Use of Emergency Medicine Milestones as Items on End-of-Shift Evaluations Results in Overestimates of Residents' Proficiency Level

Erin Dehon, PhD
Jonathan Jones, MD
Michael Puskarich, MD
John Petty Sandifer, MD
Kristina Sikes, MD

Abstract

Background The emergency medicine milestones were developed to provide more objective resident assessment than current methods. However, little is known about the best practices for applying the milestones in resident assessment.

Objective We examined the utility of end-of-shift evaluations (ESEs) constructed using the milestones in resident assessment.

Methods We developed 14 daily ESEs, each of which included 9 or 10 emergency medicine milestones. Postgraduate year (PGY)-1 and PGY-2 residents were assessed on milestone levels 1 through 3; PGY-3 and PGY-4 residents were assessed on levels 3 through 5. Each milestone was rated on a nominal scale (*yes, no,* or *not applicable*). The Clinical Competency Committee combined the ESE data with data from other assessments to determine each resident's proficiency level for the emergency medicine subcompetencies. We used descriptive statistics to summarize resident ESEs and milestone levels. We analyzed differences in ESE score across PGY levels using *t* tests and analyses of variance.

Results Faculty completed 763 ESEs on 33 residents with a range of 2 to 54 (median = 22) ESEs per resident. Faculty rarely (8%, 372 of 4633) rated a resident as not achieving a milestone on the ESEs. Analyses of variance revealed that ESE scores on level 3 milestones did not differ significantly by PGY level. There was poor agreement between ESE scores and Clinical Competency Committee ratings.

Conclusions The ESEs constructed using the milestones resulted in grade or milestone inflation. Our results do not support using milestones as a stand-alone assessment tool.

Editor's Note: The online version of this article contains a table of response data.

Introduction

In 2013, the Accreditation Council for Graduate Medical Education (ACGME) implemented the Next Accreditation System, which introduced subcompetencies and the milestones.[1] The Review Committee for Emergency Medicine and the American Board of Emergency Medicine identified 227 emergency medicine (EM) milestones that describe a resident's progression from novice to expert. Each milestone is assigned to 1 of 23 subcompetency areas, and the milestones are associated with a specific developmental stage in resident competency. Residents are expected to reach proficiency (level 4 of the 5-level trajectory) by the end of training.[2,3]

The ACGME does not specify how programs should assess residents on each milestone. Several programs have incorporated the EM milestones directly into resident assessment, yet there are little data on their reliability, validity, and general use by programs. A recent study found that, when using an assessment tool based on the milestones, 100% of interns met level 1 milestones; yet, when using a standardized observation tool those same interns were competent in only 48% to 93% of the level 1 milestones.[4]

To date, no studies have examined the use of milestone-based evaluation tools in end-of-shift evaluations (ESEs),

All authors are at the Department of Emergency Medicine, University of Mississippi Medical Center. **Erin Dehon, PhD,** is Assistant Professor and Director of Faculty Development and Behavioral Science Education; **Jonathan Jones, MD,** is Associate Professor and Residency Program Director; **Michael Puskarich, MD,** is Assistant Professor and Research Director; **John Petty Sandifer, MD,** is Assistant Professor and Assistant Program Director; and **Kristina Sikes, MD,** is Emergency Medicine Resident.

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

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

This manuscript was presented at the Society of Academic Emergency Medicine Annual Meeting in May 2014 in Dallas, Texas, and at the Council of Emergency Medicine Residency Directors Annual Meeting in April 2014 in New Orleans, Louisiana.

Corresponding author: Erin Dehon, PhD, Department of Emergency Medicine, University of Mississippi Medical Center, 2500 N State Street, Jackson, MS 39216, 504.710.5368, edehon@umc.edu

Received August 1, 2014; revisions received December 4, 2014, and January 2, 2015; accepted January 12, 2015.

DOI: http://dx.doi.org/10.4300/JGME-D-14-00438.1

also known as *shift cards*. The purpose of our study was to determine the usefulness of ESEs based on the milestones and to collect construct validity evidence to guide inferences made from using milestone-based ESEs.

What was known and gap

Use of milestones is thought to allow for more meaningful assessment, yet little is known about their utility in direct evaluation.

What is new

Use of milestones in end-of-shift evaluations of emergency medicine residents showed inflated faculty ratings and poor agreement between these evaluations and Clinical Competency Committee ratings.

Limitations

Scores may have been influenced by the nonanonymous approach to assessment; the single site study limits generalizability.

Bottom line

Milestone-based assessments may not be appropriate as a stand-alone tool.

Methods

Study Setting and Population

This was a retrospective analysis of data from 33 EM residents assessed during a 4-month data collection period (July through October 2013). All EM residents assessed were included. Assessments were completed by 19 EM faculty members.

