring Response Process Evidence

by

COL Ashley M. Maranich

Thesis submitted to the Faculty of the

Health Professions Education Program

Uniformed Services University of the Health Sciences

In partial fulfillment of the requirements for the degree of

Masters, Health Professions Education (2021)

Distribution Statement

Distribution A: Public Release.

The views presented here are those of the author and are not to be construed as official or reflecting the views of the Uniformed Services University of the Health Sciences, the Department of Defense or the U.S. Government.

UNIFORMED SERVICES UNIVERSITY, SCHOOL OF MEDICINE GRADUATE PROGRAMS
Graduate Education Office (A 1045), 4301 Jones Bridge Road, Bethesda, MD 20814

September 24, 2021

APPROVAL SHEET

Title of Dissertation: **ACGME Milestones in the Real World: A Qualitative Study Exploring Response Process Evidence**

Name of Candidate: **COL Ashley M. Maranich**, Master of Science in **Health Professions Education**

**Date 09/24/2021**

THESIS AND ABSTRACT APPROVED: DATE:

[redacted]

09/24/2021

Abigail W. Konopasky, PhD
Center for Health Professions Education
Committee Chairperson

[redacted]

09/24/2021

Alexis Battista, PhD
Center for Health Professions Education
Thesis Advisor

[redacted]

9/27/201

Sebastian Uijtdehaage, PhD
Center for Health Professions Education
Thesis Advisor

HEMMER.PAUL.A.1100211021 Digitally signed by HEMMER.PAUL.A.1100211021 Date: 2021.09.27 12:52:07 -04'00'

Paul Hemmer, MD, MPH, MACP
Department of Medicine
Center for Health Professions Education
Thesis Advisor

# ACKNOWLEDGEMENTS

First and foremost, I would like to acknowledge my amazing team of research advisors, Alexis Battista, Sebastian Uijtdehaage, and Paul Hemmer. You have all been so incredibly supportive and affirming through this journey, made even longer by a global pandemic. And, you have pushed me to continue to think and grow in all the best of ways. I look forward to a time when we can celebrate this milestone in the real (not virtual) world!

I also want to express my gratitude to my supervisor, Colonel Pamela Williams. As my "household" contact within the workplace throughout the pandemic, you were a constant support through this past 18 months and never stopped encouraging me to see this project through to the end. It seems ironic that the only days I spent tele-working were ones that were dedicated to my thesis work; thanks for always looking to take work off my list to ensure my own work remained a priority.

Finally, I want to thank Dr. Kimball Mohn and the leadership team at the Association for Hospital Medical Education (AHME). My involvement in AHME fueled my passion for Transitional Year internship training and directly led to my selection for the ACGME's Transitional Year Review Committee. You have truly built a great community of diverse and talented educators and I am forever grateful to have found my way into the AHME family.

# DEDICATION

This is dedicated to my parents who have always encouraged me to reach for the stars and been ready to catch me if I should fall. Thank you from the bottom of my heart.

# COPYRIGHT STATEMENT

The author hereby certifies that the use of any copyrighted material in the thesis manuscript entitled: ACGME Milestones in the Real World: A Qualitative Study Exploring Response Process is appropriately acknowledged and, beyond brief excerpts, is with the permission of the copyright owner.

Ashley M. Maranich
Candidate, MHPE
Uniformed Services University
23Feb2022

# DISCLAIMER

The views presented here are those of the author and are not to be construed as official or reflecting the views of the Uniformed Services University of the Health Sciences, the Department of Defense or the U.S. Government.

# ABSTRACT

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

By MHPE Candidate, COL Ashley M. Maranich, MD

Thesis directed by: Alexis Battista, PhD, Assistant Professor of Medicine, Uniformed Services University

**Statement of the Problem:** The growing body of validity evidence supporting the use of ACGME Milestones has grown, but response process remains poorly studied.

**Methods**: We conducted semi-structured interviews of Transitional Year Clinical Competency Committee members to query their response process in the application of Milestones, drawing from constructivist grounded theory.

**Results:** Themes included an absence of formal training and the use of Milestones for resident assessment, ignoring their role in program evaluation. Regarding thought processes, themes were: comparing averaged data to peers and time in training, utility of meaningful narrative comments, and assumption of average performance in the absence of data.

**Conclusions**: Our study found that the response process used by TY CCC members is not consistent with the ACGME's intent. This work suggests a number of ways to improve the application of Milestones in the "real world" as well as the importance of further investigation into evidence for response process.

# TABLE OF CONTENTS

# LIST OF TABLES

# CHAPTER 1: Introduction

## BACKGROUND

The Accreditation Council for Graduate Medical Education (ACGME) introduced Educational Milestones for Graduate Medical Education (Milestones) in 2013. The Milestones, based on the Dreyfus Model of expertise development, were independently developed for each medical specialty (e.g., Emergency Medicine, Pediatrics), and intended to assist residency and fellowship programs with operationalizing competency-based educational outcomes and providing formative feedback to trainees in support of lifelong learning. (25)

Each medical specialty has a list of 15-30 Milestones, listed as sub-competencies that are nested under the six core competencies of ACGME training (Patient Care, Systems Based Practice, Interpersonal and Communication Skills, Medical Knowledge, Problem Based Learning and Improvement, and Professionalism). Residency and fellowship programs are required to have Clinical Competency Committees (CCCs), comprised of program faculty, who are charged with assessing Milestone values for each resident twice a year. To support and guide this process, the Milestones provide common, narrative expectations ("behavioral anchors") that can be used as a criterion-referenced assessment tool to evaluate resident trainees (25). Furthermore, while Milestone values must be reported to the ACGME biannually, this process is intended to be a low-stakes evaluation with no accreditation or graduation implications.

