Preliminary Validity Evidence for a Milestones-Based Rating Scale for Chart-Stimulated Recall

Shalini T. Reddy, MD, MHPE
Ara Tekian, PhD, MHPE
Steven J. Durning, MD, PhD
Shanu Gupta, MD
Justin Endo, MD, MHPE
Brenda Affinati, MD
Yoon Soo Park, PhD

ABSTRACT

Background Minimally anchored Standard Rating Scales (SRSs), which are widely used in medical education, are hampered by suboptimal interrater reliability. Expert-derived frameworks, such as the Accreditation Council for Graduate Medical Education (ACGME) Milestones, may be helpful in defining level-specific anchors to use on rating scales.

Objective We examined validity evidence for a Milestones-Based Rating Scale (MBRS) for scoring chart-stimulated recall (CSR).

Methods Two 11-item scoring forms with either an MBRS or SRS were developed. Items and anchors for the MBRS were adapted from the ACGME Internal Medicine Milestones. Six CSR standardized videos were developed. Clinical faculty scored videos using either the MBRS or SRS and following a randomized crossover design. Reliability of the MBRS versus the SRS was compared using intraclass correlation.

Results Twenty-two faculty were recruited for instrument testing. Some participants did not complete scoring, leaving a response rate of 15 faculty (7 in the MBRS group and 8 in the SRS group). A total of 529 ratings (number of items $\times$ number of scores) using SRSs and 540 using MBRSs were available. Percent agreement was higher for MBRSs for only 2 of 11 items—use of consultants (92 versus 75, $P = .019$) and unique characteristics of patients (96 versus 79, $P = .011$)—and the overall score (89 versus 82, $P < .001$). Interrater agreement was 0.61 for MBRSs and 0.51 for SRSs.

Conclusions Adding milestones to our rating form resulted in significant, but not substantial, improvement in intraclass correlation coefficient. Improvement was inconsistent across items.

Introduction

While direct observation is central to workplace-based assessment (WBA) of clinical competence, assessors contribute to substantial variability in ratings.[1–4] Another source of variability appears to be the format of the rating instrument. Efforts to improve the reliability of WBAs have generally been directed toward manipulating forms and rating scales.[5]

Several rating scales are commonly used for WBA. Standard Rating Scales (SRSs) are often minimally anchored by words such as *unsatisfactory*, *good*, *very good*, and *outstanding*, and are in widespread use in residency education. An example of an SRS scoring form is the mini-CEX.[6]

A large scale investigation in the United Kingdom found that a rating scale aligned to descriptions of behaviors improved reliability of ratings and reduced variability in scale interpretation.[7] Expert-derived frameworks of clinical competence, like the Accreditation Council for Graduate Medical Education (ACGME) Milestone Projects, may be helpful in defining level-specific performance for rating scales.[8,9] Although preliminary studies have shown advantages to using Milestones-Based Rating Scales (MBRSs) over SRSs in differentiating among levels of learners,[10,11] there is insufficient validity evidence to recommend conversion to MBRS for *all* assessments. Despite the paucity of studies to support the use of milestones for rating forms, many residency management software systems provide program directors with the option to create assessment forms with milestones that are used verbatim as anchors on forms. There are concerns with programs' use of the milestones as anchors on rating scales without data to support validity.[12]

The goal of this study was to gather evidence to support the construct validity of MBRSs for scoring chart-stimulated recall (CSR). CSR is a modified oral examination that uses the examinee's patient, rather than a standardized case, as a stimulus for assessing clinical reasoning.[13,14] Use of CSR allows an assessor to obtain information about residents' clinical decision-making by querying them about their management of the patient. Few studies provide clear guidance on how to optimally score CSR and, to our knowledge, there are no MBRS scoring forms for CSR.[7]

DOI: http://dx.doi.org/10.4300/JGME-D-17-00435.1

Editor's Note: The online version of this article contains the survey instrument, an example of a learner profile, and a sample script.

We compared the reliability of MBRSs versus SRSs for assessing CSR using the ACGME Milestones as descriptive anchors on a rating form. We used Messick's framework to identify content and response process validity evidence.[15]

What was known and gap
Expert-derived rating scales to improve the quality of assessments are needed in a wide range of areas.