ESE Tool Development

Using only the EM milestones, the education division developed 14 ESEs (2 sets of 7), 1 for each day of the week. As level 4 and level 5 milestones correspond to functioning near an attending level or above, we assumed that residents in postgraduate year (PGY)-1 or PGY-2 had not achieved these levels, and that PGY-3 and PGY-4 residents had already achieved level 1 and level 2 milestones. Therefore, set 1 of the ESEs included only milestone levels 1 through 3 and was used for PGY-1 and PGY-2 residents, while set 2 only included milestone levels 3 through 5 and was used for PGY-3 and PGY-4 residents. The ESEs contained a total of 87 milestones from 12 of the 23 EM subcompetencies. Specific subcompetencies were chosen based on educators' perception that they were well-suited for assessment during a shift. Each ESE evaluated 9 or 10 milestones. The distribution of milestones included on the ESEs by level for set 1 was level 1, 22%; level 2, 30%; and level 3, 48%. For set 2, the distribution of milestones was level 3, 55%; level 4, 36%; and level 5, 9%. All 87 milestones included are provided as online supplemental material.

TABLE 1 lists the 12 subcompetencies assessed with the ESEs. The 6 EM procedural subcompetencies, as well as medical knowledge, accountability, patient safety, and system-based management were not assessed because they were considered to be too difficult to directly observe in a single shift or were assessed using other methods. On each ESE, the milestones are listed verbatim (eg, recognizes abnormal vital signs) and rated on a nominal scale. The FIGURE provides a representative ESE.

The 14 ESEs were entered into E*Value (Advanced Informatics). Residents were responsible for sending the designated ESE to their attending for each shift. Thus, evaluations were not anonymous, since the attending physician's identity was known to the resident. Faculty have been known to base ratings on the resident's PGY level;[3] therefore, faculty were blinded to the item's milestone level. Faculty were instructed to skip an item if it referred to a behavior that was not observed. Finally, faculty were asked to provide written recommendations for residents using a comment box (FIGURE).

Clinical Competency Committee Ratings

The ACGME requires residency programs to form a Clinical Competency Committee (CCC) to assess residents semiannually based on the EM milestones. Our CCC used ESE data as well as other data (eg, off-service rotation evaluations, procedure evaluations, test scores, unsolicited comments) to determine milestone achievement levels for each resident. The residents were assessed on each of the 23 subcompentencies from levels 1 (novice) through 5 (expert).

TABLE 1 **Total Percentage of *Yes* to *No* (*Yes* / [*Yes* + *No*]) Responses on End-of-Shift Evaluations by Subcompetency Area**

Subcompetency Area	%
Professional values	96
Emergency stabilization	95
Observation and reassessment	94
Disposition	94
Multitasking/task switching	93
Diagnosis	93
Patient-centered communication	93
Team management	93
Technology	92
Focused history and physical	91
Diagnostic studies	91
Pharmacology	86

Item	Yes/No/N/A[a]	Milestone Level[b]	Subcompetency Evaluated[b]
Performs and communicates a reliable, comprehensive history and physical examination	Yes/No/N/A	1	Focused History and Physical (PC2)
Performs appropriate bedside diagnostic studies and procedures	Yes/No/N/A	2	Diagnostic Studies (PC3)
Participates as a member of a patient care team	Yes/No/N/A	1	Team Management (ICS2)
Constructs a list of potential diagnoses based on chief complaint and initial assessment	Yes/No/N/A	1	Diagnosis (PC4)
Task switches between different patients	Yes/No/N/A	2	Task Switching (PC8)
Ensures that medical records are complete, with attention to preventing confusion and error	Yes/No/N/A	2	Technology (SBP3)
Applies medical knowledge for selection of appropriate agent for therapeutic intervention	Yes/No/N/A	2	Pharmacotherapy (PC5)
Prioritizes critical initial stabilization actions in the resuscitation of a critically ill or injured patient	Yes/No/N/A	3	Emergency Stabilization (PC1)
Manages the expectations of those who receive care in the ED and uses communication methods that minimize the potential for stress, conflict, and misunderstanding	Yes/No/N/A	3	Patient-Centered Communication (ICS1)
Performs and communicates a reliable, comprehensive history and physical examination	Yes/No/N/A	1	Focused History and Physical (PC2)
Recommendations:			

FIGURE **Sample Electronic End-of-Shift Evaluations for Residents in Postgraduate Year (PGY)-1 and PGY-2**

Abbreviations: ESE, end-of-shift evaluation; PGY, postgraduate year; PC, patient care; SBP, systems-based practice; ICS, interpersonal and communication skills; ED, emergency department.

[a] N/A is depicted here to note that not answering an item was an option, but it was not actually listed an option to select.

[b] The milestone level and subcompetency evaluated were not visible to the evaluator.

The study was considered exempt from review by the University of Mississippi Medical Center Institutional Review Board.

