A Faculty Development Program to Reduce Rater Error on Milestone-Based Assessments

Jaya M. Raj, MD, FACP
Patti M. Thorn, PhD

Abstract

Background Rater errors, such as halo/reverse halo, range restriction, and leniency errors, are frequently cited as threats to the validity of resident assessment by faculty.

Objective We studied whether participation in faculty development on the use of a new Milestone-based assessment tool reduced rater error for participants compared to individuals who did not participate.

Methods We reviewed evaluations of resident Milestones completed by faculty at the end of rotations between July 2012 and June 2013. The 2 Milestones in each competency with the greatest number of ratings were selected for analysis.

Results A total of 412 evaluations were analyzed, including 217 completed by faculty who participated in the development activity, and 240 completed by nonparticipant faculty. All evaluations that contained identical scores for all Milestones (16%) were completed by nonparticipant faculty ($\chi^2 = 37.498$, $P < .001$). Faculty who had participated in development assigned a wider range of scores and lower minimum scores to residents, and provided the highest ratings for residents less frequently ($P < .001$) than nonparticipants.

Conclusions Faculty who participated in education about the Milestones demonstrated significantly less halo, range restriction, and leniency errors than faculty members who did not participate. These findings support a recommendation to develop a cadre of "core faculty" by training them in the use of Milestone assessment tools, and making them responsible for a significant portion of resident assessments.

Introduction

Competency-based assessment of resident performance requires faculty who are prepared to provide these assessments. Studies have shown that end-of-rotation evaluations completed by faculty lack reliability and fail to identify important deficiencies in residents' performance.[1,2] A number of factors account for the less-than-acceptable reliability, including variations in clinical context or patient complexity,[3] lack of longitudinal experiences with residents and patients,[4] and rater error. Types of rater error include leniency error, in which residents' abilities are rated higher than performance merits; halo or reverse halo, when performance in 1 domain affects ratings in other domains; range restriction, in which ratings fall within a narrow range; and error of undifferentiation, in which raters fail to differentiate among different performance domains.[5–8]

In 2009 a Milestone Task Force of the Accreditation Council for Graduate Medical Education and American Board of Internal Medicine (ABIM) published its draft Milestone document, consisting of a list of 142 "curricular" Milestones (which are distinct from the 22 internal medicine reporting Milestones).[9,10] These curricular Milestones represent specific, observable skills in the 6 competencies residents should demonstrate at each level of training during the 3 years of training. Although program directors have developed new assessment tools based on the Milestones, little is known about how best to educate and support faculty in using the new tools. To date, research has not examined the effect of faculty development in the use of a Milestone-based evaluation form. We sought to determine whether faculty development would reduce common rating errors among faculty completing Milestone-based end-of-rotation evaluations. Our hypothesis was that faculty who participated in faculty development

Jaya M. Raj, MD, FACP, is Associate Professor of Medicine, Creighton University School of Medicine, and Residency Program Director, Department of Internal Medicine, St. Joseph's Hospital and Medical Center; and **Patti M. Thorn, PhD,** is Assistant Professor, Creighton University School of Medicine, and Residency Education Specialist, Department of Internal Medicine, St. Joseph's Hospital and Medical Center.

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 Pamela DiSalvo for her assistance with study design and statistical analysis and Andrea-Darby Stewart, MD, for her role in planning and facilitating the faculty development workshop.

Corresponding author: Jaya M. Raj, MD, FACP, St. Joseph's Hospital and Medical Center, 500 W Thomas Road, Suite 900, Phoenix, AZ 85013, 602.406.8798, jaya.raj@dignityhealth.org

Received February 23, 2014; revision received June 4, 2014; accepted June 30, 2014.

DOI: http://dx.doi.org/10.4300/JGME-D-14-00161.1

would demonstrate reduced halo error, range restriction error, and leniency error as compared to faculty who did not participate.

What was known

Common rater errors reduce the reliability and validity of faculty assessments of residents.

What is new

Participation in faculty development focused on Milestone assessments reduced halo, range restriction, and leniency errors.

Limitations

Single program sample limits generalizability; retrospective cohort design reduces ability to attribute improved performance to the intervention versus other attributes of participants.

Bottom line

Training faculty raters is key to good Milestone assessments. Programs may benefit from identifying key assessment faculty and providing specific training for that role.

Methods

Participants

The study was conducted in an internal medicine residency program at a university-affiliated, community-based academic health center between July 2012 and June 2013. Subjects included all 35 residents (27 categorical residents and 8 preliminary postgraduate year (PGY)–1 residents), and 60 faculty members who evaluated the residents during the study period. Faculty demographics are listed in TABLE 1. Three faculty members left during the study period. Resident composition remained stable.

