Feasibility and Reliability of a Multisource Feedback Tool for Emergency Medicine Residents

Gregory Garra, DO
Andrew Wackett, MD
Henry Thode, PhD

Abstract

Background While the Accreditation Council for Graduate Medical Education recommends multisource feedback (MSF) of resident performance, there is no uniformly accepted MSF tool for emergency medicine (EM) trainees, and the process of obtaining MSF in EM residencies is untested.

Objective To determine the feasibility of an MSF program and evaluate the intraclass and interclass correlation of a previously reported resident professionalism evaluation, the Humanism Scale (HS).

Methods To assess 10 third-year EM residents, we distributed an anonymous 9-item modified HS (EM-HS) to emergency department nursing staff, faculty physicians, and patients. The evaluators rated resident performance on a 1 to 9 scale (needs improvement to outstanding). Residents were asked to complete a self-evaluation of performance, using the same scale.

Analysis Generalizability coefficients ($E\rho^2$) were used to assess the reliability within evaluator classes. The mean score for each of the 9 questions provided by each evaluator class was calculated for each resident. Correlation coefficients were used to evaluate correlation between rater classes for each question on the EM-HS. $E\rho^2$ and correlation values greater than 0.70 were deemed acceptable.

Results EM-HSs were obtained from 44 nurses and 12 faculty physicians. The residents had an average of 13 evaluations by emergency department patients. Reliability within faculty and nurses was acceptable, with $E\rho^2$ of 0.79 and 0.83, respectively. Interclass reliability was good between faculty and nurses.

Conclusions An MSF program for EM residents is feasible. Intraclass reliability was acceptable for faculty and nurses. However, reliable feedback from patients requires a larger number of patient evaluations.

Editor's Note: The online version of this article contains the paper version of the Emergency Medicine Humanism Scale.

Introduction

In 1997, the Accreditation Council for Graduate Medical Education (ACGME) embarked on a long-term initiative known as the Outcome Project.[1] The purpose was to place increased emphasis on educational outcome assessment in the accreditation process. As part of this process, the ACGME identified a set of 6 competencies every resident should achieve before graduation.

In 2006, the ACGME required programs to assess residents in all competency domains, using at least 1 assessment tool in addition to standard global/end-of-rotation evaluations.[2] Tools were expected to be valid and reliable measures of competency-based learning objectives, focusing on a few objectives that naturally occur during the resident-patient-supervisor encounter.

Multisource feedback (MSF) involving nonphysician evaluators and patients is an ACGME-recommended assessment method.[3] Multisource feedback instruments are completed by multiple observers within a person's sphere of influence and are valuable in assessing specific behaviors or skills, especially interpersonal communication skills (ICS) and professionalism (Prof).[4] However, for MSF to be considered a reliable component of a resident's performance portfolio, it must demonstrate consistent results regardless of when it is used, who uses it, and which item is assessed.

Use of the MSF in emergency medicine (EM) programs is largely untested. The only study examining MSF in EM resident training focused solely on nurses as evaluators.[5] Our goal was to determine the feasibility and reliability of MSF in an EM training program.

All authors are at Stony Brook University Medical Center. **Gregory Garra, DO,** is Emergency Medicine Program Director; **Andrew Wackett, MD,** is Assistant Dean for Undergraduate Medical Education; and **Henry Thode, PhD,** is Emergency Medicine Biostatistician.

Funding: The authors report no external funding source.

Presented at the 2010 American College of Emergency Medicine Research Forum, September 28, 2010, Las Vegas, NV.

Corresponding author: Gregory Garra, DO, Stony Brook University Medical Center, Department of Emergency Medicine, Stony Brook, NY 11794, 631.444.3880, Gregory.Garra@StonyBrook.edu

Received September 8, 2010; revision received March 17, 2011; accepted April 24, 2011.

DOI: http://dx.doi.org/10.4300/JGME-D-10-00173.1

Methods

Our study was a prospective questionnaire-based assessment of ICS and Prof. It was conducted at a suburban, university-based EM residency program. Residents in their final year of training (EM-3) were surveyed to ascertain the feasibility of implementing an MSF program and to determine the reliability of the questionnaire.

