The Relationship Between Faculty Performance Assessment and Results on the In-Training Examination for Residents in an Emergency Medicine Training Program

James G. Ryan, MD
David Barlas, MD
Simcha Pollack, PhD

Abstract

Background Medical knowledge (MK) in residents is commonly assessed by the in-training examination (ITE) and faculty evaluations of resident performance.

Objective We assessed the reliability of clinical evaluations of residents by faculty and the relationship between faculty assessments of resident performance and ITE scores.

Methods We conducted a cross-sectional, observational study at an academic emergency department with a postgraduate year (PGY)-1 to PGY-3 emergency medicine residency program, comparing summative, quarterly, faculty evaluation data for MK and overall clinical competency (OC) with annual ITE scores, accounting for PGY level. We also assessed the reliability of faculty evaluations using a random effects, intraclass correlation analysis.

Results We analyzed data for 59 emergency medicine residents during a 6-year period. Faculty evaluations of MK and OC were highly reliable ($\kappa = 0.99$) and remained reliable after stratification by year of training (mean $\kappa = 0.68–0.84$). Assessments of resident performance (MK and OC) and the ITE increased with PGY level. The MK and OC results had high correlations with PGY level, and ITE scores correlated moderately with PGY. The OC and MK results had a moderate correlation with ITE score. When residents were grouped by PGY level, there was no significant correlation between MK as assessed by the faculty and the ITE score.

Conclusions Resident clinical performance and ITE scores both increase with resident PGY level, but ITE scores do not predict resident clinical performance compared with peers at their PGY level.

Introduction

Faculty assessment of clinical performance is a frequently used method for assessing competencies and is required by nearly all Residency Review Committees.[1] Most programs also administer annual, in-training examinations (ITEs), designed to measure each resident's medical knowledge (MK). Despite the nearly universal use of these 2 methods, little research has been done to assess the relationship of the data derived from these different methods of evaluation.

The ITEs have been shown to predict resident performance on future specialty certifying examinations,[2] yet the literature shows a poor correlation between ITE scores and resident clinical performance.[3–8] To date, no study assessing the relationship between ITE performance and faculty evaluations has been documented to be reliable. Because emergency medicine (EM) residents routinely work closely with several different attending evaluators, they offered a unique opportunity to assess the interobserver and overall reliability of their clinical evaluations. If faculty evaluations prove reliable, but yield results divergent from the ITE results, the likely reason is that the 2 evaluation methods measure different constructs.

The goal of this investigation was to assess the reliability of faculty evaluations and to determine the relationship between faculty's assessment of resident performance and residents' ITE scores. In addition, we planned to determine whether those relationships changed

James G. Ryan, MD, is Assistant Professor of Emergency Medicine in Clinical Medicine, Weill School of Medicine, Cornell University, and Residency Director of Emergency Medicine, New York Hospital Queens; **David Barlas, MD,** is Assistant Professor of Emergency Medicine in Clinical Medicine, Weill School of Medicine, Cornell University, and Associate Residency Director of Emergency Medicine, New York Hospital Queens; and **Simcha Pollack, PhD,** is Professor of Computer Information Systems and Decision Sciences, Tobin College of Business, St. John's University.

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

Corresponding author: David Barlas, MD, Department of Emergency Medicine, New York Hospital Queens, 56-45 Main Street, Flushing, NY 11355, 718.670.1426, dab9044@nyp.org

Received August 16, 2012; revision received January 5, 2013; accepted February 25, 2013.

DOI: http://dx.doi.org/10.4300/JGME-D-12-00240.1

when that data were stratified by postgraduate year (PGY) level.

What was known

Residents' medical knowledge is commonly assessed by the in-training examination (ITE) and faculty evaluations of resident performance.

What is new

Faculty evaluations and ITE scores increase with residents' postgraduate year level, and are moderately correlated.

Limitations

Single-program study and small sample size may limit generalizability.

Bottom line

Faculty assessment of resident medical knowledge may represent a construct that is distinct and separate from the "medical knowledge" assessed by the ITE.

