Milestone-Based Assessments Are Superior to Likert-Type Assessments in Illustrating Trainee Progression

Kathleen W. Bartlett, MD
Shari A. Whicker, EdD, MEd
Jack Bookman, MAT, PhD
Aditee P. Narayan, MD, MPH
Betty B. Staples, MD
Holly Hering, BS
Kathleen A. McGann, MD

Abstract

Background The Pediatrics Milestone Project uses behavioral anchors, narrative descriptions of observable behaviors, to describe learner progression through the Accreditation Council for Graduate Medical Education competencies. Starting June 2014, pediatrics programs were required to submit milestone reports for their trainees semiannually. Likert-type scale assessment tools were not designed to inform milestone reporting, creating a challenge for Clinical Competency Committees.

Objective To determine if milestone-based assessments better stratify trainees by training level compared to Likert-type assessments.

Methods We compared assessment results for 3 subcompetencies after changing from a 5-point Likert scale to milestone-based behavioral anchors in July 2013. Program leadership evaluated the new system by (1) comparing PGY-1 mean scores on Likert-type versus milestone-based assessments; and (2) comparing mean scores on the Likert-type versus milestone-based assessments across PGY levels.

Results Mean scores for PGY-1 residents were significantly higher on the prior year's Likert-type assessments than milestone-based assessments for all 3 subcompetencies ($P < .01$). Stratification by PGY level was not observed with Likert-type assessments (eg, interpersonal and communication skills 1 [ICS1] mean score for PGY-1, 3.99 versus PGY-3, 3.98; $P = .98$). In contrast, milestone-based assessments demonstrated stratification by PGY level (eg, the ICS1 mean score was 3.06 for PGY-1, 3.83 for PGY-2, and 3.99 for PGY-3; $P < .01$ for PGY-1 versus PGY-3). Significantly different means by trainee level were noted across 21 subcompetencies on milestone-based assessments ($P < .01$ for PGY-1 versus PGY-3).

Conclusions Initial results indicate milestone-based assessments stratify trainee performance by level better than Likert-type assessments. Average PGY-level scores from milestone-based assessments may ultimately provide guidance for determining whether trainees are progressing at the expected pace.

Kathleen W. Bartlett, MD, is Associate Professor of Pediatrics and Associate Program Director, Duke Pediatric Residency Training Program, Duke University Medical Center; **Shari A. Whicker, EdD, MEd,** is Assistant Professor, Medical Education and Faculty Development Specialist, Virginia Tech Carilion School of Medicine; **Jack Bookman, MAT, PhD,** is Professor Emeritus of Mathematics, Duke University; **Aditee P. Narayan, MD, MPH,** is Associate Professor of Pediatrics and Associate Program Director, Duke Pediatric Residency Training Program, Duke University Medical Center; **Betty B. Staples, MD,** is Associate Professor of Pediatrics and Program Director, Duke Pediatric Residency Training Program, Duke University Medical Center; **Holly Hering, BS,** is Program Coordinator, Duke Pediatric Residency Training Program, Duke University Medical Center; and **Kathleen A. McGann, MD,** is Professor of Pediatrics, and Vice Chair of Education, Office of Pediatric Education, Duke University Medical Center.

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

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

Study results were previously presented at the annual Association of Pediatric Program Directors Spring Meeting, April 2014 in Chicago.

The authors would like to thank the Duke University Pediatrics and Combined Internal Medicine-Pediatrics Residents, and Pediatrics Core Faculty.

Corresponding author: Kathleen W. Bartlett, MD, Duke University Medical Center, Box 3127, Durham, NC 27710, katy.bartlett@duke.edu

Received July 3, 2014; revision received October 7, 2014; accepted November 10, 2014.

