The Impact of Using Mean Versus Mode When Assessing Resident Competency

Patrick B. Barlow, PhD
Kate DuChene Thoma, MD, MME
Kristi J. Ferguson, PhD

ABSTRACT

Background The Accreditation Council for Graduate Medical Education Milestone Project was implemented in 2014 to standardize assessments and progression of residents. While it is recommended that milestones not be used as tools for direct assessments of resident competency, many programs have used or adapted milestone tools for this purpose.

Objective We sought to explore use of the most frequent milestone level at which a resident was evaluated (ie, the mode), and compared this to the standard practice of using the arithmetic mean for summarizing performance.

Methods We reviewed all Family Medicine Milestone evaluations from 1 program for the first 2 academic years of milestone implementation. Mean and mode scores were calculated across 24 unique residents, 841 evaluation forms, and 5897 measurements. The proportion of overestimation errors (where the mean is at least 0.5 larger than the mode) and underestimation errors (where the mean is at least 0.5 less than the mode) were then compared across resident training year and subcompetency.

Results For the 24 residents, an estimation error occurred in 175 of 792 of the comparisons (22%). Of these errors, 118 (67%) were overestimation errors. First-year residents accounted for 55% (96 of 175) of all estimation errors. All subcompetencies had some estimation errors, with 6 having greater than 5%.

Conclusions If the trend for using the milestones as stand-alone assessment tools is to continue, aggregating data by using frequency distributions and mode would be a more stable and appropriate approach given their nominal or, at best, ordinal nature.

Introduction

The intent of the Accreditation Council for Graduate Medical Education (ACGME) Milestone Project was to standardize assessments and progression of residents focused around 6 competencies (patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice).[1] These were divided into subcompetencies for each specialty, each with a unique set of milestones (FIGURE 1).[1] The ACGME did *not* intend for the milestones to be used as direct evaluation tools. They were intended to be developmental steps toward readiness for unsupervised practice that would be informed by data from direct assessments.[1–4] As milestones were not designed to be stand-alone evaluation instruments, they were not constructed to be a linear progression of a skill, competency, or behavior across the 5 levels. Despite the ACGME's recommendations, many programs and institutions use the milestone rubrics in frontline evaluation instruments, and use the mean values of aggregate data to represent resident performance.

This use of the milestones prompts a closer look at the validity and reliability of the resident performance data, when the milestone rubric instruments are in stand-alone assessments. The purpose of this study was to explore the impact of using the most frequent subcompetency level at which a resident was evaluated (ie, the mode), compared to using the arithmetic average subcompetency level when summarizing performance.

DOI: http://dx.doi.org/10.4300/JGME-D-16-00571.1

Methods

Program Description

Our family medicine residency program is a university program with 6 residents per year. Five core faculty members and a program director comprise the Clinical Competency Committee (CCC).

Our CCC developed new evaluation instruments utilizing the Family Medicine Milestones. Subcompetencies are evenly distributed across all rotations during the 3-year residency training, and assessment calculates a mean score for each subcompetency as its default reporting method. The assessment tools were implemented into MedHub.[5]

Study Design and Sample

We conducted a retrospective review of all resident evaluation data in MedHub that related to the Family Medicine Milestones for the first 2 academic years of

their implementation (2014–2015 and 2015–2016). This consisted of 841 evaluation forms for 24 unique residents that evaluated the 22 family medicine subcompetencies a combined 6417 times. For each form, a resident could be rated on any number of subcompetencies on a scale of 0 (not applicable) to 5 (indicating mastery). Ratings used increments of 0.5. Evaluation forms were organized by subcompetency, score, resident, and residency training year.