Methods

Our study was conducted at New York Hospital Queens, an urban emergency department that sponsors an Accreditation Council for Graduate Medical Education (ACGME)–accredited PGY-1 through PGY-3 training program approved for 10 residents per year. During the data collection period, residents were supervised by 51 board-certified or board-eligible EM faculty members.

The program administers the ITE annually to all residents. Evaluations and ITE data for a 6-year period (2005–2010) were included in this study. In our program, faculty members routinely complete anonymous, online, global, quarterly evaluations for each resident. The evaluations consisted of 10 questions designed to assess each of the 6 competencies as well as each resident's overall clinical competency (OC). The evaluation used an anchored 9-point Likert scale to rate resident performance. A score of 1 corresponded to the faculty member's opinion that the resident's performance was at the level of a medical student and a score of 9 corresponded to the expected level of performance for an attending EM physician. Only the scores for each resident's "fund of medical knowledge" and "overall clinical competence" from the third academic quarter (January through March) were used for the analysis because they coincided with the administration of the ITE in February. Each resident's raw score on the ITE was used as the measure of MK.

For each resident, we collected 4 data points: (1) the MK score, (2) the OC score, (3) the ITE score, and (4) the year of training. The MK score for each resident was the mean score across all evaluators who evaluated the resident's "fund of medical knowledge," and the OC score was the mean score across all evaluators who evaluated the resident's "overall clinical competence." The ITE score was the raw score obtained by an individual resident. The PGY was the PGY level of training for the resident at the time of those assessments.

The study was reviewed and granted exemption status by the Institutional Review Board of New York Hospital Queens.

Data analysis was performed using SAS 9.3 (SAS Institute, Cary, NC). To assess the reliability of the faculty evaluation process, we performed a random-effects model, intraclass correlation for the variables of MK and OC. This test is used to assess agreement between raters when raters who evaluate the subjects are randomly assigned from a larger pool.[9] Correlation between faculty assessment of MK, OC, PGY, and ITE scores was reported using Pearson correlation coefficients. Each of those analyses was first performed across all residents and then performed with the residents stratified by PGY level. A P value $< .05$ was considered significant.

Results

During the 6-year study period, 51 faculty members completed 1912 evaluations on 59 residents. The data set included 140 composite, third-quarter evaluations, with most residents having evaluations for multiple years of training. A mean of 13.7 (SD $\pm$ 2.9) faculty members evaluated each resident during that period. There were 12 circumstances in which the ITE scores were not available, leaving 128 complete sets of resident observations for data analysis. No residents repeated any year of training during the study period.

The random-effects, intraclass correlation analysis revealed that the faculty evaluation process was highly reliable (MK mean $\kappa = 0.99$ and OC mean $\kappa = 0.99$). We also grouped the residents by PGY level and repeated the analysis to remove any potential evaluator bias leading to falsely elevated reliabilities from evaluator knowledge of the residents' year of training. That analysis again revealed high reliabilities for both MK (PGY-1 mean $\kappa = 0.68$; PGY-2 mean $\kappa = 0.76$; PGY-3 mean $\kappa = 0.84$) and OC factors (PGY-1 mean $\kappa = 0.70$; PGY-2 mean $\kappa = 0.73$; PGY-3 mean $\kappa = 0.81$).

The mean scores for the ITE, MK, and OC increased significantly with year of training (TABLE 1). The ITE scores had more overlap across year of training than did the MK assessed by faculty evaluations (FIGURES 1 and 2). When correlation analyses were performed across all PGY levels, MK and OC had very high correlations with PGY level (MK $r = 0.97$, $P < .001$; OC $r = 0.97$, $P < .001$), whereas the ITE score correlated moderately with PGY level ($r = .60$, $P < .001$). Assessment of the relationship

TABLE 1 **Mean Score (±SD) of Residents on the In-Training Examination (ITE) and From Third-Quarter Faculty Evaluations**

	Year 1	Year 2	Year 3	*P* Value
EM ITE score	65.2 (±7.0)	73.2 (±6.5)	77.5 (±6.7)	<.001
Faculty evaluation score				
Medical knowledge	3.0 (±0.60)	5.8 (±0.52)	8.4 (±0.51)	<.001
Overall clinical competence	3.1 (±0.63)	5.8 (±0.45)	8.3 (±0.46)	<.001

Abbreviation: EM, emergency medicine.

between the ITE, MK, and OC scores showed that faculty assessments of MK correlated strongly with OC score ($r = 0.99$, $P < .001$) and moderately with the ITE score ($r = 0.61$, $P < .001$; TABLE 2).