Although many residency programs already had CCCs prior to the introduction of Milestones, the work of these committees were largely focused on

problem identification, ensuring interventions for low-performing residents, instead of a more global developmental model benefitting all trainees (18). As these committees are now charged with the responsibility for Milestones, the academic world is now working to better understand how CCCs "work" in determining consensus decisions and to identify best practices (7, 11, 12, 20). One common model used by residency programs to increase CCC efficiency is to assign individual members specific preparatory tasks in advance of group discussions (33). This means that, commonly, a single CCC member will be charged with reviewing available assessment data and assigning Milestone ratings to a single learner (or group of learners) to present to the entire group for discussion and ratification. Thus, a complete understanding of how Milestone ratings are assigned must include both the CCC as a whole and its individual members.

As Milestones were being introduced and implemented, the GME community largely focused on implementing this new assessment tool at a program level. As Milestones are increasingly becoming routine for GME residents and educators, it is important to critically assess whether they are indeed meeting the intended purposes established by the ACGME (Table 1). Central to this process is an examination of the evidence for the validity of Milestones as used by residency programs.

Table 1. The Purpose and Function of Milestones

*From The Milestones Guidebook (25)*

| Constituency or Stakeholder | Purpose/Function |
|---|---|
| Residents and Fellows | Provide a descriptive roadmap for education and training<br>Increase transparency of performance requirements<br>Encourage informed self-assessment and self-directed learning<br>Facilitate better feedback to learner<br>Encourage self-directed feedback seeking behaviors |
| Residency and Fellowship Programs | Guide curriculum and assessment tool development<br>Provide meaningful framework for CCC (e.g., help create shared mental model)<br>Provide more explicit expectations of residents and fellows<br>Support better systems of assessment<br>Enhance opportunity for early |

| | |
|---|---|
| | identification of under-performers<br>Enhance opportunity to identify advanced learners to offer them innovative educational opportunities |
| ACGME | Accreditation - enable continuous improvement of programs and lengthening of site visit cycles<br>Public Accountability – report at an aggregated national level on Competency outcomes<br>Community of practice for evaluation and research, with focus on continuous improvement |
| Certification Boards | Enable research to improve certification process |

## THEORETICAL FRAMEWORK

For this work, we chose Messick's definition of validity as the theoretical framework. As articulated in The Standards for Educational and Psychological Tests (26), this framework considers validity evidence across five categories of sources as

defined below: content, response process, internal structure, relationships with other variables, and consequences (Table 2).

Table 2. Summary of Messick's Sources of Validity Evidence

*Adapted from The Standards for Educational and Psychological Tests (26)*

| Validity Type | Operational Definition |
|---|---|
| Content | Relationship between the content of a test and the construct it is intended to measure |
| Response Process | Detailed nature of the performance or response actually engaged in by test takers |
| Internal Structure | Degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based |
| Relationship to other Variables | Relationship of test scores to variables external to the test |
| Consequences | Evaluate(s) the soundness of proposed interpretations [of scores] for their intended use |

## EXISTING VALIDITY EVIDENCE

Early publications addressing content validity were conducted by the different groups involved in the initial Milestone design for their given specialty (1, 29). Additionally, multiple studies have addressed relationships with other variables, with all but one showing a correlation of Milestones assessments with other summative assessment tools of Medical Knowledge, such as annual in-training examinations and board certification rates across multiple specialties (5, 6, 16, 21, 27, 32).

There are also several published studies addressing internal structure evidence for Milestones, although there is conflicting interpretation of similar data. For example, some studies found Milestones to be reliable in discriminating between learners at different levels of training. (4, 22, 41) However, this effect has also viewed, not as evidence of good internal structure, but instead, as proof that 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 (36)." Other concerning evidence in the domain of internal structure includes studies showing higher frequencies of "straight line scoring" (same Milestone value across all sub-competencies or within a group of sub-competencies) than expected by chance (3, 39).

Evidence regarding the consequences of Milestones is also a growing body of research. Although intended for formative assessment rather than high stakes decisions (25), studies show that the Milestones are being used for entrustment and graduation decisions (31), expanding their impact beyond the intended use. Additionally, studies of bias across gender lines with Milestones ratings have shown conflicting results (28, 30),

raising concerns about their use in supporting diversity, equity, and inclusion efforts to develop a diverse workforce of physicians.

### Response Process Evidence

Overall, response process remains poorly studied. Defined as "the detailed nature of the performance or response actually engaged in by...observers or judges to record and/or evaluate examinees' performance" (26), response process, when applied to Milestones, involves how members of a CCC understand, interpret, and operationalize tasks to "review the completed evaluations to select the Milestones levels that best describe each learner's current performance, abilities, and attributes for each sub-competency" (2). 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.

The materials and trainings developed by the ACGME to provide guidance for individuals, programs, and CCCs are among the most commonly cited response process evidence, but do not truly meet Messick's definition (Table 2). A few studies have attempted to indirectly search for evidence of response process, using retrospective analysis of Milestone ratings for large groups of residents to infer how raters assign values (19, 23). Additionally, Dzara et al (13) conducted an interview-based study that hinted at response process evidence, finding that Program Directors recognized "a benchmark approach to assigning Milestones levels," signaling a flawed approach worthy of further investigation.

## Research Questions

Our study sought to use qualitative methods to directly query individual residency CCC members about their application of Milestones in assigning ratings to a learner. The purpose of this study was to examine CCC members' understanding of Milestones' purpose, their training for assessing residents using Milestones, and thought process when assigning Milestones ratings to a resident. Our research question was, what is the response process used by individual CCC members in the application of Milestones?

# CHAPTER 2: Milestones in the Real World: A Qualitative Study Exploring Response Process Evidence

Ashley M. Maranich; Paul A. Hemmer, Sebastian Uijidehaage, Alexis Battista

Pending revision and resubmission (*Journal of Graduate Medical Education*)

**ABSTRACT**

**Background:** Since their introduction in 2013, the body of validity evidence supporting the use of the Accreditation Council for Graduate Medical Education (ACGME) educational Milestones has grown, but there is a gap with regards 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:** We drew from constructivist grounded theory to guide data collection and analysis. In semi-structured interviews of eight Transitional Year (TY) CCC members conducted after a CCC meeting from November to December 2020, participants were queried about their response process in their application of Milestones assessment. Data were analyzed iteratively.

