A Checklist to Help Faculty Assess ACGME Milestones in a Video-Recorded OSCE

L. Jane Easdown, MD, MHPE
Marsha L. Wakefield, MD, MHPE
Matthew S. Shotwell, PhD
Michael R. Sandison, MD

ABSTRACT

Background Faculty members need to assess resident performance using the Accreditation Council for Graduate Medical Education Milestones.

Objective In this randomized study we used an objective structured clinical examination (OSCE) around the disclosure of an adverse event to determine whether use of a checklist improved the quality of milestone assessments by faculty.

Methods In 2013, a total of 20 anesthesiology faculty members from 3 institutions were randomized to 2 groups to assess 5 videos of trainees demonstrating advancing levels of competency on the OSCE. One group used milestones alone, and the other used milestones plus a 13-item checklist with behavioral anchors based on ideal performance. We classified faculty ratings as either correct or incorrect with regard to the competency level demonstrated in each video, and then used logistic regression analysis to assess the effect of checklist use on the odds of correct classification.

Results Thirteen of 20 faculty members rated assessing performance using milestones alone as *difficult* or *very difficult*. Checklist use was associated with significantly greater odds of correct classification at entry level (odds ratio [OR] = 9.2, 95% confidence interval [CI] 4.0–21.2) and at junior level (OR = 2.7, 95% CI 1.3–5.7) performance. For performance at other competency levels checklist use did not affect the odds of correct classification.

Conclusions A majority of anesthesiology faculty members reported difficulty with assessing a videotaped OSCE of error disclosure using milestones as primary assessment tools. Use of the checklist assisted in correct assessments at the entry and junior levels.

Introduction

The implementation of milestone-based assessments by the Accreditation Council for Graduate Medical Education (ACGME) creates a need for residency programs to provide faculty members with training and tools to make these assessments.[1] Each specialty has developed milestones or subcompetencies based on the 6 ACGME competencies for periodic assessment of trainee performance.[2–4] Faculty members evaluate trainees' performance using the milestones, which now replace traditional global faculty assessments. Little is known about the manner in which faculty are trained to make milestone-based assessments, and whether use of milestone-based tools will improve the quality of faculty assessments.

We assessed whether use of a checklist would improve assessments of milestones by anesthesiology faculty at 3 institutions. We developed an objective structured clinical examination (OSCE) scenario around the disclosure of an adverse outcome to a standardized patient (SP), which is a patient care milestone in anesthesiology. Several residency programs use SPs for teaching this activity and have developed milestones for managing errors.[5,6]

DOI: http://dx.doi.org/10.4300/JGME-D-17-00112.1

Editor's Note: The online version of this article contains the faculty survey results, the disclosure objective structured clinical examination (OSCE) checklist, and the milestone/OSCE video evaluation tool.

Methods

In 2013, we e-mailed a description of the study to 20 faculty members from the Education and Clinical Competency Committees of the anesthesiology departments at the University of Alabama at Birmingham (UAB), Vanderbilt University, and Albany Medical Center.

Two authors (L.J.E. and M.L.W.) wrote the scenario for the OSCE: a resident is asked to make a postoperative visit to a female patient who experienced an adverse event (a loose tooth after a difficult intubation). In a 10-minute encounter, the resident must discuss the event, educate the patient about her difficult intubation, and counsel her for future surgery. This scenario is intended to allow the faculty member to assess 5 milestones in the competencies of patient care, professionalism, interpersonal and communications skills, practice-based learning and improvement, and systems-based practice. To demonstrate content validity, faculty

members at UAB and Vanderbilt University reviewed the scenario, and provided feedback as to the ideal observable behaviors based on the literature and their institutions' protocols for managing medical errors, including disclosure to patients.[7–9]

We created an itemized checklist with behavioral anchors for each of 13 items similar to checklists that are used in UAB and Vanderbilt University simulation centers (provided as online supplemental material). The scale for assessment was *adequate*, *inadequate*, or *did not observe*. We developed the milestone assessment tool around the 5 subcompetency milestones selected as the focus of the scenario (provided as online supplemental material).