What is new
A milestones-based scale for scoring a chart-stimulated recall task resulted in greater interrater agreement among internal medicine faculty compared with a standardized rating scale.

Limitations
Small sample, single specialty study, and lack of rater training limit generalizability.

Bottom line
Use of milestones in rating forms resulted in significant, but not substantial, improvement in rater agreement; improvement was inconsistent across items.

Methods

Study Design

The study consisted of (1) scoring instrument development; (2) standardized case development; and (3) instrument testing. A prospective randomized crossover design was used. The study was conducted at 4 unaffiliated internal medicine (IM) residency programs in Chicago—3 academic medical centers and 1 community hospital.

Scoring Instrument Development (Content Validity)

Two scoring instruments were developed—one with an SRS and another with an MBRS. An iterative process of blueprinting CSR to the ACGME Internal Medicine Milestones was used to establish content validity (provided as online supplemental material).[9,16] Two IM faculty and 2 chief residents independently reviewed the 22 subcompetencies and 139 associated milestones. Eleven subcompetencies and 27 milestones were identified as measurable by CSR (see TABLE 1 for a sample item).

The 11 identified subcompetencies were used as items for both forms. The SRS form mirrored the 9-point ACGME level-based scale. The 27 identified milestones were added as anchors for the MBRS. A tenth option, *insufficient information*, was added to each form. Four questions were added to the end of each form to gather comments describing how raters chose their scores.

Standardized Case Development

Six cases, representing a spectrum of examinee performance, were developed.[17,18] Two authors (S.T.R. and S.G.) created descriptions of learners using behaviors described in the 11 subcompetencies identified earlier. All depicted examinees were residents who demonstrated varying levels of ability to answer questions about the management of common inpatient cases. An example of a learner profile is provided as online supplemental material. Learners were not designated by postgraduate year since progression through the milestones is competency based rather than time based. Scripts were developed based on how 2 of the authors (S.T.R. and S.G.) thought a learner would answer a question posed by a CSR examiner. A sample script is available as online supplemental material. Scripts were reviewed by experienced clinician-educators blinded to the learner profiles to ensure that the level of the learners matched the answers they gave in the script. Reviewed scripts were used to develop videos of CSR encounters. All examinees videotaped were men to avoid possible variations in scoring due to gender.

Instrument Testing (Response Process)

Recruitment: Eligible participants were IM generalist physicians who taught and evaluated residents during inpatient rotations. We excluded IM subspecialists and faculty who spend less than 20% of their time providing patient care to avoid error from lack of familiarity with general clinical management of the depicted cases. A total of 84 faculty were invited by e-mailing the general IM and hospital medicine listservs at participating sites. No incentives were offered for participation.

Demographic Data Collection: Data were collected to determine characteristics that may influence responses to items, including gender concordance between faculty and examinee[19,20] and time spent in education and clinical work.[21]

Randomization and Testing: Participants were randomized to 1 of 2 groups (FIGURE 1). A web-based platform was created for viewing and scoring. Formal rater training was purposefully not provided for 2 reasons: (1) lack of formal rater training simulates real-life settings where faculty development is not consistently provided,[7] and (2) MBRSs may improve scoring accuracy independent of rater training by offering a shared mental model of examinee performance. After being given a definition of CSR and guidance on navigating the online platform, all participants watched 3 videos in the same sequence during each session.

During the first scoring session, 1 group used the MBRS and 1 used the SRS to score the same 3 videos.

TABLE 1
Example of Rating Scale for Chart-Stimulated Recall[a]

Please Indicate the Resident's Ability to Accomplish This Task: **Item 3: Appropriately Use Consultants**					
Standard Rating Scale	Level 1, Critically deficient	Level 2	Level 3	Level 4, ready for unsupervised practice	Level 5, Aspirational
Milestones-Based Rating Scale	Does not use consultant services when needed for patient care	Unable to justify reason(s) for consultation	Asks meaningful clinical questions that guide the input of consultants	Weighs recommendations from consultants in order to effectively manage patients	Manages discordant recommendations from multiple consultants

[a] Standard Rating Scale form contains item, first and last row; Milestones-Based Rating Scale form contains all 3 rows.

Participants were given up to 4 weeks to complete the scoring and were sent weekly reminders. Two weeks after completing the first session, all participants were invited to another scoring session using different forms.