Comparing Measures of Central Tendency

The rubrics used to assess milestones in family medicine programs are set to an ordinal scale of 0 (not applicable) to 5. The educational attainment on a specific milestone to move from a 3 to a 4, for example, is *not* a continuous scale. Instead, a resident is placed into 1 of 5 discrete levels, with specific text that may be different from what is expected at the previous or next levels. Thus, the Family Medicine Milestones are set on a *nominal* scale of measurement, which "names" qualities of a specific observation, such as a person's sex, race, or eye color, but does not place them in a numerical relationship with each other. In contrast to numerical scales, which have both magnitude (eg, 3 lb is greater than 2 lb) and exact units (eg, 3 lb is exactly 1 lb greater than 2 lb), nominal data are more meaningfully understood in terms of frequencies and percentages.[6]

To identify a resident's average or typical performance on a given subcompetency, a measure of central tendency is used to summarize many evaluations into a single value. For numerical scales, the most common measure is the arithmetic mean, which is the sum of observed values divided by the number of observations. However, it is important to understand that milestone data are nominal or, at best, ordinal, and thus computing a mean subcompetency level is equivalent to computing a resident's mean eye color. The mode, which denotes the *most frequent* level across all evaluations, may be a more appropriate measure of central tendency for nominal data.[6]

Using the mean to summarize discrete data can have a dramatic impact on how performance is calculated. For example, if a resident was evaluated 10 times on the professionalism 1 subcompetency, and receives the distribution of scores in TABLE 1, then the mean score would be 3.0, yet the resident had been most frequently placed in level 2—a full scale point below. The first annual ACGME report summarizing national milestone achievements reinforces this warning: "The mean rating . . . should be interpreted with caution given milestones are ordinal, not dimensional data."[7(p15)]

What was known and gap

Many programs use the milestone rubrics as a tool for direct assessment of residents at the bedside and in clinic.

What is new

This study evaluated the accuracy of milestone valuations of family medicine residents, comparing assessments based on mean versus modal values.

Limitations

Single institution and single specialty study design reduce generalizability.

Bottom line

Use of the modal value is thought to provide more robust and defensible milestone assessment of trainees.

The University of Iowa Institutional Review Board reviewed this study and determined it to be exempt.

Statistical Analysis

We calculated mean scores for each resident for the 22 subcompetencies to produce the "typical" evaluation data used by programs in determining performance. The calculation did *not* include the zeros (ie, not applicable). The most frequent (ie, mode) score for the same 22 subcompetencies was then calculated to compare the difference between both measures of central tendency. To make the comparison as equal as possible, zeros were also *not* included when calculating the mode. In the event a score distribution on a subcompetency was bimodal (ie, exactly 2 modes), the *lower* of the 2 values was chosen, as it represents the more conservative estimate of performance. In the event a score distribution on a subcompetency contained *greater than* 2 modes, that subcompetency was excluded from the analysis. Absolute differences between the mean and mode that were equal to or greater than 0.5 were considered practically meaningful and were coded as an *estimation error*. Finally, the frequency and proportion of times in which the mean *overestimated* or *underestimated* mode performance were tabulated for each resident class (postgraduate year 1 [PGY-1] to PGY-3).

The measurements in which the score was a zero (ie, not applicable) were filtered from the total sample to keep our mean and mode calculations consistent with the way MedHub currently calculates average ratings.

Results

Data Cleaning

Our calculations produced a final data set containing 24 unique residents, 841 forms, and 5897 measurements. Initial mean and mode values were then computed for a total of 792 subcompetencies (22 ×

PATIENT CARE

General Competency

Family physicians provide accessible, quality, comprehensive, compassionate, continuous, and coordinated care to patients in the context of family and community, not limited by age, gender, disease process, or clinical setting, and by using the biopsychosocial perspective and patient-centered model of care.