Assessment Tool

At the start of academic year 2012–2013, the program implemented new rating forms, based on the 142 curricular Milestones, to be used by faculty to provide end-of-rotation evaluations for the residents. To promote faculty ownership of the new tool, several key clinical faculty in each area (inpatient, outpatient, and subspecialty) were asked to select Milestones they thought were most relevant and important to assess during their rotations. The resulting list of Milestones was used to create evaluation forms for inpatient, outpatient, critical care, and subspecialty rotations. Of the 142 Milestones, 106 Milestones representing all 6 core competencies were used across the 8 different evaluation forms. Residents were rated on a 9-point Likert scale, which corresponded to the progression of "Novice" to "Expert" in the Dreyfus model of skills acquisition.[11] A 9-point scale was chosen to reflect the continuum of development on the path to expertise. Descriptive anchors were provided at each end, and at the midpoints of the scale. On this scale, a rating of 1 to 4 indicated a resident was not yet competent, a rating of 5 or 6 signified a resident demonstrated competence, a score of 7 meant a resident was proficient, and a rating of 8 or 9 indicated a resident was functioning above the level of proficiency.

Faculty Development

Three months before implementation of the new evaluation forms, the program offered faculty development on the value and use of Milestone-based education and assessment. The faculty development program included 2 components of rater training described by Holmboe and Hawkins[12]: "performance dimension training" to develop necessary knowledge of the Milestones, and "frame of

TABLE 1 **Faculty Demographics**

	Faculty Development Participants, No. (%)	Faculty Development Nonparticipants, No. (%)
Men	7 (38.9)	33 (78.6)
Women	11 (61.1)	9 (21.4)
Academic hospitalists	5 (27.8)	1 (2.4)
Ambulatory faculty	8 (44.4)	4 (9.5)
IM subspecialists	5 (27.8)	20 (47.6)
Other specialties[a]	0	17 (40.5)
Key clinical faculty	18 (100)	25 (59.5)
Employed by IM department	18 (100)	0
Total	18	42

Abbreviation: IM, internal medicine.

[a] Other specialties included radiology; neurology; dermatology; ear, nose, and throat; ophthalmology; family medicine; and sports medicine.

TABLE 2 **Milestones Included in Analysis**[a]

Milestone	
PC-C1	Synthesize all available data, including interview, physical examination, and preliminary laboratory data, to define each patient's central clinical problem
PC-C2	Develop prioritized differential diagnoses, evidence-based diagnostic and therapeutic plan for common inpatient and ambulatory conditions
MK-A7	Demonstrate sufficient knowledge to evaluate complex or rare medical conditions and multiple coexistent conditions
MK-B2	Understand indications for and have basic skills in interpreting more advanced diagnostic tests
PBLI-B1	Identify learning needs (clinical questions) as they emerge in patient care activities
PBLI-F1	Respond welcomingly and productively to feedback from all members of the health care team, including faculty, peer residents, students, nurses, allied health workers, patients, and their advocates
ICS-A2	Effectively use verbal and nonverbal skills to create rapport with patients/families
ICS-D3	Engage in collaborative communication with all members of the health care team
P-B1	Demonstrate empathy and compassion to all patients
P-F1	Dress and behave appropriately
SBP-B2	Work effectively as a member within the interprofessional team to ensure safe patient care
SBP-E2	Minimize unnecessary care, including tests, procedures, therapies, and ambulatory or hospital encounters

Abbreviations: PC, patient care; MK, medical knowledge; PBLI, practice-based learning and improvement; ICS, interpersonal and communication skills; P, professionalism; SBP, systems-based practice.

[a] http://www.im.org/p/cm/ld/fid=561.

reference training'' to define behavioral examples associated with different levels of performance. Three stand-alone performance dimension training sessions included (1) two 30-minute didactic sessions during which the program director distributed and explained the Milestones and led a discussion at regularly scheduled faculty meetings, and (2) a 1½-hour interactive workshop held jointly with the Department of Family Medicine, in which participants practiced Milestone-based assessment by rating the performance of a PGY-1 resident in a scripted video simulation (developed by J.M.R.) presenting a chest pain admission to an attending physician. The use of videotaped scenarios in faculty development has been described in the literature as an effective teaching tool.[1,13] Frame-of-reference training was integrated and distributed among curriculum meetings, annual program review, and clinical competency committee meetings, during which examples and consistency of observable behaviors related to the Milestones were discussed and related to descriptive anchors on the rating scale. To be designated as a ''faculty development participant,'' an individual needed to attend at least 1 of the performance dimension training and 1 of the frame-of-reference training sessions.

Of the 60 faculty members who completed evaluations, 18 met the criteria to be called ''faculty development participants.'' Starting in July 2012, all faculty members were required to complete the Milestone-based evaluation forms electronically at the end of the rotation and provide the resident with face-to-face feedback.

The study was reviewed by the Institutional Review Board and deemed exempt as a quality assurance project.