Introduction

In 2010, the Accreditation Council for Graduate Medical Education (ACGME) and the American Board of Pediatrics (ABP) introduced the Pediatrics Milestone Project,[1] a new method of evaluating program effectiveness based on residents' aggregated performance on educational milestones.[2] The milestones are narrative anchors describing observable behaviors that provide a framework for measuring learner progression from novice to expert through the subcompetencies of the 6 ACGME competencies.[3] Starting in June 2014, pediatrics residency programs are required to submit semiannual milestone reports on 21 subcompetencies for each trainee to the ACGME as part of the Next Accreditation System.[4]

National collaborative efforts are underway to develop tools with validity evidence for the assessment of Pediatrics

DOI: http://dx.doi.org/10.4300/JGME-D-14-00389.1

Milestones, but results from these efforts were not available before June 2014.[5,6] As noted by Schumacher et al,[7] "in the absence of nationally developed assessment tools with sufficient validity and reliability, individual programs must . . . develop and/or identify currently available tools and mechanisms for assessing milestones in their programs until such tools are available."

This study sought to determine if assessments that incorporated milestone-based behavioral anchors (narrative descriptions of observable behaviors) stratified learners by level better than assessments scored on a Likert scale.

What was known and gap

Likert scales are commonly used in assessment, and it is not known whether milestone-based assessments are superior in stratifying pediatrics resident performance compared to Likert-type assessments.

What is new

Milestone-based assessments showed different means by trainee level across 21 subcompetencies.

Limitations

Single institution, single specialty study, and historical comparison limit generalizability.

Bottom line

Milestone-based assessments demonstrated improved stratification by PGY level, and may provide better guidance on whether trainees are progressing as expected.

Methods

System Development

Starting in February 2013, the Duke University Pediatrics Curriculum Committee and core faculty used expert consensus and a modified Delphi process[8] to map the 21 subcompetencies required by the ACGME to resident rotations based on where each subcompetency could be best assessed. This process ensured that all 21 subcompetencies were assessed at multiple points in the curriculum.[6] Additional subcompetencies from the Pediatrics Milestone Project[1] beyond the 21 required "reporting" milestones were included if requested by the core faculty for the rotation. In all, 35 of the 48 subcompetencies were represented on the assessments. Each rotation had a separate assessment form that included 5 to 10 subcompetencies. Peer assessments incorporated 6 subcompetencies and were administered during neonatology, critical care, and inpatient rotations, where trainees worked closely with peers. Rating scales were changed from a 5-point Likert scale (unsatisfactory performance, needs development, meets expectations, exceeds expectations, exceptional performance) to milestone-based anchors taken verbatim from the Pediatrics Milestone Project.[1] Prior to implementing the new assessments, residency leadership presented at a departmental faculty meeting and then met with each division individually to provide faculty development on the subcompetencies and Pediatrics Milestones, using a pilot version of a module created for that purpose.[9] The module incorporated a video scenario for scoring practice using milestone-based behavioral anchors.[6] In the past, faculty did not receive specific instruction on how to score Likert-type assessments, although program directors conducted periodic faculty development sessions on giving robust formative and summative feedback to trainees.

Study Population

The study population included the categorical pediatrics and combined internal medicine-pediatrics (med-peds) residency programs at Duke University Medical Center. There are 16 categorical residents and 6 combined med-peds residents per year. The aggregate Likert assessment results of the 2012–2013 cohort were compared to the aggregate milestone-based results of the 2013–2014 cohort. Because all end-of-rotation and peer assessments were included, each resident had multiple assessments across different settings. We compared demographic data and in-training examination standard scores for the 2 cohorts to establish their similarities.

Procedure

Both the Likert-type and milestone-based assessments were housed in MedHub, an online commercial residency management system and assessment repository.[10] Using 2 consecutive years of resident end-of-rotation assessments, we compared assessment data from July 2013–February 2014 to those of July 2012–February 2013. We included data from the first 7 rotations of the academic year only because we expected resident performance in the second half of the year to differ from resident performance in the first half of the year. We compared results for 3 subcompetencies after changing from Likert-type to milestone-based assessments: interpersonal and communication skills 1 (ICS1), patient care 1 (PC1), and patient care 6 (PC6). These subcompetencies were well preserved in the conversion from Likert-type to milestone-based assessments (BOX), whereas most of the other subcompetencies were not represented on the Likert-type assessments. We also sought to compare results by PGY level on the milestone-based assessments for all 21 reportable subcompetencies and the 35 chosen by Duke University faculty.