Subcompetency

PC-1 Cares for acutely ill or injured patients in urgent and emergent situations and in all settings					
Has not achieved Level 1	**Level 1**	**Level 2**	**Level 3**	**Level 4**	**Level 5**
	Gathers essential information about the patient (history, examination, diagnostic testing, psychosocial context) Generates differential diagnoses Recognizes role of clinical protocols and guidelines in acute situations	Consistently recognizes common situations that require urgent or emergent medical care Stabilizes the acutely ill patient utilizing appropriate clinical protocols and guidelines Generates appropriate differential diagnoses for any presenting complaint Develops appropriate diagnostic and therapeutic management plans for acute conditions	Consistently recognizes complex situations requiring urgent or emergent medical care Appropriately prioritizes the response to the acutely ill patient Develops appropriate diagnostic and therapeutic management plans for less common acute conditions Addresses the psychosocial implications of acute illness on patients and families Arranges appropriate transitions of care	Coordinates care of acutely ill patient with consultants and community services Demonstrates awareness of personal limitations regarding procedures, knowledge, and experience in the care of acutely ill patients	Provides and coordinates care for acutely ill patients within local and regional systems of care
☐	☐	☐	☐	☐	☐
Comments:					

Specific Milestone

FIGURE 1
Example Milestone Rubric Showing the Components of a Core Competency, Subcompetency, and Milestone

TABLE 1
Example Distribution of a Resident's Milestone Scores for Professionalism 1

Example Scores										Mean Score	Mode Score
1	2	2	2	3	3	3.5	5	5	4	3.0	2.0

18 residents per year) to identify bimodal and multimodal cases. A total of 164 of 792 of the distributions (21%) were bimodal, and 88 of 792 (11%) were excluded for being multimodal.

Estimation Errors by Class

TABLE 2 displays the results of our primary comparisons between resident mean and mode scores for each residency year. For each academic year, the mean and mode scores from each resident's 22 subcompetencies were compared for estimation errors. Across all residents, there were 175 of 792 subcompetency estimation errors (22%), in which the mean either overestimated or underestimated the mode by at least 0.5. Of these errors, 118 (67%) were considered overestimation errors. Both types of errors were most frequently seen in the PGY-1 residents, with a combined 55% (96 of 175) of all errors occurring in this class (62 overestimates and 34 underestimates). The fewest errors were seen in PGY-3 residents. When the mean was *higher* than the mode by at least 0.5, the degree of difference was highest among PGY-1 residents who had overestimation errors of 0.77 (SD = 0.27), a difference of more than three-fourths of a milestone level.

Subcompetency data were compiled across both years to look for subcompetencies especially prone to overestimation or underestimation errors (TABLE 3). While the proportion of estimation errors tended to decrease the more frequently a subcompetency was measured, every subcompetency had *some* estimation errors. There were 6 subcompetencies that had total estimation errors greater than 5%, including 3 of the 4 professionalism subcompetencies (ICS-4, PROF-4, MK-1, PROF-3, SBP-3, and PROF-1). Thirteen subcompetencies had between 2% and 5% total errors, and ICS-3, PBLI-1, and MK-2 each had less than 2% errors.

We also saw a link between the way in which the milestones were written and the prevalence of estimation errors. TABLE 4 illustrates the milestones for 2 subcompetencies (ICS-3 and ICS-4), which had a very low and a very high prevalence of estimation errors, respectively. The milestones within ICS-3 can be more easily recognized as ordinal progressions across different levels, which could have led to fewer errors. In contrast, the progression from level 1 to level 5 may be ordinal for ICS-4; however, the individual milestone statements within each level are not continuations of one another. If a resident's mode was 2.0, but a mean of 3.0 was reported as his or her average performance, then the most typically achieved milestones would denote a different level of skill.

Estimation Errors and Their Impact on Resident Performance Summaries

While quantifying the prevalence of estimation errors is important on a global level, the impact these errors may have in accurately representing a resident's performance brings this measurement issue to a practical context. We chose 3 examples (1 from each resident class) to represent how the mode as a measure of central tendency, when compared to the mean, provides a different depiction of overall performance and the milestones that were achieved. FIGURE 2 displays these 3 cases where the resident's mean score for each subcompetency (circle) is compared to their mode for that subcompetency (square). The same pattern that was seen with the aggregate data reveals itself in these individual cases. Use of the mean score for these *discrete* milestone levels resulted in a dramatic *overestimation* of performance among first- and second-year residents, and was more likely to *underestimate* performance as the resident reached the final year of training.