Survey Content and Administration

A previously validated MSF instrument, the Humanism Scale[6] (HS), was used to obtain feedback on ICS and Prof. The original HS was reviewed by the residency administration and revised to avoid rater misinterpretation. The original HS question, ''Cooperates with paramedical staff,'' was refined into 2 questions: ''Cooperates with nurses'' and ''Cooperates with ancillary medical staff.'' The modified HS was pilot tested on different raters (faculty, nurses, and patients) and further revised upon feedback before full implementation. The final MSF instrument, the emergency medicine Humanism Scale (EM-HS), consisted of 9 questions with ratings on a 9-point continuum from ''needs improvement'' to ''outstanding.'' Electronic evaluation forms were distributed to faculty and nurses. Paper forms were generated for self-evaluations and patient evaluations. Both versions of the EM-HS included the resident's name and photo. The EM-HS provided to each of the rater groups was identical except that questions 1 to 3 were eliminated from the patient EM-HS.

Residents were informed at our annual orientation conference that performance feedback would be obtained from patients, staff, and faculty. Self-evaluations were collected before faculty, nurse, and patient data collection. The EM-HS was distributed to all EM nurses and faculty via e-mail list serve with a link to the electronic survey (Survey Monkey, Palo Alto, CA). Evaluators were informed that the anonymously completed evaluation would be included in the resident performance portfolio and reviewed/discussed with each resident. The survey was open for 21 days and a weekly e-mail reminder was sent via the faculty and nurses list serve.

Paper copies of the EM-HS were generated for patient evaluations. Patients were approached by the research personnel at the conclusion of an emergency department encounter and asked to complete an anonymous evaluation of the resident physician involved in their care. Surveys were collected from those patients for whom emergency department care, in its entirety, was provided by a single EM resident.

Data Analysis

All data were entered into SPSS version 18 (SPSS, Inc, Armonk, NY). Exploratory factor analyses were performed within each evaluator class to identify the dimensionality of the questions. Generalizability coefficients[7] ($E\rho^2$) were calculated to evaluate consistency within the attending and nurse evaluator classes, with residents and raters as random facets and survey items as fixed facets; a nested design was necessary for patient evaluations. Generalizability coefficients could not be calculated for self-evaluations since there was only 1 per resident. The average of the total scores provided by each evaluator class was calculated for each resident. Pearson correlations (r) were used to assess correlation between pairs of evaluator classes, using ''resident'' as the unit of analysis. Values of $E\rho^2$ and $r > 0.70$ were deemed acceptable.

The study was granted a waiver of consent by the institution's Institutional Review Board.

Results

Requests to evaluate residents were sent to 16 faculty members and 65 nurses. Complete evaluations on all residents were obtained from 12 faculty members (75%) and 44 nurses (68%). Nine of 10 EM-3 residents (90%) completed a self-evaluation. Patient evaluations were obtained on 9 residents. Complete evaluations were available from 119 patients, or an average of 13 per resident.

The mean score for each question per rater group is shown in TABLE 1. The $E\rho^2$ for faculty was 0.79. The variance components for calculating $E\rho^2$ are shown in TABLE 2. Nurses showed similar reliability, with an $E\rho^2$ of 0.83, while patients demonstrated fair reliability with $E\rho^2$ of 0.59.

The factor analyses resulted in only a single factor for each evaluator class. Owing to the homogeneity of responses by question, correlations on an item-by-item basis were deemed unnecessary and between-rater reliability was assessed with correlations, using mean resident scores.

The interrater reliability of nurses and faculty for the mean score on all 9 questions was fair; $r = 0.62$ (TABLE 3). The correlation of faculty and nurses with resident self-assessment was negative but not significant. Comparison of patient scores with faculty, nurses, and resident self-assessment was based upon the mean score of questions 4 through 9. Correlations between patients and the other evaluator classes were positive but poor and not statistically significant.