This study was determined to be exempt by Duke University Medical Center's Institutional Review Board.

BOX **Subcompetencies that are Well-Represented on Likert-Type and Milestone-Based Assessments**

Likert-Type Assessments	Milestone-Based Assessments
1. (ICS1) Demonstrate positive attitudes, behaviors, and interpersonal skills in relation to patients and families.	1. (ICS1) Communicate effectively with patients, families, and the public, as appropriate, across a broad range of socioeconomic and cultural backgrounds.
2. (PC1) Efficiently and appropriately gather patient data through medical history, physical examinations, and diagnostic tests.	2. (PC1) Gather essential and accurate information about the patient.
3. (PC6) Formulate an effective plan based on the gathered data.	3. (PC6) Make informed diagnostic and therapeutic decisions that result in optimal clinical judgment.

Abbreviations: ICS, interpersonal and communication skills; PC, patient care.

Analysis

For the 3 preserved subcompetencies, we compared mean scores for PGY-1s on the Likert scale assessments to mean scores on the newer milestone-based assessments. We also compared mean scores across PGY levels on Likert-type and milestone-based assessments. Finally, we averaged the mean scores across the 21 reportable subcompetencies and all 35 subcompetencies used in our assessments by PGY level. Two-sample *t* tests were performed to determine *P* values and 95% confidence intervals (CIs). This method of analysis allowed for comparison of means for PGY-1 versus PGY-2, PGY-2 versus PGY-3, and PGY-1 versus PGY-3. We compared mean scores by class for each of the 3 subcompetencies on Likert-type and milestone-based assessments and mean scores by class across all subcompetencies on the milestone-based assessments. Analyses were conducted in aggregate because the intent of the study was to assess the assessment tool not the individual. All data were anonymized.

TABLE 1 **Characteristics of 2 Cohorts**

Characteristic	2012–2013 Cohort (n = 71), No. (%)	2013–2014 Cohort (n = 72), No. (%)
Categorical pediatrics	47 (66)	48 (67)
Combined med-peds	24 (34)	24 (33)
Men	23 (32)	22 (31)
PGY-1	22 (31)	22 (31)
PGY-2	22 (31)	22 (31)
PGY-3	21 (30)	22 (31)
PGY-4	6 (8)	6 (8)
ITE standard scores[a]	**Class Average**	**Class Average**
PL-1/MP-1	213	209
PL-2/MP-2/MP-3	346	342
PL-3/MP-4	370	391

Abbreviations: meds-peds, categorical pediatrics and combined internal medicine and pediatrics residency program; PGY, postgraduate year; ITE, in-training examination; PL-1, first-year categorical pediatrics residents; MP-1, first-year combined med-peds residents; PL-2, second-year categorical pediatrics residents; MP-2, second-year combined med-peds residents; MP-3, third-year combined med-peds residents; PL-3, third-year categorical pediatrics residents; MP-4, fourth-year combined med-peds residents.

[a] American Board of Pediatrics (ABP) ITE is conducted in July of each academic year. The average of standard scores for each training level for each cohort is reported. In 2013, the ABP moved from a standard score (0–800) based on the average performance of the reference group taking the certifying examination to a new criterion-based scaled score (1–300). To allow for comparisons between the 2 years, average 2013–2014 ITE scores have been converted from the scaled score to the previous standard scores (0–800) by using a table provided by the ABP for that purpose. Differences in class average between the 2 cohorts were not statistically significant, using a 2-sample *t* test.

Results

Characteristics of the 2 cohorts are presented in TABLE 1. The proportion of combined med-peds to categorical pediatrics residents was stable from year to year, and the balance between sexes in the cohorts was similar. Standard scores on the ABP in-training examination were similar between the 2 cohorts.