Because both the examination score and faculty assessment scores increased with year of training, we also assessed the relationship of those variables by calculating Pearson correlation coefficients within each PGY level. Faculty assessment of MK did not have a significant correlation with ITE scores at any PGY level (PGY-1 $r = 0.06$, $P = .70$; PGY-2 $r = 0.09$, $P = .56$; PGY-3 $r = 0.38$, $P = .06$). The correlation between MK and OC remained very high (PGY-1 $r = 0.94$, $P < .001$; PGY-2 $r = 0.91$, $P < .001$; PGY-3 $r = 0.97$, $P < .001$).

Discussion

Our study is the first, to our knowledge, to evaluate the relationship between ITE scores and faculty evaluations of resident clinical performance using faculty evaluations that were assessed for reliability. We found that faculty evaluations and ITE scores increased with resident PGY level and were moderately correlated. However, stratification by PGY level showed that ITE scores could not be used to predict resident clinical performance compared with peers at the same PGY level. Any relationship between these 2 factors was due to covariance with PGY level of training.

Several prior studies have found a low to moderate correlation between the results of examinations and faculty assessment of MK.[3–8] One study found that improved prediction of resident ITE scores by faculty occurred for residents in internal medicine with increased levels of training (and faculty exposure).[10] Studies of radiology and pediatrics residents have had mixed results, with some authors reporting significant improvement in correlation with advancement,[4,11,12] and others reporting no change.[5,8] The divergent results found in prior studies may be

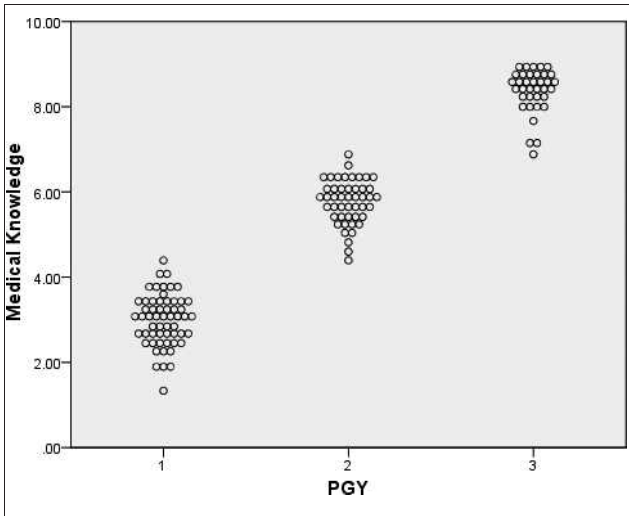

FIGURE 1 **Change With Postgraduate Year (PGY) for Faculty Assessment of Medical Knowledge**

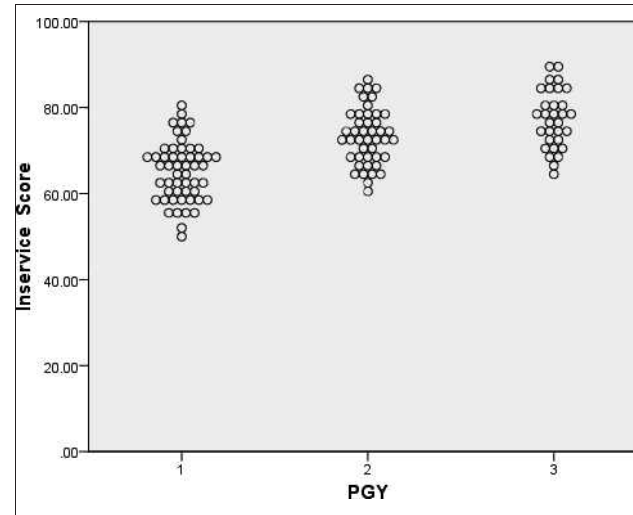

FIGURE 2 **Change with Postgraduate Year (PGY) for In-Training Examination Score**