**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 Milestones ratings. Meaningful narrative comments, when available, differentiated resident performance from peers. When assessment data was absent, participants assumed an average performance.

**Conclusions**: Our study found that the response process used by TY CCC members to assign Milestone ratings is not consistent with the ACGME's intent that they support competency-based education as a means for formative assessment and program

evaluation. If these findings are reproducible across other specialties, the validity of inferences made based upon Milestone ratings would be in question.

## INTRODUCTION

In 2013, the Accreditation Council for Graduate Medical Education (ACGME) introduced Milestones as a way for residency programs to "...monitor and iteratively improve educational outcomes…at the level of the individual learner and the program (25)." Since then, numerous research efforts seeking evidence of Milestones' validity, defined as "...the degree to which evidence and theory support the interpretation of... (26)" assessment data have examined the categories of content (1, 17), internal structure (3, 36), correlation (6), and consequences(28, 31). However, there remains a dearth of evidence for response process. Response process is "the detailed nature of the performance or response actually engaged in by...observers or judges to record and/or evaluate examinees' performance (26)." Applied to Milestones, 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 sub-competency (40)." Without response process evidence showing that the "real world" use of Milestones aligns with the ACGME's intent, evidence supporting validity of Milestone remains incomplete.

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 upon "a benchmark approach to assigning Milestones levels," assuming time-based

achievement in the absence of complete assessment data (13), signaling a flawed interpretation of, or approach to, Milestones ratings. Furthermore, three additional studies, with the aim of examining response process evidence (19, 23, 39), 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 nor 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 implementation of Milestones. The purpose of this study was to examine CCC members' understanding of Milestones' purpose, their training for assessing residents using Milestones, and thought process when assigning Milestones ratings to a resident. Our research question was, what is the response process used by individual CCC members in the application of Milestones?

## METHODS

We selected a qualitative research approach using cognitive interviewing strategies near the time of a CCC meeting because it requires participants to describe their actual thought process and allows an interviewer to probe for details to gain a fuller understanding of responses (34). We drew from constructivist grounded theory (8), which acknowledges that research is co-constructed by researchers and participants and can provide a structured approach to investigating phenomena that are not well described. We used reflexivity throughout the design and conduct of the study to recognize the

interviewer's prior experiences as a 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 one-year TY residency programs because they are unique in their resident and faculty diversity, with faculty representing many specialties, and allowed us to sample from a broad range of backgrounds and experiences in committee membership. Additionally, at the time this study was conceived, TY programs had implemented the revised Milestones (2.0) (40), ensuring that the specialty specific Milestones were the most current and remained consistent throughout the study.

Initial recruitment communication was directed to Program Directors, ensuring their assent for program inclusion in the study. After receiving approval, we 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 eight participants, two from each of four TY programs (the CCC Chair and one other member).

**Interview Questionnaire and Data Collection**

We used a semi-structured interview guide (Appendix A) developed by the research team and piloted with an experienced CCC member who was not a part of the study team nor a study participant. The guide was developed drawing on the perspectives of Cook and colleagues on what can represent response process evidence (9), including questions about prior training, assessor reflexivity, and thought processes. The interview guide also utilized initial and follow-up probing questions designed to investigate participants' response process for two 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 (34)."

The Milestones used as the basis of discussion were in the Patient Care and Systems Based Practice domains. These examples were chosen because we judged them to represent sub-competencies that were either readily directly observed (i.e., Patient Care) or better determined through indirect methods (i.e., 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 (AMM) virtually observed a live CCC meeting for each program during which Milestone values were determined for all trainees in the program. Within one week, this was followed with a 1:1 interview via video conferencing with each study participant from the program. The time frame was chosen so that the Milestone assessment process remained

recent in the participants' memory, although we also incorporated stimulated recall from investigator observations and notes.

### Data Analysis

All interviews were audio recorded and transcribed, with removal of identifiers and pauses, "ahs" and "uhms". Then, each transcript was read and reread by one team member (AMM) to gain familiarity with the data and begin coding. Other members of the research team (AB, SU, PH) reviewed and discussed the codes as part of their contribution to the interactive and iterative analytic process. During subsequent coding passes, we focused on passages where participants discussed their a) perspectives on Milestones' purpose, b) training relevant to Milestones, and c) thought processes in assigning Milestone values. NVivo Version 12 software was used to assist with coding analysis. (QSR International Pty Ltd.)

Notes of the observed CCC meetings as well as memoranda of reflections and research team discussions were maintained by the interviewer (AMM) to inform analysis. Understanding the role of researchers and their perspectives in the production of grounded theory, we assembled a diverse research team that included PhD researchers with expertise in qualitative research, assessment, and clinical physician educators with GME leadership experience.

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 geographical location. The eight participants represented seven different medical specialties of practice with half reporting 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 3).

## Response process evidence

Our analysis led to five themes surrounding evidence for response process: (1) the absence of formal training; (2) Milestones primarily as a tool for resident assessment; (3) making the data fit; (4) moving the needle; and (5) "no data? no problem." Theme definitions and example participant quotations are detailed in Table 4.

### *Theme #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 Guide for Clinical Competency Committees 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 #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 resident expectations, 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 at the time of graduation.

*"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]

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

***Theme #3 – Making any Data Fit: The Importance of Average***

In this theme, the concept of "average" was a driving force in decision-making. In the paragraphs that follow, we describe three sub-themes that comprise this theme; *a priori* assumption of average, use their average, and comparison to peer average.

"*a priori* assumption of average": Participants described a step-wise process using available data to generate a Milestone value. Before reviewing any data, all participants reported holding a preconceived assumption of where a resident *should be* on the Milestone overall rating scale based on their time in training, yet none explicitly referenced the Milestones behavioral anchors as a starting point for this assumption.