For PGY-1s, there was a range of 59 to 92 total responses to the assessment questions for the 3 subcompetencies. PGY-2s and PGY-3s had a total response range of 16 to 78. For each of the 3 subcompetencies (ICS1, PC1, and PC6), the mean score for PGY-1s on milestone-based assessments was significantly lower than that for the PGY-1 score on the Likert-type assessments (TABLES 2 and 3). For ICS1 and PC1, there were no differences in mean scores across PGY levels on the Likert-type assessments. By comparison, on the milestone-based assessments for ICS1 and PC1, there were significant increases in mean score between PGY-1s and PGY-2s; differences between PGY-2s and PGY-3s were not statistically significant. For PC6, there was a small increase in mean score of 0.34 points ($P = .02$, 95% CI 0.06–0.62) overall between PGY-1 and PGY-3 on the Likert-type assessments. On the milestone-based assessment for PC6, there was a larger increase in mean score between PGY-1s and PGY-3s of 0.83 points overall ($P < .01$, 95% CI 0.64–1.04), with most of this increase occurring between PGY-2s and PGY-3s, and a nonstatistically significant difference between PGY-1s and PGY-2s. Analysis of this PC6 subcompetency was limited by the low number of responses for PGY-2s.

TABLE 2 **PGY-1 Mean Scores on Likert-Type Versus Milestone-Based Assessments**

Subcompetency	Mean (n)[a] Likert-Type Score	Mean (n)[a] Milestone-Based Score	*P* Value
(ICS1) Communicate effectively with patients, families, and the public	3.99 (92)	3.06 (73)	< .01[b]
(PC1) Gather essential and accurate information about the patient	3.60 (88)	2.93 (74)	< .01[b]
(PC6) Make informed diagnostic and therapeutic decisions	3.55 (91)	2.87 (59)	< .01[b]

Abbreviations: ICS, interpersonal and communication skills; PC, patient care.

[a] n = No. of assessments completed for given subcompetency.
[b] Indicates statistically significant differences between groups to a level of $P < .01$ on a 2-sample *t* test.

Analysis of the average of mean scores across all 21 reportable subcompetencies revealed clear stratification by PGY level, with significant increases in mean scores between PGY-1s versus PGY-2s and PGY-2s versus PGY-3s (TABLE 4 and FIGURE). This stratification was maintained when data for all 35 subcompetencies chosen by the core faculty for inclusion in the milestone-based assessments were analyzed.

Discussion

We found that milestone-based assessments resulted in lower average scores for PGY-1s and significant increases in mean scores between the first-year and third-year trainees. This stratification by training level did not occur with Likert-type rating scales. The difference suggests that the milestone-based approach, including use of milestone anchors and targeted faculty development, was more effective in stratifying resident performance.

Prior to this study, there were no published results of the performance of the Pediatrics Milestones in practice. Our results provide preliminary feasibility experience and demonstrate a trajectory in scores over time with greater differentiation among learners as they progress through training. Milestone-based assessments also provide the program with preliminary level-specific averages for trainee performance. Other GME specialties and subspecialties could replicate this process as they develop and implement specialty-specific milestones.[11]

There were some drawbacks to implementing milestone-based assessments in our residency program. The faculty found that the verbatim behavioral anchors were

TABLE 3 **Stratification by PGY Level for 3 Subcompetencies**

Subcompetency	PGY Level	Likert-Type Assessment			Milestone-Based Assessment		
		n[a]	Mean	*P* value	n[a]	Mean	*P* value
(ICS1) Communicate effectively with patients, families, and the public	PGY-1	92	3.99		73	3.06	
	PGY-2	53	3.74	.06 (1 vs. 2)	60	3.83	< .01 (1 vs. 2)[b]
	PGY-3	68	3.98	.07 (2 vs. 3)	73	3.99	.24 (2 vs. 3)
				.98 (1 vs. 3)			< .01 (1 vs. 3)[b]
(PC1) Gather essential and accurate information about the patient	PGY-1	88	3.60		74	2.94	
	PGY-2	20	3.65	.79 (1 vs. 2)	63	3.81	< .01 (1 vs. 2)[b]
	PGY-3	21	3.71	.77 (2 vs. 3)	78	3.97	.23 (2 vs. 3)
				.54 (1 vs. 3)			< .01 (1 vs. 3)[b]
(PC6) Make informed diagnostic and therapeutic decisions	PGY-1	91	3.55		59	2.87	
	PGY-2	25	3.56	.94 (1 vs.. 2)	16	2.91	.78 (1 vs. 2)
	PGY-3	36	3.89	.09 (2 vs.. 3)	44	3.71	< .01 (2 vs. 3)[b]
				.02 (1 vs. 3)[c]			< .01 (1 vs. 3)[b]