TABLE 2 **Correlation Matrix of Postgraduate Year (PGY), Overall Clinical Competency (OC), Medical Knowledge (MK), and In-Training Examination (ITE) Score**

	PGY	OC	MK	ITE Score
PGY				
Pearson correlation	1	0.969[a]	0.969[a]	0.595[a]
Significance (2-tailed)		0.000	0.000	0.000
N	140	140	140	128
OC				
Pearson correlation	0.969[a]	1	0.996[a]	0.608[a]
Significance (2-tailed)	0.000		0.000	0.000
N	140	140	140	128
MK				
Pearson correlation	0.969[a]	0.996[a]	1	0.605[a]
Significance (2-tailed)	0.000	0.000		0.000
N	140	140	140	128
ITE score				
Pearson correlation	0.595[a]	0.608[a]	0.605[a]	1
Significance (2-tailed)	0.000	0.000	0.000	
N	128	128	128	128

[a] Correlation is significant at the 0.01 level (2-tailed).

explained by the covariance of the ITE score and the faculty assessment of resident skills with PGY level. In our study, we observed a moderate correlation between the ITE and faculty assessments of MK when the analysis was performed across all years of training, which disappeared when the analysis was repeated with the residents grouped by PGY level. Studies that correlate the data across all years of training may find a moderate relationship because of this covariance, yet when the analysis is performed within a given year of training, the relationship is largely absent.

The rating scale we used to evaluate our residents may have contributed to the moderate correlation found between ITE score and clinical performance when all PGY levels were included. Our program uses a behavior-anchored rating scale that rates resident performance from 1 to 9 based on the expected progression from medical student to attending physician. As expected with a scale of this type, resident evaluation scores increase with PGY level. Little research has been done to assess rating scales in resident clinical evaluation, but our prior study[13] demonstrated that this type of scale provided greater reliability when compared with a traditional Likert scale (poor to excellent). Two other studies also found that behavior-anchored rating scales provided highly reliable results of operative skills in surgery and obstetrics-gynecology residents.[14,15] However, this type of scale may artificially increase reliability because evaluators may assign resident scores to certain ranges based on PGY level. Concerned about this potential evaluator bias, we repeated the reliability analysis with residents grouped by class to remove the effect of PGY level on the results. The faculty evaluations remained highly reliable for both MK and OC.

There are several other potential explanations for the poor relationship between ITE scores and faculty assessments. First, the significant correlation between the MK and OC assessment by faculty suggests that faculty may not discriminate between the various dimensions of clinical competence in their evaluation of residents. Similar findings have been reported in internal medicine,[16,17] with the potential that faculty evaluations of MK may not be as "pure" or precise as we thought.

Second, and related, is that a clinical manifestation of the MK competency includes much more than a simple factual recall. Summative evaluations by faculty may also include interpretative skills (ie, translating knowledge to a clinical scenario), rapid recall (ie, using a critical care algorithm from memory), disease pattern recognition, clinical teaching proficiency when working with junior

residents, and other observations. Additionally, presentation and documentation skills, preparation, and even attitude all may contribute to perceived MK competency. The ITE remains objective and focused, possibly resulting in a differing interpretation of a resident's MK competency, and it is possible that the cognitive knowledge assessed by the ITE may not directly apply to performance in a busy emergency department. In clinical practice, most emergency care is related to a limited number of common complaints and diseases. Absence of the broader knowledge assessed by the ITE may not be obvious to faculty if a resident is proficient in the management of commonly encountered conditions.