Data Analysis

Descriptive statistics were used to summarize resident ESEs and milestone levels. Each resident was assigned a mean ESE score for each milestone and subcompetency. The ESE scores were calculated as percentages of total *yes* to *no* responses. Secondary post hoc analyses were conducted using ESE scores; *N/A* (*not applicable*) responses were recoded as *no*.

The mean CCC score was calculated by averaging the residents' milestone achievement level on the subcompetencies included in the ESEs. Similar to other studies, to collect construct validity evidence we examined the ability of the ESEs to differentiate between resident levels of training by comparing performances of residents by PGY level using analyses of variance (ANOVAs) and *t* tests.[5]

Results

Over a period of 4 months, 19 faculty members completed 763 ESEs on 33 residents. The number of ESEs completed per resident varied from 2 to 54 (median = 22). The number of ESEs completed by faculty members ranged from 1 to 68 (median = 28). The median percentage of shift cards completed/number of resident shifts was 87%. Of the 5091 total responses, 84% (4261) were *yes*, 7% (372) were *no*, and 9% (458) were *N/A*.

An analysis of *yes* to *no* responses revealed that faculty rarely (8%, 372 of 4633) rated a resident as not achieving a milestone (TABLE 1). The median percentage of *yes* to *no* responses for all 87 milestones is provided as online supplemental material. The median percentage of *yes* to *no* responses for 84 of 87 milestones was 100%. It is worth noting that 5 residents (including 2 in PGY-1) did not receive a *no* response on any of the ESEs.

TABLE 2 shows the mean ESE scores and mean CCC scores (SD) by PGY level. ESE mean scores for lower-level milestones (ie, 1 through 3) did not differ significantly by PGY level, while mean ESE scores on level 4 milestones were significantly higher for PGY-4s than for PGY-3s. Mean CCC score was consistent with PGY level. A 1-way ANOVA revealed that the mean CCC scores differed significantly ($P < .001$) by PGY level; post hoc analyses using Tukey tests showed that all the groups were significantly different from one another ($P < .001$).

TABLE 3 displays the results of the secondary post hoc analysis in which a new ESE score was calculated using the *N/A* items. The *N/A* responses were counted the same as *no* responses. A 1-way ANOVA revealed that the ESE score differed significantly by PGY level, with upper-level residents receiving higher scores when the unanswered items were counted as *no* responses.

Discussion

Our examination of the utility of ESEs that directly incorporate milestones found that faculty often (92%, 4261 of 4633) rated residents as achieving milestones irrespective of the resident's level of training or the level of the milestone. For instance, residents in PGY-3 and PGY-4 were often rated as achieving level 5 milestones, a level reserved for physicians with expertise in a particular area. Our findings support the perspective of Carter,[6] who recently highlighted several myths and misperceptions about the milestones. He emphasized the limitations of using the milestones as direct assessment tools, including noting that milestones are "as vulnerable to grade inflation as other tools . . . because it is nearly impossible for a single observation to yield enough information to accurately assess each of the level anchors."[6]

In addition to finding inflated scores, we were unable to demonstrate that ESEs discriminated between residents at different levels of training. Although milestone levels are

TABLE 2 **Mean End-of-Shift Evaluation and Clinical Competency Committee (CCC) Scores by Postgraduate Year (PGY) and Milestone Level**

Milestone Level	PGY-1 (n = 6)	PGY-2 (n = 9)	PGY-3 (n = 9)	PGY-4 (n = 9)	Statistic	*P* Value
1	94	96	. . .	. . .	T = 0.65	.31
2	88	94	. . .	. . .	T = 1.30	.06
3	88	91	93	98	F = 2.40	.09
4	. . .	. . .	87	96	T = 2.21	.04[a]
5	. . .	. . .	78	92	T = 1.70	.12
CCC Score (SD)	1.02 (0.04)	1.91 (0.39)	3.01 (0.24)	4.00 (0.06)	F = 276.8	.001

[a] Statistically significant at $P < .05$.

not intended to directly correspond to level of training, one would expect that, on average, residents in PGY-1 and those in PGY-4 would differ significantly in terms of overall proficiency. Nonetheless, we were unable to find a difference in ESE scores on level 3 milestones across residents in PGY-1 through those in PGY-4. Furthermore, the overall mean scores obtained from the ESEs were not in agreement with the ratings determined by the CCC during the residents' semiannual evaluations. The CCC was aware of each resident's PGY level, which may have biased the scores.[3] Nonetheless, findings were consistent with a study that found that EM interns were often rated as proficient when using other milestone-based evaluation tools.[4]

The observed overestimates are a direct reflection of a lack of *no* responses by faculty. In designing our measure, we considered the tendency of faculty to rate residents based on their PGY (versus ability) level; therefore, we chose a nominal scale and blinded raters to milestone levels.[4] Rather than improve objectivity, however, this approach resulted in inflated scores for residents across all PGY levels. Several faculty reported anecdotally that they did not appreciate the all-or-nothing (*yes*, *no*, or *N/A*) method of assessment; they "felt bad" responding *no* to an item. In addition, residents frequently sought explanations for any *no* response assigned to them, even though they had been notified to expect *no* ratings. This could also have been a factor in the elevated scores.