Abbreviations: ICS, interpersonal and communication skills; PC, patient care.

[a] n indicates No. of assessments completed for given PGY level for a given subcompetency.
[b] Statistically significant difference between groups to level of $P < .01$ on a 2-sample *t* test.
[c] Statistically significant difference between groups ($P < .05$) on a 2-sample *t* test.

TABLE 4 **Stratification Across All Subcompetencies With Milestone-Based Assessments**

PGY Level	Mean for 21 Subcompetencies	*P* Value	Mean for 35 Subcompetencies[a]	*P* Value
PGY-1	2.75		2.65	
PGY-2	3.36	< .01 (1 vs. 2)[b]	3.30	< .01 (1 vs. 2)[b]
PGY-3	3.73	< .01 (2 vs. 3)[b]	3.72	< .01 (2 vs. 3)[b]
		< .01 (1 vs. 3)[b]		< .01 (1 vs. 3)[b]

[a] In addition to the 21 subcompetencies reportable to the ACGME, 14 extra subcompetencies (from the 48 in the Pediatrics Milestone Project[1]) were included in the milestone-based assessments at the request of core faculty overseeing resident rotations.
[b] Indicates statistically significant differences between groups to level of $P < .01$ on a 2-sample *t* test.

cumbersome to read, making the assessments quite lengthy. As a result, narrative comments on assessments seemed to decrease, and some grade inflation still occurred. Clearly, faculty development efforts need to be ongoing, and our program plans to use these study results to review the concepts of milestone-based assessment with faculty. We are also considering abbreviating the language in our milestone-based anchors. Faculty felt that not all important subcompetencies were reflected in the 21 milestones currently reported to the ACGME. Therefore, additional subcompetencies and non–milestone-based questions were included in the assessments. Finally, there is no evidence to suggest that milestone-based assessments should replace existing assessment tools completely.

Our study has several limitations. It occurred at a single institution in a single training program with only 7 months of data for each assessment type, and there were relatively few responses to some of the assessment questions of interest. The assessments did not compare the same individuals. Rather, for convenience, the results of the old assessments from 2012 to 2013 were compared to results of the new assessments from 2013 to 2014 in a similar population of residents. In addition, there were only 3 subcompetencies that had compatible enough language on the milestone-based and Likert-type assessments to allow for direct comparison. Although the 5-point Likert scale correlates with the 5 milestone levels for most pediatric subcompetencies, the highest level on the Likert scale (exceptional performance) may have been easier to achieve than the aspirational behaviors described by the highest level of the milestones. Finally, faculty development that was provided prior to the implementation of the milestone-based assessments may have affected the results by creating the expectation that residents' scores should increase as they progress though training. Given the fact that no specific faculty development on Likert-type assessments was provided in the past, it is not possible to determine whether the milestone-based anchors or the faculty development were responsible for the differences between the old and new ratings.