"Use their average": Participants described that they viewed and used faculty-completed assessments in the form of summarized reports when determining Milestone values. These reports were first collated by the program using a software program and consisted of averaged numerical data and all available narrative comments. Participants described using this averaged value for each Milestone sub-competency, but did not express taking into consideration any individual inputs such as range, outliers, number of data points, rotation setting, or timing across the six months being considered. These averaged scores were the main source of data used by participants to guide their rating choices, with this score viewed as a single data point representing the resident's current sub-competency achievement.

*"Whatever they have gotten on their cumulative evaluations to date is going to determine where they fall on this Milestone."* [Participant 4B]

Some indicated that it was difficult to assign Milestone scores as they had little or no interaction with the residents in patient care settings. Additionally, most participants lacked knowledge of the specifics of the source assessment forms (i.e., what the checklists entailed or the scales that were used) and assumed the faculty assessments utilized Milestone language directly. The reliance on averaged scores may also be influenced by the fact that all of the included programs used a web-based program to collect and collate resident assessment data.

"Comparison to peer average": In determining a specific Milestone rating, participants compared resident "performance," (the average score on completed assessments) to the collective peer average on the same item. Participants then described

overlaying this information onto 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.

*"I usually started people at 1.5* [expected mid-year 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]

***Theme #4 - Moving the Needle: Meaningful Narrative Comments***

Participants described that they incorporated *meaningful* narrative comments, 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 adjust assigned values for a specific sub-competency. They found descriptive comments useful, but also wished for *more* useful comments documenting specific behaviors. Participants reported that some competencies were more likely to draw helpful narratives (e.g. Patient Care, Interpersonal Communication Skills, and Professionalism) while others (e.g. 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 the majority of sub-competencies.

*"...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]

***Theme #5 – No Data? No Problem***

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 assessment data, participants indicated they made the assumption that a resident was performing at the level 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."

*"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]

**COVID-19 Impact**

After the start of the 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 that led 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. Prior studies signaled possible concerns, including works showing straight-line scoring (residents receiving the same value across all Milestones or across all in a single competency) (3, 39). Our work found additional, problematic evidence that CCC members' thought processes 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, and inappropriate use of, Milestones. Regular training of CCC members could better assist member 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" (25) and provides resources for this training, to include quizzes and case studies.

In the early years of Milestone implementation, many CCC members may have received training at educational conferences. However, as Milestones have become more routine, these workshops and presentations have dissipated, leaving newer CCC members

without these training opportunities. 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 (24), and may mitigate the inappropriate reliance upon 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 performance to the Milestones as a whole, not just to the performance of other or 'typical' residents in the program (25)." 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 (36)." 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.

Assigning Milestones values for residents is a difficult task that is time consuming for already busy educators. 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. This reduces the data to a single number, without the richness captured by multiple faculty perspectives or an accurate picture of the resident's growth over time. The ACGME intends for

Milestones to be "narratives, not numbers" (14); 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 sub-competencies (10) or may lack sufficient detail (37), 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 Milestones framework and suggests that the "benchmarking" approach found by Dzara and colleagues persists (13), even after years of experience with Milestones.

We found it striking that participants described assigning a rating 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 (25). Nonetheless, what message is 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 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 improve resident education and assessment. 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. We purposefully selected TY residency programs, which are unique, and recognize they may not fully reflect the response process of members of categorical residency programs. 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 and should be explored in future work.

## CONCLUSION

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

is needed in the areas of training and education as well as best practices to conduct assessment for sub-competencies not captured by current evaluation strategies. If our finding of a response process that departs significantly from the intent of Milestones holds up in larger samplings of CCC members, it complicates the interpretation of Milestones data and threatens the validity of the decisions based upon them.

Table 3. Participant Demographic and Program Information

| Participant | Gender | Specialty | Years in GME Education | Years on CCC | 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 Specialty | 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 |

Table 4. 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)<br><br>"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)<br><br>"I don't know if it's necessarily considered training but the person who was on the CCC last year...he kinda took me through how it works or how he used to do it." (Participant 2B)<br><br>"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 Milestones in assessment of residents or evaluation of the residency program | "...it's to show that they're [residents] progressing towards becoming independent competent providers." (Participant 2B)<br><br>"...I think it offers a standardized approach to evaluating them [residents]" (Participant 3A)<br><br>"...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) |
| Making any Data Fit: The Importance of Average | 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)<br><br>"...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 their average | -How individual assessment inputs are used | "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)<br><br>"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 average | -Use of peer comparison in assessing Milestones | "...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)<br><br>"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) |
| Moving the Needle: | Reference to narrative | "...Patient Care and Communication might be easier because my sense is that they get mentioned a lot on the |

| | | |
|---|---|---|
| Meaningful Narrative Comments | comments in assessment of residents | evaluation comments so we get a lot more information about that...versus something like Practice-Based Learning or System Based that doesn't get mentioned in the comments nearly as much." (Participant 1A)<br><br>"There's always an evaluation comment although it is not always good [useful]." (Participant 1A)<br><br>"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)<br><br>"... they're just going to always say 'Oh, you're great in [specific speciality], we love you!'" (Participant 4A) |
| No Data? No Problem | 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)<br><br>"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) |

# CHAPTER 3: Discussion

## RESPONSE PROCESS

One of the most likely reasons that the response process evidence for Milestones has been poorly studied is that it is perhaps the most complex domain of validity to examine. Response process evidence is an attempt to access thought process(es). And, with Milestones, it includes numerous individuals as well as groups of individuals, further complicating the effort. However, this complexity is also what makes it so important in the overall evaluation of the implementation of Milestones. If the response process being used by those responsible for assigning ratings is not aligned with the ACGME's intent, then the use of Milestones data must be considered in this larger context.