Data Collection and Analysis

Electronic evaluation forms were distributed and completed online by using Residency Management Suite software (New Innovations Inc). Data were downloaded in tabular form, imported into Microsoft Excel, and recoded for analysis by using SPSS statistical software (IBM Corp). Pairwise comparisons by Milestone were tabulated by evaluator status. Statistical differences were tested by using χ^2 analyses. Analysis of variance (1-way ANOVA) was performed on evaluator mean rating scores. Two Milestones in each competency with the greatest number of ratings were selected for analysis; these were the Milestones that appeared most frequently across all evaluation forms. In total, a sample of 12 Milestones were analyzed, all of which were rated by evaluators at least 250 times (TABLE 2).

Results

Between July 2012 and June 2013, a total of 487 evaluations were completed and returned by faculty for the 35 residents in the program. Evaluations missing greater than 25% of the requested data (1% of evaluations in the

TABLE 3 **Percentage of Evaluations in Which Residents Received the Same Score for All Milestones Within and Across Competencies (Halo Error)**

	Percentage of Evaluations in Which All Questions in This Competency Received the Same Numerical Answer		χ^2
Competency	**Faculty Development Participants, %**	**Faculty Development Nonparticipants, %**	
Patient care	22	38	12.500[a]
Medical knowledge	57	72	10.37[a]
PBLI	36	51	9.445[b]
ICS	30	33	0.415
Professionalism	43	57	8.707[b]
Systems-based practice	55	51	0.699
All competencies	0	16	37.498[a]

Abbreviations: PBLI, practice-based learning and improvement; ICS, interpersonal and communication skills.

[a] Statistically significant, $P \leq .001$.
[b] Statistically significant, $P < .01$.

faculty development participant group and 18% in the nonparticipant group) were excluded from analysis.

Of the 412 evaluations analyzed (217 completed by faculty development participants and 240 by nonparticipants), 66 (16%) contained identical ratings for all the Milestones across all competencies (ie, the same rating for every item on the evaluation form). All of these were completed by faculty who had not participated in faculty development (TABLE 3).

In 8 of the 12 Milestones analyzed, the range of scores in evaluations completed by faculty who had participated in faculty development was wider and the minimum score was lower than the ratings by nonparticipant faculty. ANOVA found significant differences in mean rating scores between the participant and nonparticipant groups for all 12 Milestones. Evaluators who participated in faculty development assigned residents scores of 8 or 9 less frequently than nonparticipant faculty across all the Milestones analyzed (TABLE 4).

Discussion

Our study demonstrated clear differences in the way faculty development participants and nonparticipants rated resident performance on Milestones, with participants less likely than nonparticipants to give a resident identical ratings for all Milestones across all competencies (reduced halo error). In addition, the faculty development participants assigned a wider range of scores to residents (less range restriction), including a minimum score of 2 compared to 3 in the nonparticipant group. Participant faculty members were less likely to give residents ratings of 8 or 9 (reduced leniency error). Our findings are consistent with other research that shows rater training results in changes in how faculty rate residents when using the ABIM global rating form[13] and the Mini Clinical Evaluation Exercise (mini-CEX).[14] Our findings are inconsistent with research that reported no improvement in interrater reliability or accuracy of mini-CEX scores after training.[15] Our literature search did not identify any studies that specifically examined the effect of faculty development on reducing rater error in Milestone-based evaluations. A study by Ginsburg et al[16] used structured faculty interviews and analyzed the findings by using grounded theory to elucidate how faculty conceptualize residents' performance in different domains. Similar methods could be applied to determine how faculty conceptualize the Milestones and use them to assess resident performance.

Our study has several limitations, including its single program sample and use of a retrospective cohort design. There likely were differences between the 2 groups at baseline (ie, a selection bias) that could have caused or exaggerated the differences observed. In addition, faculty development participants were a self-selected group of physicians employed by an academic institution, who were thereby expected to engage in continuous learning, adopt new methods of teaching and assessment, and adapt to changing educational paradigms. They were more likely than nonparticipants to be involved in multiple aspects of the residency program, and potentially were more invested in the program, suggesting the reduced rater error observed for participant faculty may be due to immersion in the culture of resident education and assessment. The relative

TABLE 4 **Scoring Behavior for Specific Milestones (Range Restriction and Leniency Error)**