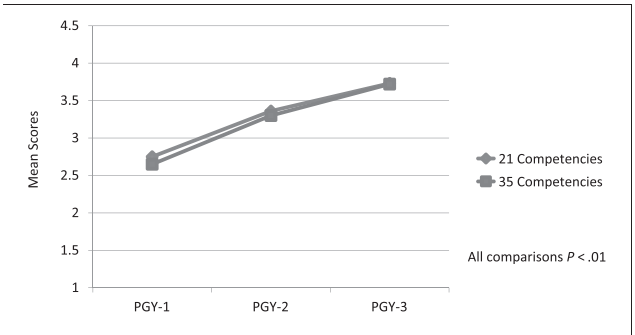

FIGURE **Stratification Across Competencies By PGY Using Milestone-Based Assessments**

The verbatim Pediatrics Milestones were not intended for use as assessment tools, and multi-institution, multi-year, prospective studies of assessment tools designed to inform milestone ratings are needed.[6] The Pediatrics Milestone Project Working Group is also looking to embed the milestones in entrustable professional activities as described by ten Cate and Scheele.[6,12] In the interim, many pediatrics programs are using some form of milestone-based assessment to inform ACGME reporting. In a survey performed by the Association of Pediatric Program Directors, 27% of programs reported using the verbatim milestone anchors, and 61% reported using modified milestone anchors for faculty evaluations of residents.[13]

Although our study demonstrated significant increases in mean scores on milestone-based assessments from PGY-1 to PGY-3, scores were clustered within a relatively small range (2.87–3.99). Few trainees will score at the extreme ends of the milestone spectrum as it is currently constructed, given that they span the continuum of training and practice from entry level to an aspirational level.[2,11] This may indicate a need to further differentiate the

behavior rubrics in the middle of the spectrum in order to better describe learner progression through the narrower range of milestones that are relevant to residency training.

Conclusion

Traditional Likert-type assessment tools did not distinguish among learners by training level as well as milestone-based assessments for 3 subcompetencies. Incorporating milestone-based behavioral anchors in trainee assessments may better stratify performance by training level, which may be helpful in determining benchmarks for resident progression in the future.

References

1 Pediatrics Milestone Working Group. The Pediatrics Milestone Project. *Acad Pediatrics*. 2014;14(suppl 2):1–97.
2 Hicks PJ, Schumacher DJ, Benson BJ, Burke AE, Englander R, Guralnick S, et al. The pediatrics milestones: conceptual framework, guiding principles, and approach to development. *J Grad Med Educ*. 2010;2(3):410–418.
3 Carraccio C, Benson B, Burke A, Englander R, Guralnick S, Hicks P, et al. Pediatrics milestones. *J Grad Med Educ*. 2013;5(suppl 1):59–73.
4 Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366(11):1051–1056.
5 American Board of Pediatrics. Foundation Support of Milestones Study. 2013. https://www.appd.org/pdf/foundationSupportofMilestonesStudy.pdf. Accessed June 24, 2014.
6 Hicks PJ, Englander R, Schumacher DJ, Burke A, Benson BJ, Guralnick S, et al. Pediatrics milestone project: next steps toward meaningful outcomes assessment. *J Grad Med Educ*. 2010;2(4):577–584.
7 Schumacher DJ, Spector ND, Calaman S, West DC, Cruz M, Frohna JG, et al. Putting the pediatrics milestones into practice: a consensus roadmap and resource analysis. *Pediatrics*. 2014;133(5):898–906.
8 Dalkey NC, Helmer O. An experimental application of the Delphi methods to the use of experts. *Manage Sci*. 1963;9(3):458–457.
9 Bhavaraju V, Bartlett K, Schumacher D, Guillot A. Faculty development series on assessment in graduate medical education: the milestone project. MedEdPORTAL Publications. 2014. www.mededportal.org/publication/9898. Accessed November 26, 2014.
10 MedHub Inc. http://medhub.com/. Accessed June 24, 2014.
11 Swing SR, Beeson MS, Carraccio C, Coburn M, Lobst W, Selden NR, et al. Educational milestone development in the first 7 specialties to enter the next accreditation system. *J Grad Med Educ*. 2013;5(1):98–106.
12 ten Cate O, Scheele F. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Acad Med*. 2007;82(6):542–547.
13 Fagan HSB, Perkins K. APPD grassroots forum for program directors. Oral presentation at the Association of Pediatric Program Directors Spring Meeting, Chicago, IL, 2014.