Discussion

We found that using an MSF tool to assess EM resident ICS and Prof is feasible. Typically, development of an MSF tool is a time-consuming process requiring psychometric design and testing[4] and a pilot period to determine reliability and establish validity.[8] We used an assessment tool that had documented reliability and validity. In our experience, the electronic and paper survey setup was not time-consuming. Collection of data from nurses and faculty yielded a reasonable response rate with little more effort than sending out weekly reminders. Collection of data from patients proved to be a more arduous task, requiring assistance

TABLE 1 **Mean (Standard Deviation) Rater Scores by Question**

	Score, Mean (SD)		
Question	**Faculty**	**Nurses**	**Patients**
Ability to cooperate with medical colleagues	7.5 (1.1)	7.5 (1.3)	N/A
Ability to cooperate with nurses	7.4 (1.2)	7.3 (1.5)	N/A
Ability to cooperate with ancillary medical staff	7.3 (1.2)	7.3 (1.5)	N/A
Quality of physician-patient relationship	7.4 (1.1)	7.3 (1.5)	8.2 (1.2)
Ability to render comfort and empathy	7.3 (1.3)	7.1 (1.6)	8.1 (1.3)
Involvement of patient in decision making	7.2 (1.3)	7.1 (1.5)	8.1 (1.2)
Consideration of patients' concerns	7.3 (1.3)	7.0 (1.6)	8.1 (1.2)
Ability to place patients at ease	7.3 (1.2)	6.9 (1.6)	8.1 (1.2)
Ability to admit one's own errors	7.3 (1.3)	6.9 (1.7)	8.2 (1.2)

Abbreviation: N/A, not applicable.

from trained research personnel to obtain the requisite number of patient evaluations for each of the 10 EM-3 residents. Studies suggest that acceptable reliability for the instrument and individual physician can be obtained by 8 to 10 coworkers, 8 to 10 medical colleagues, and 25 patients.[9] We used 12 faculty members, 44 nurses, and approximately 13 patients to rate our residents. The results of our generalizability study indicate that MSF evaluations from 8 faculty members and 21 nurses would have provided sufficient reliability ($E\rho^2 > 0.70$). Multisource feedback evaluations from 22 patients per resident would be needed to attain minimum reliability.

TABLE 2 **Variance Components of the Generalizability Study of Faculty Raters**

Source of Variation	Variance, %
Resident	15.8
Faculty	17.5
Item	0[a]
Resident-faculty	45.5
Resident-item	1.8
Faculty-item	1.8
Error	17.7

[a] Fixed factor, contributing variance = 0.

Our results suggest that the EM-HS is a reliable tool for assessing ICS and Prof of EM residents. Adequate intrarater reliability was demonstrated. Other studies have yielded similar findings. Joshi et al[10] demonstrated excellent intrarater correlation for all but patient raters of obstetrics-gynecology residents. Wood et al[11] demonstrated high internal consistency for all raters of radiology residents but low correlations between rater groups. In contrast, Weigelt et al[12] reported that 360-degree evaluations of surgical residents provided no new information not already available from faculty evaluations. Our finding of poor interrater correlation provides further support for incorporating multiple perspectives into feedback systems. Overemphasizing agreement between raters misses one of the primary purposes of MSF: to obtain different perspectives concerning performance.

Limitations

As with all questionnaire-based studies, multiple limitations apply to our findings. Although our study used a previous validated survey, we made a minor modification to the original HS. As such, it is a new questionnaire and requires strategic development. We attempted to minimize instrument bias by following methodology for survey research.[13] Despite our attempts to limit instrument bias, it

TABLE 3 **Pearson Correlations of Mean Total Resident Scores Between Evaluator Classes**

	Correlation of Total Score (Questions 1–9)	Correlation of Total Score (Questions 4–9)
Faculty-nurse	0.62	0.69
Faculty-resident	−0.08	0.01
Faculty-patient	0[a]	0.13
Nurse-resident	−0.35	−0.29
Nurse-patient	0[a]	0.13
Resident-patient	0[a]	0.21

[a] Fixed factor, contributing variance = 0.

is possible that the questions and/or response scales were misunderstood by raters.