We recorded five 10-minute videos of the OSCE scenario set at advancing levels of training and competency: *entry* (prior to first year of residency); *junior* (prior to subspecialty training); *midlevel* (subspecialty training); *senior* (ready to graduate); and *advanced* (aspirational). Three Vanderbilt trainees (a medical student, a resident, and a fellow) participated by performing the 5 roles. Trainees complied with the institutional consent process for creating videos. The same SP performed the role of the patient in all 5 videos.

We used a video capture system for medical simulation (B-line Medical, Washington, DC) and placed all 5 videos into a password-protected website randomizing the order of viewing.

Participants were randomized to 2 groups with 10 faculty in each group. Both groups used the milestone assessment tool but 1 group (N = 10) used the checklist in addition to the milestone assessment tool. Each participant received assessment instructions and tools, a description of the OSCE scenario, and a survey. The faculty at 1 institution viewed the videos as a group and completed all 5 assessments before discussion. Faculty at the other 2 institutions viewed and assessed the videos without group discussion. Participants completed an 8-question survey about their teaching experience, prior exposure to OSCEs and milestones, and ease of using the tools.

The Institutional Review Board at Vanderbilt University granted this project exempt status.

Statistical Analysis

All participants viewed and scored each of the 5 video performances. Each participant provided a score (entry, junior, mid, senior, or advanced) for each of the 5 milestones, an overall performance rating, and required level of support rating. We classified video ratings for each milestone, as well as overall performance, as either correct or incorrect. We analyzed the data for all raters, videos, and milestone/competency simultaneously using logistic regression to estimate the odds of correct classification, adjusting for milestone/competency, training level portrayed in the video, checklist use by the faculty rater, interaction of portrayed training level and checklist use, and interaction of milestone/competency and checklist use.

We used the interactions to assess whether the effects of checklist use varied by portrayed training level or milestone competency. For each of the portrayed training levels, the odds ratio associated with checklist use was presented with 95% confidence interval (CI). We used a likelihood ratio (LR) "chunk" test to assess the significance of explanatory variables and their interactions. We omitted nonsignificant interactions from the final regression model. Using postassessment survey data, a 95% CI (Wilson score methods) was created for the proportion of participants who felt that the checklist aided them in picking the appropriate milestone. We generated interrater reliability statistics for the 13 checklist items using Fleiss' kappa statistic to determine the degree of agreement in each checklist item as a measure of interrater reliability.[10]

What was known and gap
Faculty are tasked with making accurate milestone-based assessments, but may lack training and appropriate tools.

What is new
A study randomizing anesthesiology faculty to a milestone-based assessment versus one using a checklist finds the checklist superior for assessing entry level and junior learners.

Limitations
Small sample reduces generalizability; video objective structured clinical examination lacks validity evidence.

Bottom line
A checklist may be an appropriate tool for assessing performance early in training, but not with more advanced learners.

Results

Twenty faculty members (18% of 110 total anesthesiology faculty at the 3 institutions) participated in the study. Five of the 20 faculty members previously had assessed a learner in an OSCE, 11 had been OSCE participants, and 7 had prior training in the use of milestones. When asked about the difficulty of using the milestones as a tool for OSCE assessment, 13 of 20 felt that it was difficult or very difficult. Of that group using the checklist as a tool, only 4 of 10 found it difficult and 6 of 10 (95% CI 3.1–8.3) felt that it aided them to pick the appropriate milestones (provided as online supplemental material).