TABLE 2
Number of Estimation Errors by Type of Error and Resident Training Year

Type of Estimation Error and Resident Training Year	No. (%) of Estimation Errors
Overestimation errors	
PGY-1	62 (53)
PGY-2	37 (31)
PGY-3	19 (16)
Total	118 (100)
Underestimation errors	
PGY-1	34 (60)
PGY-2	12 (21)
PGY-3	11 (19)
Total	57 (100)

Abbreviation: PGY, postgraduate year.

TABLE 3
Proportion of Estimation Errors by Subcompetency Rank Ordered by Total Percentage Errors

Subcompetency	Total Measurements	Percentage Over	Percentage Under	Total Percentage Errors[a]
ICS-4	77	9.1	0	9.1
PROF-4	68	8.8	0	8.8
MK-1	111	6.3	0.9	7.2
PROF-3	164	5.5	0.6	6.1
SBP-3	82	6.1	0	6.1
PROF-1	118	5.1	0.9	5.9
SBP-1	133	2.3	2.3	4.5
PBLI-3	217	2.8	0.9	3.7
PBLI-2	308	2.3	1.3	3.6
PC-3	237	1.7	1.7	3.4
PC-4	208	3.4	0	3.4
PC-5	215	2.8	0.5	3.3
PC-2	245	2.0	0.8	2.9
SBP-4	528	1.1	1.5	2.7
ICS-2	227	2.2	0.4	2.6
PC-1	513	0.8	1.4	2.1
SBP-2	375	1.6	0.5	2.1
PROF-2	611	0.7	1.3	2
ICS-1	409	0.7	1.2	2
ICS-3	402	1.0	0.8	1.7
PBLI-1	290	1.4	0.3	1.7
MK-2	432	0.9	0.7	1.6

Abbreviations: ICS, interpersonal and communication skills; PROF, professionalism; MK, medical knowledge; SBP, systems-based practice; PBLI, practice-based learning and improvement; PC, patient care.
[a] Percentages are calculated as (number of estimation errors)/(total number of measurements for a subcompetency).

Discussion

Our results suggest that the ACGME's original intent regarding milestones (ie, that they should not be used as scales to directly assess resident competency) was a wise recommendation. Unfortunately, in many programs, milestone assessments do not follow this recommendation. Residency programs often have taken the subcompetencies, converted them to a 5-point scale with intermediate points on the continuum, and have raters place residents on that scale based on a single observation or limited time spent with the ratee. These scores are then averaged across ratings using a mean score, whereby the CCCs then use this information to create its reports.

According to our study, the use of mean performance on scales developed using these developmental categories as the scale points is ill-advised and unpredictably inaccurate. The mean score either overestimates where residents are on a subcompetency, or underestimates their level of performance almost 25% of the time when compared with the level at which they were most frequently scored (ie, mode). Presenting a CCC with a frequency distribution and mode level for each resident's performance in each subcompetency offers a more holistic view of the resident's competence.

The main limitation of this study is that it is from a single residency program, reducing generalizability. The study data were gathered during the first 2 years of implementing the milestone evaluations, so there may have been a more variable set of measurements as evaluators adjusted to the new system. Another potential limitation to the study was excluding 11% of the data for being multimodal, and selecting the *lower* of the 2 modes when a score distribution was bimodal. These decisions may have had a greater or lesser impact on error proportions for certain subcompetencies. Further research is needed to determine the benefits of using modal values for direct assessments of trainees using the milestones.