Multisource feedback is especially subject to multiple memory biases such as halo effect,[14] context effect, mood congruent memory bias,[15] and distinctive encoding.[16] Preconceptions about a resident's prior performance or previous experiences are known to influence judgment about a person or situation. We did not train raters before assessment implementation nor did we seek to determine if the rater had a recent positive or negative experience with the trainee.

We used 2 different methods to obtain MSF: paper and web based. It is unknown whether responses collected via one method are psychometrically equivalent to responses collected by a different method. There are few published studies empirically assessing measurement equivalence between web-based and paper assessment methods. It is unknown whether direct comparisons of the different MSF collection methods will result in equivalent outcomes.

Our study was not designed to assess validation of the EM-HS. Validation is a research process demonstrating that an instrument actually measures what it was designed to measure. Although Linn et al[6] demonstrated reliability and validity of the HS for assessment of internal medicine residents, we cannot assert that our modified, albeit very similar, instrument maintains the same validity.

Finally, this was a single-center study of EM residents, conducted in a suburban tertiary care facility. Therefore, generalizability to other settings or specialty programs is limited.

Conclusion

The EM-HS is relatively easy to administer and provides a reliable assessment of EM resident ICS and Prof from a minimum of 8 faculty members, 21 nurses, and 22 patients. Further studies are necessary to determine validity of the EM-HS, frequency of administration, and its effect on performance.

References

1 Accreditation Council for Graduate Medical Education. ACGME Outcome Project. http://www.acgme.org/outcome/. Accessed May 7, 2010.

2 Accreditation Council for Graduate Medical Education. ACGME Outcome Project: toolbox of assessment methods. http://www.acgme.org/Outcome/assess/Toolbox.pdf. Accessed May 7, 2010.

3 Accreditation Council for Graduate Medical Education. ACGME Program Requirements for Graduate Medical Education in Emergency Medicine. http://www.acgme.org/acWebsite/downloads/RRC_progReq/110emergencymed07012007.pdf. Accessed March 8, 2011.

4 Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof*. 2003;23(1):4–12.

5 Tintinalli JE. Evaluation of emergency medicine residents by nurses. *Acad Med*. 1989;64(1):49–50.

6 Linn LS, DiMatteo MR, Cope DW, Robbins A. Measuring physicians' humanistic attitudes, values, and behaviors. *Med Care*. 1987;25(6):504–513.

7 Brennan RL *Generalizability Theory*. New York, NY: Springer-Verlag; 2001.

8 Rodgers KG, Manifold C. 360-degree feedback: possibilities for assessment of the ACGME core competencies for emergency medicine residents. *Acad Emerg Med*. 2002;9(11):1300–1304.

9 Ramsey PC, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA*. 1993;269(13):1655–1660.

10 Joshi R, Ling FW, Jaeger J. Assessment of a 360-degree instrument to evaluate resident's competency in interpersonal and communication skills. *Acad Med*. 2004;79(5):458–463.

11 Wood J, Collins J, Burnside ES, et al. Patient, faculty, and self-assessment of radiology resident performance: a 360-degree method of measuring professionalism and interpersonal/communication skills. *Acad Radiol*. 2004;11(8):931–939.

12 Weigelt JA, Brasel KJ, Bragg D, Simpson D. The 360-degree evaluation: increased work with little return? *Curr Surg*. 2004;61(6):616–628.

13 Totten VY, Panacek EA, Price D. Basics of survey research (part 14), survey research methodology: designing the survey instrument. *Air Med J*. 1999;18(1):26–34.

14 Thorndike EL. A constant error in psychological ratings. *J Appl Psychol*. 1920;4:25–29.

15 Rholes WS, Riskind JH, Lane JW. Emotional states and memory biases: effects of cognitive priming and mood. *J Pers Soc Psychol*. 1987;52:91–99.

16 Mantyla T. Recollection of faces: remembering differences and knowing similarities. *J Exp Psychol Learn Mem Cogn*. 1997;23(5):1203–1216.