In Heath et al's (23) study of Professionalism and Communication Milestone values for fellows (post-residency trainees), ratings were compared with equivalent Milestones from recently completed residency training, showing learner regression across these domains. The authors argue that it is unlikely that fellows actually lost competency in these areas, but that a score restriction exists in the early phases of a program in order to allow a trainee to demonstrate progression over the time of a program. This exposes CCC assessments as not being truly competency-based, but instead using a requirement that time in training is directly dependent upon attainment of skills. This is also seen in our findings; insomuch as CCC members assume that learners at a particular time point have (or, as in Heath's work, have not) met a specific developmental Milestone, the assessment is not truly competency-based. This was summarized by one of our participants as, *"I know what they want; they want progression."* (Participant 2B)

### Assessment Data

Other findings from our interviews suggested some tension in the artifacts being used by the CCC members in their assessment processes. Members were relying heavily on the rotation assessments and narrative comments provided by faculty members, passing along the ratings without additional analysis or interpretation of the provided data. This indicates that Milestone assessment is actually occurring by individual faculty members with an assumption by the CCC members that individual faculty are appropriately recording their observations of the trainees. There are benefits to allowing front-line supervisors to drive the assessment process, to include gaining a more complete view of a resident's ability through repeated assessments by multiple observers. However, a single Milestone rating is meant to capture resident competency at a given timepoint which requires some interpretation of aggregated data. Should a learner who has demonstrated steady growth in one sub-competency over six months receive a rating that reflects the average of that time? Or, should the rating reflect their current level of performance? These are the interpretations that CCCs need to be making with the data that is available.

### The Transitional Year Internship

Interestingly, the current body of literature on Milestones does not include studies of TY interns nor their training programs. A one year residency program, TY programs train a diverse group of learners who will move on to a myriad of roles following completion, from independent practice to follow on residency training. In comparison to categorical residency programs, which have three or more years for residents to progress

along the Milestone continuum, TY programs have a compressed time period of a single year for this same developmental progression. When considering this short time in training, the Milestones scale, and TY intern diversity, it is probable that a single class of residents will be at different stages of Milestone progression at the prescribed mid-year and end of year reporting times.

TY programs are also unique in that TY CCCs are composed of faculty members across many medical specialties, coming from a myriad of GME disciplines with different Milestones for trainees in their own specialty. As we sought to investigate the response thought process of individual CCC members, TY programs allowed us to sample from a diverse range of backgrounds and experiences in committee membership. This makes the uniformity of responses across our participants that much more striking. Half of our participants also serve on another specialty CCC. While, it is possible that their personal approach to Milestones could differ across their two different CCCs, especially if there truly is a shared mental model within each committee, it does suggest that the themes we identified may persist across other specialties.

### MILITARY RELEVANCE

The military relies heavily on the use of TY internship training to support a number of physician pathways, with military trainees making up more than 10% of all TY residents across the nation, both in civilian and Active Duty programs (personal communication, Cheryl Gross, ACGME). Many military TY graduates will immediately move into advanced residency programs, to include Anesthesiology, Diagnostic Radiology, Ophthalmology, Dermatology, Preventive Medicine, and Physical Medicine

and Rehabilitation. However, where military medicine is unique from our civilian counterparts is in its immediate assignment of TY graduates to General Medical Officer (GMO) positions, where they function as independent, licensed physicians. These GMOs are providing front-line care to our Active Duty members and their dependents; their competency across the defined Milestones, and readiness for independent practice, is critical to the delivery of high-quality medical care and to ensuring our troops remain medically ready for any mission that may arise.

Across the full spectrum of residency and fellowship programs, an improved understanding of how Milestones are used for resident assessment, formative development, and program evaluation will benefit the military medical system and its beneficiaries. The military medical system employs 100% of its graduates as physicians responsible for the care of the beneficiary population, providing incentive to ensure that the assessment tools that we are employing are serving their stated purpose in determining readiness for practice as well as supporting continued professional development. Additionally, since our Active Duty residency programs are supported by a large resource commitment of personnel and funding, the program evaluation aspect of Milestones, which we found to be underutilized, have the potential to identify and assess program improvement initiatives.

Finally, the ACGME is now sharing Milestone data with medical schools, providing aggregated data on graduates and opening up the possibility of Milestones as an evaluative tool at the Undergraduate Medical Education level. This means that Milestones could ultimately serve as a tool for the training provided to our students at the Uniformed Services University (USU). However, before this data can be applied to new

curricular offerings or strategies for learning and assessment, it will be crucial to really understand what it represents. A better understanding of the response process of CCCs and their individual members will provide context for interpreting the performance of USU graduates in residency, and guide targeted improvement in the training of future military medical officers.

## LIMITATIONS

Our work intentionally focused on the response process of an individual. However, decisions made by CCCs in assigning Milestones values are an amalgam of, a) individual-level inputs, b) the interactions of these individuals within the consensus committee, c) the training provided to committee members, d) the sources of data utilized by the members, and e) the processes for approaching the evaluation process. Each of these could impact the way the Milestones are applied to a specific learner and it is important to remember that even a full understanding at the individual level provides only one component of the many overlapping elements involved in Milestone assessment of learners. In fact, our findings suggest that the response process of faculty who complete evaluations is another important aspect of consideration given that our participants relied upon ratings provided by clinical supervisors with little of their own interpretation or synthesis. It is our hope that our study will be the first of many on this topic and that continued research will better elucidate the response process of not only individual CCC members, but all those involved in the application of Milestones in a GME program.

We chose to study members of TY CCCs for many reasons, capitalizing on their diversity of learners, CCC members, and compressed time period in gaining a broad viewpoint on individual response process evidence. However, the elements that make a TY program unique may also make the member response process different from other categorical residency programs. Participants in our study responded to some questions by volunteering comparisons between their TY CCC and specialty of practice CCC, suggesting that there are differences at any given institution. One such difference that may be particularly relevant for TY CCC members is the limited personal interactions with residents in a clinical setting. Prior studies of CCCs have shown that personal interactions with a learner likely impact summative assessment in a way that is difficult to articulate or pinpoint (38). The limited clinical supervisory interactions between TY CCC members and their residents minimizes this aspect of the decision-making process and allowed us to focus on identifying elements where interventions might be made to better meet the goal of competency-based education, such as member training.