Institutional Review Boards at each participating site reviewed and approved this study. Participants provided consent according to the requirements of each study site.

Data Analysis

Chi-square tests were used to examine differences between the 2 randomized groups to test for potential bias. Interrater reliabilities between the SRS and MBRS were compared between cases and across levels of performance. Intraclass correlation coefficients (ICCs) were calculated to account for chance agreement.[22,23] Comments were analyzed using the constant comparative method without an a priori framework.

Results

Script Review

Nine of 10 invited expert clinician-educators reviewed the scripts. Final distribution of expert scores showed a range of performance levels across the 6 cases, ranging from level 1 to level 4 (FIGURE 2).

Instrument Testing

Eleven clinician-educators enrolled in each group; some participants did not complete scoring, leaving a total of 7 faculty in the MBRS group and 8 in the SRS group (FIGURE 1). TABLE 2 shows demographic characteristics.

A total of 529 ratings (number of items × number of scores) using the SRS and 540 using the MBRS were available (TABLE 3). Thirty-seven items scored as *insufficient information* for MBRS, and 27 were excluded for the SRS group. Percent agreement between the SRS and MBRS was higher with MBRS for the following items: use of consultants (92 versus 75, $P = .019$); unique characteristics of patients (96 versus 79, $P = .011$); and overall score (89 versus 82, $P < .001$). Interrater agreement was higher for the MBRS than SRS (ICC 0.61 versus 0.51). Using MBRS

TABLE 2
Participant Characteristics

Characteristics	SRS First[a]	MBRS First[a]	Between-Group Differences *P* Value
Gender			
Male	4	5	.80
Female	6	6	
Type of practice			
Academic medical center	8	8	.69
Community hospital	2	3	
Years since residency			
< 1	2	0	.36
1–5	6	6	
6–10	2	3	
11–15	0	1	
% time spent on education			
< 25	7	7	.22
26–50	3	1	
51–75	0	2	
Clinical workload index[b]			
Low	2	3	.29
Moderate	3	6	
Heavy	5	2	

Abbreviations: SRS, standard rating scale; MBRS, Milestones-Based Rating Scale.

[a] Participant numbers may not add up to 11 in each group as not all participants answered all demographic questions. MBRS first: 1 participant dropped out after enrollment; 1 participant skipped questions "years since residency" and "% time spent on education."

[b] The clinical workload index was calculated by separating percentage of time spent in patient care and average daily census into quartiles and summing the 2 quartiles to get a "clinical workload" index.

TABLE 3
Overall Interrater Agreement

Item	Standard Rating Scale (SRS)			Milestones-Based Rating Scale (MBRS)			Between-Group Differences % Agreement (MBRS-SRS)	*P* Value[b]
	n	% Agreement[a]	% Discrepant	n	% Agreement[a]	% Discrepant		
1. Gather and present essential clinical information	51	88	12	52	88	12	0	.97
2. Develop a comprehensive management plan for a patient	51	86	14	52	88	12	2	.74
3. Appropriately use consultants	51	75	25	50	92	8	17	.019[c]
4. Apply knowledge of clinical medicine	51	88	12	50	88	12	0	.97
5. Demonstrate a working knowledge of diagnostic testing/procedures	50	90	10	50	96	4	6	.24
6. Recognize and respond to the unique characteristics and needs of a patient	48	79	21	50	96	4	17	.011[c]
7. Recognize errors in the system and advocate for system improvement	46	74	26	46	89	11	15	.06
8. Identify forces that impact the cost of health care	36	81	19	43	91	9	10	.19
9. Manage patient transitions from inpatient to outpatient care settings	49	82	18	50	84	16	2	.76
10. Self-reflect with a goal for improvement	50	78	22	50	80	20	2	.81
11. Actively pursue knowledge to improve patient care	46	80	20	47	87	13	7	.37
Overall	529	82	18	540	89	11	7	$< .001$[c]
Intraclass correlation	0.51			0.61				

[a] % exact + adjacent agreement.
[b] % *P* values taken from chi-square test comparing differences in proportions between % exact + adjacent and % discrepant between SRS and MBRS. Significant differences may be spurious as analysis did not include correction for multiple comparisons.
[c] % $P < .05$.