TABLE 1
Counts of Observed Checklist Items for Faculty Who Used a Checklist

Item	Entry Level	Junior Level	Midlevel	Senior Level	Advanced Level
1. Introduces self and health care role to patient	6/4/0[a]	**10/0/0**	**10/0/0**	**10/0/0**	**10/0/0**
2. Presents reason for the visit	8/2/0	8/1/1	**10/0/0**	**10/0/0**	9/1/0
3. Communicates routine information effectively (explains the procedure and events surrounding difficult intubation)	4/6/0	6/4/0	9/1/0	**10/0/0**	**10/0/0**
4. Uses language appropriate for patient's educational level and cultural context (does not talk down to patient or use excess medical terminology)	3/7/0	5/5/0	8/2/0	9/1/0	**10/0/0**
5. Shows sensitivity and respect for patient's concerns (makes eye contact, sits to talk)	1/9/0	8/2/0	**10/0/0**	**10/0/0**	**10/0/0**
6. Demonstrates listening to patient's needs and concerns (does not interrupt patient)	8/2/0	**10/0/0**	**10/0/0**	**10/0/0**	**10/0/0**
7. Acknowledges conflict (patient's frustration) and responds appropriately	4/6/0	7/3/0	**10/0/0**	**10/0/0**	**10/0/0**
8. Answers patient's questions directly	2/8/0	7/2/1	9/1/0	**10/0/0**	**10/0/0**
9. Recognizes when to involve/defer to supervisor	4/2/4	6/0/4	3/2/5	8/0/2	5/0/5
10. Instructs patient in related safety issues (emphasizes importance of patient awareness with difficult airway for future care; teaches patient about airway issues)	0/10/0	5/4/1	9/1/0	8/2/0	**10/0/0**
11. Checks for understanding (does more than ask, "Do you have any questions?")	5/4/1	7/1/2	8/0/2	8/1/1	8/1/1
12. Includes the patient or family with developing appropriate plan for follow-up	0/7/3	9/1/0	8/1/1	**10/0/0**	**10/0/0**
13. Coordinates care within the health care system (makes appropriate referrals [dentistry and/or risk management])	0/9/1	7/2/1	**10/0/0**	**10/0/0**	**10/0/0**

Note: TABLE 1 presents the counts of observed checklist items for the 10 faculty members who used a checklist in addition to the milestone tool. Note that the number of observations rated as *adequate* increased as the performance level increased (see bold entries).
[a] Read the numbers listed by each item as *adequate/inadequate/did not observe*. For example, "6/4/0" should be read as "6 adequate/4 inadequate/0 did not observe."

TABLE 1 shows the counts of observed checklist items (*adequate*, *inadequate*, *did not observe*) for the 10 faculty members who used the checklist. As the portrayed level of performance increased, the number of observations rated as *adequate* increased (see numbers in boldface [and shading] in TABLE 1). Participants scored 2 items (*recognizes when to involve/defer to supervisor* and *checks for understanding*) more often as *inadequate* or *did not observe*. Interrater agreement (TABLE 2) for 3 items showed substantial agreement, 2 items showed moderate agreement, and 5 items showed fair or slight agreement, beyond the level of agreement that is expected due to chance. Three items did not show any agreement.

We classified video ratings as either correct or incorrect according to level of training portrayed in the video. Averaging across all video performances (ie, ignoring the interaction of checklist use and video training level), the odds of correct classification were greater by a factor of 1.4 (95% Cl 1.0–2.0) when the checklist was used. However, there was significant evidence of interaction between checklist use and video training level (LR test P value $< .001$); the effectiveness of checklist use was inversely related to the training level portrayed in the video. The improvement in classification was largest for the entry level performance. TABLE 3 lists the odds ratio of correct classification associated with checklist use, stratified by video training level. For example, the odds of correctly classifying the entry level video were increased by a factor of 9.2 (95% CI 4.0–21.2) with checklist use. The milestone category being rated (eg, patient care, professionalism) was not significantly associated with the odds of correct classification (LR test P value $= .35$), nor was there evidence of an interaction with checklist use (LR test P value $= .68$). TABLE 4 lists the percentage (count) of correct video ratings, stratified by training level and checklist use.

TABLE 2
Interrater Reliability on Checklist Items

Checklist Item	Adequate/ Inadequate/ Did Not Observe	Fleiss Kappa	*P* Value
1. Introduces self and health care role to patient	46/4/0[a]	0.27	< .001
2. Presents reason for the visit	45/4/1	–0.018	.74
3. Communicates routine information effectively (explains the procedure and events surrounding difficult intubation)	39/11/0	0.26	< .001
4. Uses language appropriate for patient's educational level and cultural context (does not talk down to patient or use excess medical terminology)	35/15/0	0.24	< .001
5. Shows sensitivity and respect for patient's concerns (makes eye contact, sits to talk)	39/11/0	0.67	< .001
6. Demonstrates listening to patient's needs and concerns (does not interrupt patient)	48/2/0	0.07	.27
7. Acknowledges conflict (patient's frustration) and responds appropriately	41/9/0	0.32	< .001
8. Answers patient's questions directly	38/11/1	0.42	< .001
9. Recognizes when to involve/defer to supervisor	24/4/20	–0.07	.30
10. Instructs patient in related safety issues (emphasizes importance of patient awareness with difficult airway for future care; teaches patient about airway issues)	32/17/1	0.49	< .001
11. Checks for understanding (does more than ask, "Do you have any questions?")	36/7/7	–0.024	.62
12. Includes the patient or family with developing appropriate plan for follow-up	36/9/4	0.61	< .001
13. Coordinates care within the health care system (makes appropriate referrals [dentistry and/or risk management])	36/11/2	0.61	< .001