TABLE 4
Representative Example of Subcompetencies with Low and High Error Frequencies

Example	Level 1	Level 2	Level 3	Level 4	Level 5
Low error frequency (1.7%)					
ICS-3: Develops relationships and effectively communicates with physicians, other health professionals, and health care teams.	Understands the importance of the health care team and shows respect for the skills and contributions of others.	Demonstrates consultative exchange that includes clear expectations and timely, appropriate exchange of information. Presents and documents patient data in a clear, concise, and organized manner.	Effectively uses EHR to exchange information among members of the health care team. Communicates collaboratively with the health care team members by listening attentively, sharing information, and giving and receiving constructive feedback.	Sustains collaborative working relationships during complex and challenging situations, including transitions of care. Effectively negotiates and manages conflict among members of the health care team in the best interest of the patient.	Role models effective collaboration with other providers that emphasizes efficient patient-centered care.
High error frequency (9.1%)					
ICS-4: Utilizes technology to optimize communication.	Recognizes effects of technology on information exchange and the physician-patient relationship. Recognizes the ethical and legal implications of using technology to communicate in health care.	Ensures that clinical and administrative documentation is timely, complete, and accurate. Maintains key patient-specific databases, such as problem lists, medications, health maintenance, and chronic disease registries. Uses technology in a manner which enhances communication and does not interfere with the appropriate interaction with the patient.	Ensures transitions of care are accurately documented, and optimizes communication across systems and continuums of care.	Effectively and ethically uses all forms of communication, such as face-to-face, telephonic, electronic, and social media. Uses technology to optimize continuity of care of patients and transitions of care.	Stays current with technology and adapts systems to improve communication with patients, other providers, and systems.

Abbreviations: ICS, interpersonal and communication skills; EHR, electronic health record.