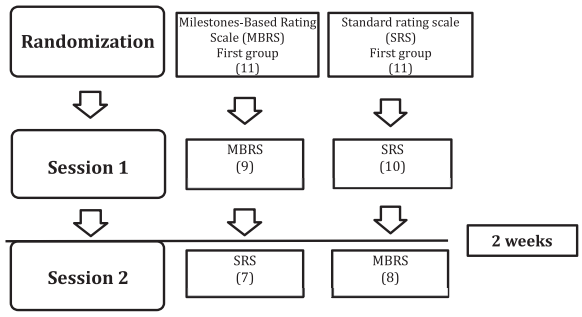

FIGURE 1
Randomization Scheme

first did not lead to an increased likelihood of rating learners higher during a second session using the SRS.

Analysis of feedback from raters showed that, in addition to using their practices and other learners, faculty used words on the rating scales on both forms as frames of reference. Faculty struggled with the cognitive load of scoring learners while observing encounters.

Our results showed that the SRS had poor ICC for scoring CSR, and adding the MBRS was only marginally helpful in improving the reliability of scoring. Adding milestones to our rating form resulted in significant, but not substantial, improvement in ICC, and improvement was inconsistent across items.

Discussion

Our findings are consistent with assertion by Williams and colleagues[12] that adding descriptive anchors to assessment forms may result in cognitive overload without attendant improvement in assessment validity. Our findings diverge from those of Crossley and colleagues,[7] who noted consistent and substantial variation between a "construct-aligned" assessment scale, analogous to our MBRS and a conventional scale. The way in which anchors in that study were developed was not described, and it is possible that their anchors were developed using a more detailed process. Furthermore, their study included 2000 trainees, which likely resulted in an improved ability to detect differences. It is possible that the assessors were more familiar with the assessment methods used in their study.[7]

A disappointing finding is that there were no substantial differences for items 7 through 11 measuring systems-based practice (SBP) and practice-based learning and improvement (PBLI). Prior studies