Milestone	Range of Scores Residents Received in Individual Milestones		ANOVA	Percentage of Evaluations in Which Residents Received a Rating of 8 or 9		Percentage Difference	χ^2
	Evaluations Completed by Faculty Development Participants	Evaluations Completed by Nonparticipants		Evaluations Completed by Faculty Development Participants	Evaluations Completed by Nonparticipants		
PC-C1	3–9	3–9	$F_{(1,385)} = 18.505$[a]	16	32	16	13.392[b]
PC-C2	2–9	4–9	$F_{(1,385)} = 22.838$[a]	15	30	15	12.277[b]
MK-A7	3–9	3–9	$F_{(1,248)} = 26.633$[a]	18	34	16	8.365[a]
MK-B2	2–9	4–9	$F_{(1,248)} = 26.633$[a]	19	40	21	14.021[b]
PBLI-B1	2–9	4–9	$F_{(1,248)} = 26.633$[a]	13	35	22	27.714[b]
PBLI-F1	3–9	4–9	$F_{(1,248)} = 26.633$[a]	22	43	21	18.042[b]
ICS-A2	3–9	4–9	$F_{(1,248)} = 26.633$[a]	22	48	26	30.821[b]
ICS-D3	3–9	4–9	$F_{(1,359)} = 34.318$[a]	22	44	22	20.765[b]
P-B1	4–9	4–9	$F_{(1,410)} = 19.329$[a]	36	52	16	10.754[b]
P-F1	4–9	4–9	$F_{(1,410)} = 19.329$[a]	41	54	13	7.126[a]
SBP-B2	2–9	4–9	$F_{(1,353)} = 30.124$[a]	17	43	26	27.455[b]
SBP-E2	2–9	4–9	$F_{(1,273)} = 45.087$[a]	14	41	27	25.674[b]

Abbreviations: ANOVA, analysis of variance; PC, patient care; MK, medical knowledge; PBLI, practice-based learning and improvement; ICS, interpersonal and communication skills; P, professionalism; SBP, systems-based practice.

[a] Statistically significant, $P < .01$.
[b] Statistically significant, $P \leq .001$.

role of faculty development versus residency program immersion in determining rater behavior is a topic for future research.

Conclusion

We conclude that it is not enough to develop better assessment tools; we also need to select and educate evaluators who will use the tools with the intent and sophistication they warrant. It may be unrealistic to expect all faculty members who teach and assess residents to participate in faculty development, and 70% of our faculty did not participate. We recommend that programs identify a targeted cadre of faculty committed to learn and apply new methods of performance rating, and make them responsible for the major portion of resident assessments. Our findings also support an appeal for institutions to direct resources to faculty development, and for organizations to promote longitudinal faculty development programs on a national scale.

References

1 Herbers JE Jr, Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ. How accurate are faculty evaluations of clinical competence? *J Gen Intern Med*. 1989;4(3):202–208.
2 Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? *Acad Med*. 2004;79(5):453–457.
3 Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Acad Med*. 2011;86(suppl 10):1–7.
4 Holmboe ES, Ward DS, Reznick RK, Katsufrakis PJ, Leslie KM, Patel VL, et al. Faculty development in assessment: the missing link in competency-based medical education. *Acad Med*. 2011;86(4):460–467.
5 Thorndike EL. A constant error in psychology ratings. *J Appl Psychol*. 1920;4:25–29.
6 Ryan JG, Mandel FS, Sama A, Ward MF. Reliability of faculty clinical evaluations of non-emergency medicine residents during emergency department rotations. *Acad Emerg Med*. 1996;3(12):1124–1130.
7 Silber CG, Nasca TJ, Paskin DL, Eiger G, Robeson M, Veloski JJ. Do global rating forms enable program directors to assess the ACGME competencies? *Acad Med*. 2004;79(6):549–556.
8 Thomas MR, Beckman TJ, Mauck KF, Cha SS, Thomas KG. Group assessments of resident physicians improve reliability and decrease halo error. *J Gen Intern Med*. 2011;26(7):759–764.
9 Green ML, Aagaard EM, Caverzagie KJ, Chick DA, Holmboe E, Kane G, et al. Charting the road to competence: developmental milestones for internal medicine residency training. *J Grad Med Educ*. 2009;1(1):5–20.
10 Iobst W, Aagaard E, Bazari H, Brigham T, Bush RW, Caverzagie K, et al. Internal medicine milestones. *J Grad Med Educ*. 2013;5(1 suppl 1):14–23.
11 Dreyfus SE, Dreyfus HL. *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*. Berkeley, CA: University of California, Operations Research Center; 1980.
12 Holmboe E, Hawkins R. Improving the accuracy of raters: direct observation workshop. http://dev.im.org/toolbox/FacultyDev/GIMFDP/Documents/Strategies%20Dallas%20Holmboe%20Huot.doc. Accessed July 9, 2014.
13 Holmboe ES, Fiebach NH, Galaty LA, Huot S. Effectiveness of a focused educational intervention on resident evaluations from faculty a randomized controlled trial. *J Gen Intern Med*. 2001;16(7):427–434.
14 Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 2004;140(11):874–881.
15 Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. *J Gen Intern Med*. 2009;24(1):74–79.
16 Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Acad Med*. 2010;85(5):780–786.