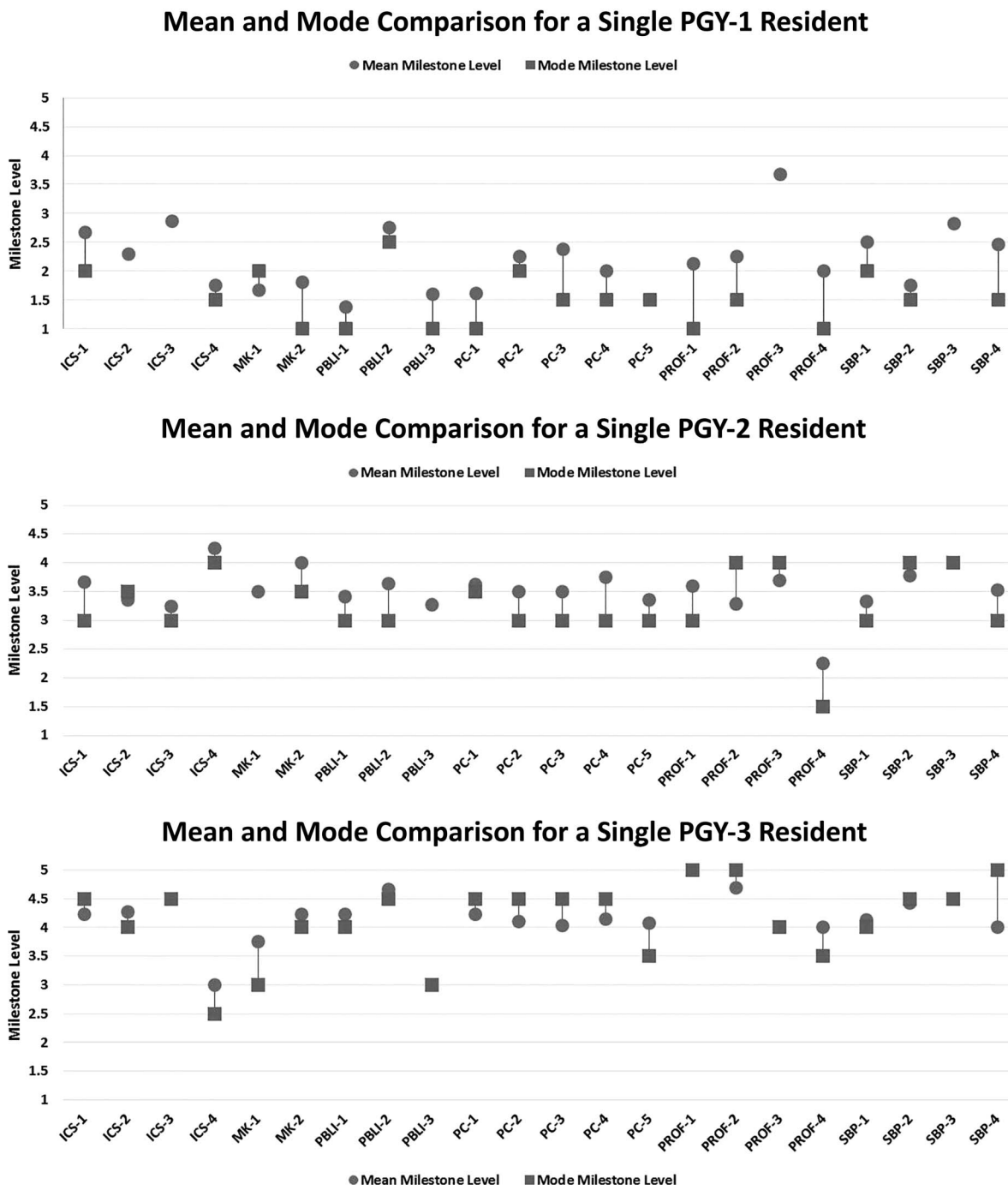

FIGURE 2
Example Cases of a Single Resident's Mean Versus Mode Summary for Each Subcompetency

Conclusion

If the current trend for using the milestone rubrics as stand-alone instruments for direct assessments is to continue, aggregating the data using a combination of frequency distributions and modal values would produce a more stable and level-appropriate approach, given the nominal or, at best, ordinal nature of the milestones. As programs struggle to develop and implement new assessment instruments for nationally developed milestones, it will be important to identify measures that are psychometrically sound as well as theory based.

References

1. The family medicine milestone project. *J Grad Med Educ*. 2014;6(1 suppl 1):74–86.
2. Allen S. Development of the family medicine milestones. *J Grad Med Educ*. 2014;6(1 suppl 1):71–73.
3. Holmboe ES, Yamazaki K, Edgar L, et al. Reflections on the first 2 years of milestone implementation. *J Grad Med Educ*. 2015;7(3):506–511.
4. Holmboe ES, Edgar L, Hamstra S. The milestones guidebook: version 2016. http://www.acgme.org/Portals/0/MilestonesGuidebook.pdf. Accessed January 25, 2017.
5. MedHub. Graduate medical education management, designed for the GME enterprise from day one. http://www.medhub.com/wp-content/uploads/2016/11/MedHub_GME2.pdf. Accessed March 15, 2017.
6. Dawson B, Trapp RG. *Basic & Clinical Biostatistics*. New York, NY: Lange Medical Books-McGraw-Hill; 2004.
7. Hamstra SJ, Edgar L, Yamazaki K, et al. Milestones annual report 2016. http://www.acgme.org/Portals/0/PDFs/Milestones/MilestonesAnnualReport2016.pdf. Accessed March 15, 2017.

All authors are at Carver College of Medicine, University of Iowa. **Patrick B. Barlow, PhD,** is Assistant Professor, General Internal Medicine, and Program Evaluation Consultant, Office of Consultation and Research in Medical Education; **Kate DuChene Thoma, MD, MME,** is Associate Clinical Professor and Residency Program Director, Department of Family Medicine; and **Kristi J. Ferguson, PhD,** is Professor, General Internal Medicine, and Director, Office of Consultation & Research in Medical Education.

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

Conflict of interest: The authors declare they have no competing interests.

Corresponding author: Patrick B. Barlow, PhD, University of Iowa, Carver College of Medicine, Office of Consultation and Research in Medical Education, 1204 Medical Education Building, Iowa City, IA 52242, 319.384.4291, patrick-barlow@uiowa.edu

Received September 7, 2016; revision received December 22, 2016; accepted January 23, 2017.