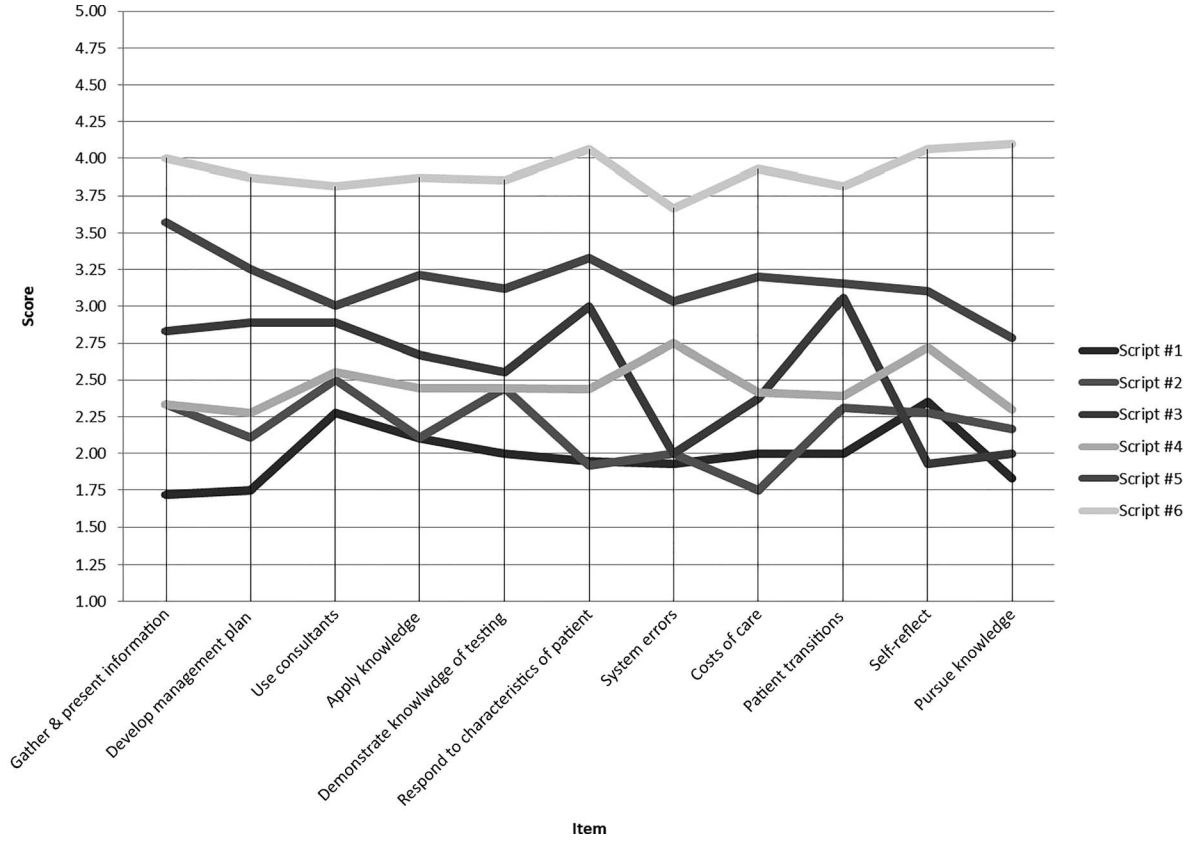

FIGURE 2
Expert Ratings of Scripts

have reported confusion regarding measuring these 2 competencies. Providing descriptions of behaviors that demonstrate dimensions of performance in SBP and PBLI could help evaluators develop a shared mental model.[24,25] It is possible that behavioral descriptors from the ACGME Milestones added clarity to items that asked about using consultants and responding to characteristics of patients. Variation in the effectiveness of the MBRS in improving reliability suggests that raters may already have shared mental models for some competencies such as data gathering.

The study engaged faculty in different settings, lending support for generalizability to diverse groups of faculty. Use of standardized videos limited the potential for distractions related to the work environment.[26] The randomized crossover design allowed us to determine whether the MBRS impacted future scoring with the SRS.

This study has limitations. Formal rater training was not conducted, consistent with common practice as instructions on how to use the form may be more commonplace, and contrasted with faculty development on how to use the form to assess the individual.[27] It is possible that brief in-person training sessions would have improved scoring accuracy for both groups. Additionally, scoring a video may be sufficiently different from scoring a learner in actual practice, and may limit the generalizability of our findings to ratings in the field.

Assessing scoring accuracy by comparing the scores of raters to master coders would allow for the collection of internal validity evidence for an MBRS scoring instrument. Formal rater training and deployment of the CSR instrument in the field to determine the utility of the scale when a rater is both the examiner posing the questions and the rater scoring the performance would be a next step. Comparing scores to the level of training could provide further validity evidence.

Conclusion

Although the ACGME Milestones used as narrative anchors for CSR resulted in improvement in interrater reliability, the improvement was small and inconsistent, and interrater reliability remained poor overall.

References

1. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990;65(9):63–67.
2. Albanese AM. Challenges in using rater judgements in medical education. *J Eval Clin Pract*. 2000;6(3):305–319.
3. Gingerich A, Kogan J, Yeates P, et al. Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–1068.
4. Govaerts M. Workplace-based assessment and assessment for learning: threats to validity. *J Grad Med Educ*. 2015;7(2):265–267.
5. Ilgen JS, Ma IW, Hatala R, et al. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ*. 2015;49(2):161–173.
6. American Board of Internal Medicine. Mini-CEX. http://www.abim.org/program-directors-administrators/assessment-tools/mini-cex.aspx. Accessed April 26, 2018.
7. Crossley J, Johnson G, Booth J, et al. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. *Med Educ*. 2011;45(6):560–569.
8. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. *Med Teach*. 2007;29(7):642–647.
9. Iobst W, Aagaard E, Bazari H, et al. Internal medicine milestones. *J Grad Med Educ*. 2013;5(1 suppl 1):14–23.
10. Bartlett KW, Whicker SA, Bookman J, et al. Milestone-based assessments are superior to Likert-type assessments in illustrating trainee progression. *J Grad Med Educ*. 2015;7(1):75–80.
11. Friedman KA, Balwan S, Cacace F, et al. Impact on house staff evaluation scores when changing from a Dreyfus- to a milestone-based evaluation model: one internal medicine residency program's findings. *Med Educ Online*. 2014;19(1):25185.
12. Williams RG, Dunnington GL, Mellinger JD, et al. Placing constraints on the use of the ACGME milestones: a commentary on the limitations of global performance ratings. *Acad Med*. 2015;90(4):404–407.
13. Goulet F, Jacques A, Gagnon R, et al Assessment of family physicians' performance using patient charts interrater reliability and concordance with chart-stimulated recall interview. *Eval Health Prof*. 2007;30(4):376–392.
14. Schipper S, Ross S. Structured teaching and assessment: a new chart-stimulated recall worksheet for family medicine residents. *Can Fam Physician*. 2010;56(9):958–959, e352–e954.
15. Messick S. Validity of test interpretation and use. *ETS Res Rep Series*. 1990;1990(1):1487–1495.
16. Caverzagie KJ, Iobst WF, Aagaard EM, et al. The internal medicine reporting milestones and the next accreditation system. *Ann Intern Med*. 2013;158(7):557–559.
17. Regehr G, Ginsburg S, Herold J, et al. Using "standardized narratives" to explore new ways to