Finally, our study was not designed to reach theme saturation and, as such, is limited by a small sample size. At the time of our study design, Department of Defense regulations would not allow for a larger number of participants and we recognize that this work would have been stronger had we been able to interview more CCC members. However, even among the small number of interviewees, the themes we identified were clearly articulated by a majority of participants. These themes accurately reflect the response process of this group and are important to the overall body of validity evidence for the use of Milestones. While there may be other themes that could be discovered with

a larger sample size, those that we list remain relevant, actionable, and critical to the growing understanding of the use of Milestones across residency programs.

### PANDEMIC IMPACT

This study was designed and approved shortly before the world was introduced to the novel coronavirus now known as COVID-19. Initial data collection was delayed from the end of year Milestone assessments (summer 2020) to mid-year assessments (winter 2020-2021) to minimize the impact the pandemic may have had on normal CCC meetings and processes. During this time, many group meetings moved to virtual platforms, which ultimately allowed for observation of participants' CCC meetings - and likely made that observation less obtrusive by minimization of physical presence through turning off video and muting audio. The ability to complete this observation allowed for a better understanding of the milieu of participants and allowed for the use of stimulated recall techniques during interviews.

During interviews, participants were queried about their perception of the possible impact of COVID-19 on both learner assessment and CCC deliberations. Participants acknowledged that learners had less direct interaction with faculty members during calendar year 2020, be it due to personal illness or exposure, re-assignment of residents to alternate clinics or inpatient services, or increased use of virtual encounters and didactic activities. Thus, the pandemic likely did impact individual faculty member's ability to confidently assess learners across a full range of Milestone sub-competencies. Additionally, from a committee standpoint, participants believed that virtual platforms diminished group attention and discussion. This made the "pre-work" and initial

Milestones ratings assigned by each individual CCC member more likely to be ratified by the group. The identified impacts would likely have had more of an impact on direct faculty evaluation and the group decision-making process with less of an effect on the response process of an individual. While the three are likely enmeshed with one another, it is the individual perspective that would have been least directly affected by COVID-19.

### Future Research

As already discussed, validity evidence for response process has not been a focus of Milestone research. This work is the first to use cognitive interviews in an attempt to directly assess the thought process of those involved in the assignment of Milestones rating. Hopefully, future work will build upon our findings using a similar study design. Studies with a larger number participants can determine if there are other relevant themes that need attention. We have used only a single residency specialty; many other residency specialties and fellowships should be explored to determine what similarities and differences exist across different disciplines. Beyond the individual CCC member response process, the collective response process of a CCC needs to be considered and may naturally arise from ongoing work being done by others on group decision-making processes. Additionally, there remains a question about why programs are under-utilizing the "Not Yet Assessable" option on Milestone ratings which may be an interesting topic for future investigation.

### Practical Implications

Despite a limited sample size, this work raises a number of concerns about the use of Milestones in the "real world" and points to a number of practical interventions. CCC

Chairs should review their own policies and procedures to ensure that regular training regarding Milestones is in place for new and existing CCC members. Residency programs should consider where program evaluation opportunities are being missed and find out from their CCC where assumptions are feeding Milestone ratings in order to better identify curricular and assessment areas needing programmatic improvement. Sponsoring institutions as well as national organizations need to be aware of the time that is required of CCC members to appropriately carry out Milestone assessment. Many of the concerns about response process identified in this study have arisen from efforts within programs to increase the efficiency of this assessment process. Medical educators are increasingly asked to do more with less; if Milestones are to work in the reality of a busy residency program, educators need to be appropriately provided with both time and resources. Additionally, as Milestone ratings become available to a wider group of stakeholders, it is important to consider this data in the fuller context of validity evidence. Fellowship programs, other follow-on residency programs, and medical schools should interpret values with caution and not rush to judgements based solely on residency Milestone information.

## CONCLUSIONS

As intended, this work has provided some initial information about the response process used by CCC members in assigning Milestone values to learners. While the findings cast doubt on how well the implementation of Milestones in the "real world" truly aligns with the intentions of the ACGME, they also suggest some practical steps for multiple stakeholders that might help move the GME community closer to the ideal of

competency-based educational assessment. This study has only scratched the surface of response process evidence, laying an initial foundation of knowledge while generating many more questions and potential avenues for future projects.

# APPENDIX

## APPENDIX A. INTERVIEW GUIDE: ACGME MILESTONES IN THE REAL WORLD

Thank you for volunteering to participate in this interview. This is an IRB exempt study being done as part of my thesis project for a Master's degree in Health Professions Education. While my interest in Transitional Year Milestones stems from my own work as both a Program Director and TYRC member, this study is independent of any accrediting or sponsoring institution involved in decisions that affect TY programs or learners. My objective is to learn how Clinical Competency Committee (CCC) members understand and apply the ACGME Milestones in learner assessment. This interview should last between 30 and 60 minutes.

Your participation is completely voluntary and you may stop the interview at any time or decline to answer specific questions. Everything said in this interview will be confidential and anonymous. Nothing said in this discussion will impact you, your residents, your program, or your institution. I am recording our conversation so that it can be accurately transcribed for analysis, but there will be no personal identifiable information stored with these transcripts. Access to the data will be restricted to the members of the research team and not used for any other purpose. Are you willing to continue as a participant in this study?

<u>Participant Demographics</u>

What is your medical specialty of practice?

How long have you been out of GME training?

How long have you been teaching in a GME setting?

How long have you been on a Transitional Year CCC either at your current or prior institutions?

Are you now, or have you ever been, a member of a CCC in another specialty?