[a] Read the numbers listed by each item as *adequate/inadequate/did not observe*. For example, "46/4/0" should be read as "46 adequate/4 inadequate/0 did not observe."

Despite randomization, there was a degree of imbalance across study groups in years of experience and prior experience in assessing a learner in an OSCE. To address the possibility of chance confounding by these factors, we performed a sensitivity analysis in which we additionally adjusted for the years of experience category and prior OSCE assessment experience on the odds of correctly rating the portrayed performance level, as well as their interaction with checklist use. Although experience in OSCE assessment was found to be positively associated with the odds of correct video rating, the effect of checklist use was robust after adjustment for these factors.

Discussion

This study demonstrates a faculty development exercise designed to compare the use of milestones alone as a tool and the use of a checklist, the conventional tool. Although we assumed that the faculty with an itemized checklist would choose the correct milestone more often, this was only true for the performances portrayed at the entry and junior

TABLE 3
Odds Ratio of Correct Classification Associated With Checklist Use, Stratified by Video Training Level

Training Level	Odds Ratio	95% Confidence Interval
Entry	9.2	(4.0–21.2)
Junior	2.7	(1.3–5.7)
Midlevel	1.7	(0.8–3.4)
Senior	0.5	(0.3–1.0)
Advanced	0.3	(0.1–0.7)

TABLE 4
Percentage (Count) of Correct Video Ratings, Stratified by Training Level and Checklist Status

Training Level	Checklist	No Checklist
Entry	84 (59)	39 (27)
Junior	44 (31)	23 (16)
Midlevel	46 (32)	33 (23)
Senior	44 (31)	60 (42)
Advanced	16 (11)	39 (27)

levels. In all other cases, the use of the checklist added no advantage.

Use of a checklist is the most common method to assess OSCE performance.[11,12] Others have noted that it is more difficult to observe expertise using an OSCE examination, especially with a binary scale (adequate/inadequate) checklist,[13,14] and there is evidence that it is best to use global assessments or entrustable professional activities (EPAs) when working with more advanced learners.[15–18] Videos of standardized performances by trainees have been used in other studies for the purpose of setting standards, training faculty, and determining reliability of assessments.[19–21] Although we developed 5 videos for this study, for subsequent faculty and resident training sessions involving time constraints, we used 1 junior and 1 advanced video for assessment and discussion with positive results. The faculty participants in this study commented that viewing the video performances and assessing with the tools provided was an effective introduction to the milestone concept and performance assessment.

There are limitations to this study, including its small sample, which may reduce generalizability to faculty who did not participate. The videos were not piloted in advance.

We will be creating new OSCE stations based on EPAs and milestones that are difficult to assess, giving trainees and faculty live opportunities to practice and assess using the actual subcompetency milestones as the assessment tool. Videos of individual performances will be used for classroom use and standard setting. The goal of this research is to improve the quality of faculty assessment of trainees in actual clinical care.[22,23]

Conclusion

In this study, faculty members were able to accurately assign milestones in most cases to a video performance. A checklist aided the assessment of entry level and junior resident performers. Global or EPA-based assessments may be more effective for more advanced trainees.