To explore the hypothesis that faculty may have failed to follow instructions and skipped items because they were reluctant to answer *no*, we conducted a secondary analysis in which we recoded all *N/A* items as *no* responses. After recoding, ESE scores did differ significantly by PGY level. However, it is worth noting that the percentage of responses was still overwhelmingly *yes*, even across milestone levels above the PGY training level. Because we can only speculate about alternative explanations as to what the *N/A* items imply, we recommend interpreting the *N/A* items as intended (ie, descriptions of behaviors not observed during a shift). It is certainly feasible that some milestones (eg, identify rare patient conditions based solely

TABLE 3 **Mean End-of-Shift Evaluation (With *N/A* [*Not Applicable*] Responses Coded as *No*) and Clinical Competency Committee Scores by Postgraduate (PGY) and Milestone Level With *N/A* + *No***

Milestone Level	PGY-1 (n = 6)	PGY-2 (n = 9)	PGY-3 (n = 9)	PGY-4 (n = 9)	Statistic	*P* Value
1	83	84	. . .	. . .	T = 0.76	.32
2	78	87	. . .	. . .	T = 2.41	.03[a]
3	73	81	87	90	F = 6.18	< .01[a]
4	. . .	. . .	74	89	T = 2.78	.01[a]
5	. . .	. . .	69	89	T = 2.92	.01

[a] Statistically significant at $P < .05$.

on the history and physical) may not apply to a shift and therefore an *N/A* response is most appropriate.

Our finding that nominal ratings are poor at discriminating between levels of training is consistent with the findings of previous studies. Compared with nominal scales, global rating scales (often utilizing a Likert scale) have been shown to better discriminate between different levels of training and have superior interrater reliability.[4,7,8] Furthermore, the consistency and accuracy of subjective global rating scores can be significantly improved with examiner training.[8] Given our findings and those of previous studies, a better assessment tool may be one that uses global ratings not based on PGY level and incorporates extensive faculty training.

Several limitations should be considered. First, ESE scores could have been influenced by the nonanonymous assessment format, and an anonymous format may have resulted in a wider distribution of scores (ie, less inflated). Second, the ESEs did not assess all residents on milestone levels 1 through 5, even though it is possible that some junior residents may have achieved advanced milestones and some senior residents may be at the novice level for certain milestones. Third, data were collected from a single institution, which limits generalizability. Fourth, although the median ESE completion rate was relatively high, some residents may have chosen not to assign an evaluation after a particularly bad shift, which may have skewed the results. Finally, there remains a lack of objective data supporting the appropriateness of the levels assigned to each EM milestone. Further research is needed to examine whether the milestones truly differentiate between learners at different levels of training.

Conclusion

Our findings highlight the ongoing challenge of incorporating the EM milestones into resident assessment. Although there remains a need for feasible instruments that can accurately assess learner performance in order to inform CCC competency decisions, we do not find evidence to support the direct incorporation of milestones into shift evaluations in the emergency department.

References

1 Nasca TJ, Philibert I, Brigham T, Flynn TC. The next accreditation system rationale and benefits. *N Engl J Med*. 2012;366(11):1051–1056.
2 Korte RC, Beeson MS, Russ CM, Carter WA, Emergency Medicine Milestones Working Group, Reisdorff EJ. The emergency medicine milestones: a validation study. *Acad Emerg Med*. 2013;20(7):730–735.
3 Beeson MS, Carter WA, Christopher TA, Heidt JW, Jones JH, Meyer LE, et al. The development of the emergency medicine milestones. *Acad Emerg Med*. 2013;20(7):724–729.
4 Hauff SR, Hopson LR, Losman E, Perry MA, Lypson ML, Fischer J, et al. Programmatic assessment of level 1 milestones in incoming interns. *Acad Emerg Med*. 2013;21(6):694–698.
5 Panait L, Bell RL, Roberts KE, Duffy AJ. Designing and validating a customized virtual reality-based laparoscopic skills curriculum. *J Surg Educ*. 2008;65(6):413–417.
6 Carter WA. Milestone myths and perceptions. *J Grad Med Educ*. 2014;6(1):18–20.
7 Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Med Educ*. 2003;37(11):1012–1016.
8 Malau-Aduli BS, Mulcahy S, Warnecke E, Otahal P, Teague PA, Turner R, et al. Inter-rater reliability: comparison of checklist and global scoring for OSCEs. *Creative Educ J*. 2012;3:937–942.