Does your hospital have an affiliation with a medical school?

Does your program have a military affiliation?

Committee demographic information

How many members are on your CCC?

How often does your CCC meet? How do they meet? *Probe, as needed, for details about frequency (scheduled and/or ad hoc), format (in-person, tele-conference, etc), and length.*

Do you receive any training specific to your role on the CCC?

If yes: describe this training. *Probe, as needed, for details about timing (one-time or recurrent), length, format, and content.*

When did your CCC last meet?

Keeping this most recent CCC meeting in mind, describe what data inputs were used to determine milestone values for a learner.

Describe the process your committee uses to work through this assessment.

*Probe as needed to get an idea of what data is used and how the process works. Examples of data inputs include rotation evaluations, observed encounters, evaluations from patients and/or peers, exam data, etc. The process will include how those inputs are compiled and used in the group decision. Examples include having one member summarize the data and present it to the group, having each member review the data individually, and reviewing the data together at a meeting.*

If not previously mentioned, does your program ask residents to complete self-assessment using Milestones? If yes, how are these utilized?

Response process

I now want to shift to discussing your thoughts and approach to Milestones as an individual member of the CCC. While this may be influenced by prior trainings and discussions with others both in and out of your CCC meetings, I am interested in your personal opinions.

In your own words, what is the purpose of Milestones?

Can you share your mental model of how a specific learner should progress along the Milestone continuum? *Probe as needed to determine if there connections to time in training as well as if there is expected variability between each Competency and/or individual Milestones.*

These next questions will focus on the specific Milestone of Patient Care 1. Please take a minute to review this Milestone (screen share) before we proceed.

In your own words, what does this Milestone mean to you?

What are the ways in which you could determine the learner's level on this Milestone? *Probe as needed for specific evaluations, assessments, observations, etc. that could inform the decision-making for the individual participant.*

Thinking specifically about your most recent CCC meeting:

-What were you thinking when assigning a level to a learner on this specific Milestone?

-How did you arrive at the assigned level? How confident are you in your determination?

These next questions will focus on the specific Milestone of Systems Based Practice 3.

Please take a minute to review this Milestone (screen share) before we proceed.

In your own words, what does this Milestone mean to you?

What are the ways in which you would determine the learner's level on this Milestone? *Probe as needed for specific evaluations, assessments, observations, etc. that could inform the decision-making for the individual participant.*

Thinking specifically about your most recent CCC meeting:

-What were you thinking when assigning a level to a learner on this specific Milestone?

-How did you arrive at the assigned level? How confident are you in your determination?

How does the format of this Milestone (as compared to the prior one) impact your approach? How does it impact your confidence level? *Probe as needed to address effect of multiple listed tasks in single Milestone.*

Pandemic impact

Describe how the pandemic has impacted assessment of your residents.

How has it changed the deliberations of your CCC?

*This is the end of the interview questions. Ask any relevant clarifying questions before concluding.*

Is there anything else you would like to add? Thank you for your time and participation.

# REFERENCES


1. Aagaard E, Kane GC, Conforti L, Hood, S. 2013. Early feedback on the use of the internal medicine reporting milestones in assessment of resident performance. *J Grad Med Educ*. 5(3):433-8.
2. Andolsek K, Padmore J, Hauer KE, Ekpenyong A. *Clinical Competency Committees: A Guidebook for Programs (3rd Ed).* Accreditation Council for Graduate Medical Education. https://www.acgme.org/Portals/0/ACGMEClinicalCompetencyCommitteeGuidebook.pdf.
3. Beeson MS, Hamstra SJ, Barton MA, Yamazaki K. 2017. Straight Line Scoring by Clinical Competency Committees Using Emergency Medicine Milestones. *J Grad Med Educ*. 9(6):716-720.
4. Beeson MS, Holmboe ES, Korte RC, Nasca TJ. 2015. Initial Validity Analysis of the Emergency Medicine Milestones. *Acad Emerg Med*. 22(7):838-44.
5. Bienstock JL, Shivraj P, Yamazaki K, Connolly AM. 2020. Correlations between Accreditation Council for Graduate Medical Education Obstetrics and Gynecology Milestones and American Board of Obstetrics and Gynecology Qualifying Examination Scores: An Initial Validity Study. *Am J Obstet Gynecol.* 223(3):308.e1-308.e25.
6. Cassaro S, Jarman BT, Joshi ART, Goldman-Mellor S. 2020. Mid-Year Medical Knowledge Milestones and ABSITE Scores in First-Year Surgery Residents. *J Surg Educ*. 77(2):273-280.
7. Chahine S, Cristancho S, Padgett J, Lingard L. 2017. How do Small Groups Make Decisions? *Perspect Med Educ*. 6:192-198

8. Charmaz K. 2010. Grounded Theory: Objectivist and Constructivist Methods. In *Qualitative Educational Research: Readings in Reflexive Methodology and Transformative Practice*. ed W Luttrell, pp. 183-207. New York, NY: Routledge.
9. Cook DA, Kuper A, Hatala R, Ginsburg S. 2016. When Assessment Data Are Words: Validity Evidence for Qualitative Educational Assessments. *Acad Med*. 91(10):1359-1369.
10. Diller D, Cooper S, Jain A, Lam CN. 2019. Which Emergency Medicine Milestone Sub-competencies are Identified Through Narrative Assessments? *West J of Emerg Med*. 21(1):173-179.
11. Doty CI, Roppolo LP, Asher S, Seamon JP. 2015. How do Emergency Medicine Residency Programs Structure their Clinical Competency Committees? A Survey. *Acad Emerg Med*. 22(11):1351-4.
12. Duitsman ME, Fluit CRMG, van Alfen-van der Velden JAEM, de Visser M. 2019. Design and Evaluation of a Clinical Competency Committee. *Perspect Med Educ.* 8: 1-8.
13. Dzara K, Huth K, Kesselheim JC, Schumacher DJ. 2019. Rising to the Challenge: Residency Programs' Experience with Implementing Milestones-Based Assessment. *J Grad Med Educ*. 11(4):439-446.
14. Edgar, L. 2020. *TY Milestones 2.0 - Program Director Panel Discussion.* Presented at Association for Hospital Medical Education Institute, virtual meeting.
15. Edgar L, McLean S, Hogan S, Hamstra SJ, Holmboe E. 2020. *The Milestones Guidebook.* Accreditation Council for Graduate Medical Education. https://www.acgme.org/Portals/0/MilestonesGuidebook/pdf.