There are several limitations of our study. Our study was performed at a single program that uses selected items from the ACGME toolbox for resident evaluation. In addition, the generalizability of our findings is limited by the small sample size, and the weak and nonsignificant correlation we observed when the data was grouped by year of training may be a result of type II (β) error. However, even if those correlations were significant with a larger sample size, the coefficient of determination would be very small, suggesting that ITE service scores and faculty evaluation of resident performance have little, if any, relationship. Finally, our study did not seek to determine which assessment method was the better predictor of clinical performance after completion of residency. Further study into which assessment method correlates best with practice performance is needed.

Conclusion

Faculty assessment of EM residents' clinical performance is highly reliable and correlates moderately with EM residents' ITE score when measured across different years of training. However, when assessed within a given year, ITE scores do not predict resident clinical performance compared with peers at their PGY level. Faculty assessment of resident MK may represent a construct that is distinct and separate from the "medical knowledge" assessed by the ITE.

References

1 Accreditation Council for Graduate Medical Education. ACGME program requirements for graduate medical education in emergency medicine. http://www.acgme.org/acgmeweb/Portals/0/PFAssets/2013-PR-FAQ-PIF/110_emergency_medicine_07012013.pdf. Accessed September 18, 2013.

2 American Board of Emergency Medicine. In-training examination overview. https://www.abem.org/public/emergency-medicine-training/in-training-examination/in-training-examination-overview. Accessed August 7, 2013.

3 Lertkhachonsuk R, Wangsaturaka D. Correlation of residents' knowledge and clinical performance with their teaching skill and attitude. *J Med Assoc Thai*. 2009;92(8):995–998.

4 Wise S, Stagg PL, Szucs R, Gay S, Mauger D, Hartman D. Assessment of resident knowledge: subjective assessment versus performance on the ACR in-training examination. *Acad Radiol*. 1999;6(1):66–71.

5 Nuovo J, Bertakis KD, Azari R. Assessing resident's knowledge and communication skills using four different evaluation tools. *Med Educ*. 2006;40(7):630–636.

6 Tabuenca A, Welling R, Sachdeva AK, Blair PG, Horvath K, Tarpley J, et al. Multi-institutional validation of a web-based core competency assessment system. *J Surg Educ*. 2007;64(6):390–394.

7 Kolars JC, McDonald FS, Subhiyah RG, Edson RS. Knowledge base evaluation of medicine residents on the gastroenterology service: implications for competency assessments by faculty. *Clin Gastroenterol Hepatol*. 2003;1(1):64–68.

8 Quattlebaum TG, Darden PM, Sperry JB. In-training examinations as predictors of resident clinical performance. *Pediatrics*. 1989;84(1):165–172.

9 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420–428.

10 Hawkins RE, Sumption KF, Gaglione MM, Holmboe ES. The in-training examination in internal medicine: resident perceptions and lack of correlation between resident scores and faculty predictions of resident performance. *Am J Med*. 1999;106(2):206–210.

11 Adusumilli S, Cohan RH, Korobkin M, Fitzgerald JT, Oh MS. Correlation between radiology resident rotation performance and examination scores. *Acad Radiol*. 2000;7(11):920–926.

12 Althouse LA, McGuinness GA. The in-training examination: an analysis of its predictive value on performance on the general pediatrics certification examination. *J Pediatr*. 2008;153(3):425–428.

13 Ryan JG, Madden JF, Sama AE, Risucci D, LaMantia J, Ward MF. The interrater reliability of faculty clinical evaluations of emergency medicine residents comparing a fixed reference with a relative reference performance scoring system [Abstract 301]. *Acad Emerg Med*. 1997;4(5):441.

14 Chou B, Bowen CW, Handa VL. Evaluating the competency of gynecology residents in the operating room: validation of a new assessment tool. *Am J Obstet Gynecol*. 2008;199(5):571.e1–e5.

15 Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery*. 2005;138(4):640–647.

16 Haber RJ, Avins AL. Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? *J Gen Intern Med*. 1994;9(3):140–145.

17 Thompson WG, Lipkin M Jr, Gilbert DA, Guzzo RA, Roberson L. Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. *J Gen Intern Med*. 1990;5(3):214–217.