represent faculty opinions of resident performance. *Acad Med*. 2012;87(4):419–427.
18. Rawlings A, Knox A, Park YS, et al. Development and evaluation of standardized narrative cases depicting the general surgery professionalism milestones. *Acad Med*. 2015;90(8):1109–1115.
19. Ehrenberg RG, Goldhaber DD, Brewer DJ. Do teachers' race, gender, and ethnicity matter? Evidence from the National Educational Longitudinal Study of 1988. *Ind Labor Relat Rev*. 1995;48(3):547–561.
20. Mullola S, Ravaja N, Lipsanen J, et al. Gender differences in teachers' perceptions of students' temperament, educational competence, and teachability. *Brit J Educ Psychol*. 2012;82(2):185–206.
21. Kogan JR, Hess BJ, Conforti LN, et al. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Acad Med*. 2010;85(suppl 10):25–28.
22. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas*. 1973;33(3):613–619.
23. Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess*. 1994;6(4):284.
24. Didwania A, McGaghie WC, Cohen E, et al. Internal medicine residency graduates' perceptions of the systems-based practice and practice-based learning and improvement competencies. *Teach Learn Med*. 2010;22(1):33–36.
25. Chen EH, O'Sullivan PS, Pfennig CL, et al. Assessing systems-based practice. *Acad Emerg Med*. 2012;19(12):1366–1371.
26. Calaman S, Hepps JH, Bismilla Z, et al. The creation of standard-setting videos to support faculty observations of learner performance and entrustment decisions. *Acad Med*. 2016;91(2):204–209.
27. Holmboe ES, Ward DS, Reznick RK, et al. Faculty development in assessment: the missing link in competency-based medical education. *Acad Med*. 2011;86(4):460–467.

Shalini T. Reddy, MD, MHPE, is Professor, University of Chicago Pritzker School of Medicine, and Associate Program Director, Internal Medicine Residency, Mercy Hospital and Medical Center; **Ara Tekian, PhD, MHPE,** is Professor, Department of Medical Education, University of Illinois at Chicago; **Steven J. Durning, MD, PhD,** is Professor, Uniformed Services University of the Health Sciences; **Shanu Gupta, MD,** is Assistant Professor, Internal Medicine, Rush University Medical Center; **Justin Endo, MD, MHPE,** is Assistant Professor, University of Wisconsin–Madison; **Brenda Affinati, MD,** is Associate Professor of Medicine, Chicago Medical School, and Associate Program Director, Advocate Lutheran General Hospital; and **Yoon Soo Park, PhD,** is Associate Professor, Department of Medical Education, University of Illinois at Chicago.

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

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

The authors would like to thank Sejal Prachand and Jordan Affinati for their videography and video editing assistance, all of the faculty who participated in the study, and the learners who volunteered to act in the videos.

Corresponding author: Shalini T. Reddy, MD, MHPE, University of Chicago Pritzker School of Medicine, Section of Hospital Medicine, Department of Medicine, MC 5000, 5841 S Maryland Avenue, Chicago, IL 60637, 773.834.5216, sreddy@uchicago.edu

Received June 21, 2017; revisions received November 3, 2017, and January 29, 2018; accepted March 5, 2018.