16. Francisco GE, Yamazaki K, Raddatz M, Sabharwal S. 2021. Do Milestone Ratings Predict Physical Medicine and Rehabilitation Board Certification Examination Scores? *Am J Phys Med Rehabil* 100:S34-9.
17. Hamstra SJ, Yamazaki K, Barton MA, Santen SA. 2019. 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*. 94(10):1522-1531.
18. Hauer KE, Chesluk B, Iobst W, Holmboe E. 2015. Reviewing Residents' Competence: A Qualitative Study of the Role of Clinical Competency Committees in Performance Assessment. *Acad Med*. 90(8): 1084-92.
19. Hauer KE, Clauser J, Lipner RS, Holmboe ES. 2016. The Internal Medicine Reporting Milestones: Cross-sectional Description of Initial Implementation in U.S. Residency Programs. *Ann Intern Med*. 165(5):356-62.
20. Hauer KE, ten Cate O, Boscardin CK, Iobst W. 2016. 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*. 8(2):156-64.
21. Hauer KE, Vandergrift J, Hess B, Lipner RS. 2016. 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*. 316(22):2253-62.
22. Hauer KE, Vandergrift J, Lipner RS, Holmboe ES. 2018. National Internal Medicine Milestone Ratings: Validity Evidence from Longitudinal Three-Year Follow Up. *Acad Med.* 93(8):1189-1204.

23. Heath JK, Dine CJ. 2019. ACGME Milestones within Subspecialty Training Programs: One Institution's Experience. *J Grad Med Educ*. 11(1):53-59.

24. Hemmer PA, Dadekian GA, Terndrup C, Pangaro LN. 2015. Regular Formal Evaluation Sessions are Effective as Frame-of-Reference Training for Faculty Evaluators of Clerkship Medical Students. *J Gen Intern Med*. 30(9):1313-8.

25. Holmboe E, Edgar L, Hamstra S. 2016. *The Milestones Guidebook.* Accreditation Council for Graduate Medical Education. https://www.acgme.org/Portals/0/MilestonesGuidebook.pdf?ver=2016-05-31-113245-103.

26. Joint Committee on the Standards for Educational and Psychological Testing of the American Educational Research Association, and the National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

27. Kimbrough MK, Thrush CR, Barrett E, Bentley FR. 2018. Are Surgical Milestone Assessments Predictive of In-Training Examination Scores? *J Surg Educ.* 75(1):29-32.

28. Klein R, Ufere NN, Rao SR, Koch J. 2020. Association of Gender With Learner Assessment in Graduate Medical Education. *JAMA Netw Open*. 3(7):e2010888.

29. Korte RC, Beeson MS, Russ CM, Carter WA. 2013. The Emergency Medicine Milestones: A Validation Study. *Acad Emerg Med*. 20(1&):730-5.

30. Kwasny L, Shebrain S, Munene G, Sawyer R. 2021. Is There a Gender Bias in Milestones Evaluations in General Surgery Residency Training? *Am J Surg.* 221(3):505-8.

31. Li ST, Schwartz A, Burke AE, Guralnik S. 2020. Pediatric Program Director Minimum Milestone Expectations Before Allowing Supervision of Others and Unsupervised Practice. *Acad Pediatr*. 20(8):1063-1065.

32. Lozado KN, Ferrandino RM, Teng MS, Colley PM. 2019. Are Otolaryngology Milestones Predictive of Otolaryngology Training Examination Scores? *Ear Nose Throat J*. 98(3):139-42.

33. Nabors C, Forman L, Peterson SJ, Gennarelli M. 2017. Milestones: A Rapid Assessment Method for the Clinical Competency Committee. *Arch Med Sci*. 13(1): 201-209.

34. Padilla JL, Benitez I. 2014. Validity evidence based on response processes. *Psicothema*. 26(1):136-44.

35. Patton MQ. 1987. *How to Use Qualitative Methods in Evaluation (4th Ed.)*. Los Angeles, CA: Sage Publications.

36. Peabody MR, O'Neill TR, Peterson LE. 2017. Examining the Functioning and Reliability of the Family Medicine Milestones. *J Grad Med Educ*. 9(1):46-53.

37. Raaum SE, Lappe K, Colbert-Getz JM, Milne CK. 2019. Milestone Implementation's Impact of Narrative Comments and Perception of Feedback for Internal Medicine Residents: a Mixed Methods Study. *J Gen Intern Med*. 34(6):929-35.

38. Schwartz A, Balmer DF, Borman-Shoap E, Chin A. 2020. Shared Mental Models Among Clinical Competency Committees in the Context of Time-Variable, Competency-Based Advancement to Residency. *Acad Med*. 95(11S):S95-102.

39. Tanaka P, Park YS, Roby J, Ahn K. 2020. Milestone Learning Trajectories of Residents at Five Anesthesiology Residency Programs. *Teach Learn Med*. 1-10.

40. Transitional Year Milestones Work Group. 2019. *Transitional Year Milestones.* Accreditation Council for Graduate Medical Education. https://www.acgme.org/Portals/0/PDFs/Milestones/TransitionalYearMilestones.pdf.

41. Turner TL, Bhavaraju VL, Luciw-Dubas UA, Hicks PJ. 2017. Validity Evidence from Ratings of Pediatric Interns and Subinterns on a Subset of Pediatric Milestones. *Acad Med.* 92(6):809-19.