References

1. Nasca TJ, Philibert I, Brigham T, et al. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366(11):1051–1056.
2. Leep Hunderfund AN, Reed DA, Starr SR, et al. Ways to write a milestone: approaches to operationalizing the development of competence in graduate medical education. *Acad Med*. March 28, 2017. Epub ahead of print.
3. Schartel SA, Kuhn C, Culley DJ, et al. Development of the anesthesiology educational milestones. *J Grad Med Educ*. 2014;6(1 suppl 1):12–14.
4. Meade LB, Borden SH, McArdle P, et al. From theory to actual practice: creation and application of milestones in an internal medicine residency program, 2004–2010. *Med Teach*. 2012;34(9):717–723.
5. Sukalich S, Elliott JO, Ruffner G. Teaching medical error disclosure to residents using patient-centered simulation training. *Acad Med*. 2014;89(1):136–143.
6. Stroud L, McIlroy J, Levinson W. Skills of internal medicine residents in disclosing medical errors: a study using standardized patients. *Acad Med*. 2009;84(12):1803–1808.
7. Gallagher TH, Studdert D, Levinson W. Disclosing harmful medical errors to patients. *N Engl J Med*. 2007;356(26):2713–2719.
8. Hobgood C, Peck CR, Gilbert B, et al. Medical errors—what and when: what do patients want to know? *Acad Emerg Med*. 2002;9(11):1156–1161.
9. Mazor KM, Simon SR, Gurwitz JH. Communicating with patients about medical errors: a review of the literature. *Arch Intern Med*. 2004;164(15):1690–1697.
10. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174.
11. Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Med Educ*. 2003;37(11):1012–1016.
12. Regehr G, Macrae H, Reznick RK, et al. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med*. 1998;73(9):993–997.
13. Hodges B, Regehr G, Mcnaughton N, et al. OSCE checklists do not capture increasing levels of expertise. *Acad Med*. 1999;74(10):1129–1134.
14. Hodges B, McNaughton N, Regehr G, et al. The challenge of creating new OSCE measures to capture the characteristics of expertise. *Med Educ*. 2002;36(8):742–748.
15. Prideaux D. The emperor's new wardrobe: the whole and the sum of the parts in curriculum design. *Med Educ*. 2016;50(1):10–12.
16. Touchie C, ten Cate O. The promise, perils, problems and progress of competency-based medical education. *Med Educ*. 2016;50(1):93–100.
17. ten Cate O. Nuts and bolts of entrustable professional activities. *J Grad Med Educ*. 2013;5(1):157–158.
18. Sibbald M, De Bruin AB, Van Merrienboer JJ. Finding and fixing mistakes: do checklists work for clinicians with different levels of experience? *Adv Health Sci Educ*. 2014;19(1):43–51.
19. 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.
20. Kane KE, Weaver KR, Barr GC Jr, et al. Standardized direct observation assessment tool: using a training video. *J Emerg Med*. 2017;52(4):530–537.
21. Shayne P, Gallahue F, Rinnert S, et al. Reliability of a core competency checklist assessment in the emergency department: the standardized direct observation assessment tool. *Acad Emerg Med*. 2006;13(7):727–732.
22. Carraccio C, Burke AE. Beyond competencies and milestones: adding meaning through context. *J Grad Med Educ*. 2010;2(3):419–422.
23. 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.

L. Jane Easdown, MD, MHPE, is Associate Professor of Anesthesiology, Vanderbilt University School of Medicine, Department of Anesthesiology, Vanderbilt University Medical Center; **Marsha L. Wakefield, MD, MHPE,** is Associate Professor, Department of Anesthesiology, University of Alabama at Birmingham School of Medicine; **Matthew S. Shotwell, PhD,** is Assistant Professor of Biostatistics, Vanderbilt University School of Medicine, Department of Biostatistics, Vanderbilt University Medical Center; and **Michael R. Sandison, MD,** is Professor and Vice Chair for Education, Department of Anesthesiology, Albany Medical Center.

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

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

These results were presented at the Society for Education in Anesthesiology 29th Spring Meeting, Boston, Massachusetts, May 30–June 1, 2014, and at Gerald S. Gotterer Health Professions Education Research Day, Vanderbilt University Medical Center, Nashville, Tennessee, September 22, 2014.

The authors would like to thank Ms Martha Tanner for her editorial assistance and Dr Mark Rice for his guidance in preparing the manuscript.

Corresponding author: L. Jane Easdown, MD, MHPE, Vanderbilt University Medical Center, Department of Anesthesiology, Division of Neuroanesthesiology, 4648 The Vanderbilt Clinic, 1301 Medical Center Drive, Nashville, TN 37232-5614, 615.343.9419, fax 615.936.6493, jane.easdown@vanderbilt.edu

Received February 14, 2017; revision received May 10, 2017; accepted May 31, 2017.