Original Article

SAGE Open Medicine

SAGE Open Medicine
3: 2050312115589648

DOI: 10.1177/2050312115589648
smo.sagepub.com
SAGE

The validity and reliability of attending evaluations of medicine residents

Jeffrey L Jackson[1], Cynthia Kay[2] and Michael Frank[3]

Abstract

Objectives: To assess the reliability and validity of faculty evaluations of medicine residents.

Methods: We conducted a retrospective study (2004–2012) involving 228 internal medicine residency graduates at the Medical College of Wisconsin who were evaluated by 334 attendings. Measures included evaluations of residents by attendings, based on six competencies and interns and residents' performance on the American Board of Internal Medicine certification exam and annual in-service training examination. All residents had at least one in-service training examination result and 80% allowed the American Board of Internal Medicine to release their scores.

Results: Attending evaluations had good consistency (Cronbach's α = 0.96). There was poor construct validity with modest inter-rater reliability and evidence that attendings were rating residents on a single factor rather than the six competencies intended to be measured. There was poor predictive validity as attending ratings correlated weakly with performance on the in-service training examination or American Board of Internal Medicine certification exam.

Conclusion: We conclude that attending evaluations are poor measures for assessing progress toward competency. It may be time to move beyond evaluations that rely on global, end-of-rotation appraisals.

Keywords

Education, residents, evaluation

Date received: 18 March 2015; accepted: 11 May 2015

In the United States, to become board-certified in internal medicine, candidates must complete an accredited 3-year program, be certified by their program director as competent and qualified, and pass a written examination. While residency program directors have a variety of tools to evaluate trainees, including annual in-service training examinations (ITEs), medical record audits, clinical evaluation exercises, peer review, 360 review, and standardized patients,[1–3] monthly attending evaluations are the most common evaluation approach.[3]

The accreditation of US residency programs and the evaluation of residents have changed over time. Before 1999, internal medicine residency programs evaluated their residents using the American Board of Internal Medicine Evaluation Form (ABIM-EF) that consisted of 23 questions intended to assess resident performance in seven domains. We and several others showed that this rating system lacked predictive ability, poorly predicting residence performance on licensing examinations.[4–8] In addition, we demonstrated that attending evaluations lacked validity, by rating internal medicine residents in just two domains (knowledge and professionalism), rather than on the seven domains intended to be assessed.[8] In 1999, the Accreditation Council for Graduate Medical Education (ACGME) revised its accreditation process by defining six general competencies (patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice) that served as the basis for accrediting all US residency programs.[9] When the core competencies were first released in 1999, there was considerable confusion among attendings about what some of them meant, particularly systems-based practice and practice-based learning. The ACGME rolled out this system in phases, with the

[1]C, GIM Section, Zablocki VAMC, Department of Medicine, Medical College of Wisconsin, Milwaukee, WI, USA
[2]Department of Medicine, Medical College of Wisconsin, Milwaukee, WI, USA
[3]Program Director, Internal Medicine Residency, Department of Medicine, Medical College of Wisconsin, Milwaukee, WI, USA

Corresponding author:
Jeffrey L Jackson, C, GIM Section, Zablocki VAMC, Department of Medicine, Medical College of Wisconsin, 5005 W Washington Blvd, Milwaukee, WI 53295, USA.
Email: jjackson@mcw.edu

second phase (that occurred between July 2002 and June 2006) focused on defining these core competencies and a national attending education program.[10]

A recent systematic review of global evaluation forms in the era of these ACGME competencies identified only five studies that addressed the issue of validity and reliability, none of which focused on internal medicine.[10] One study at two centers from multiple training programs found that attendings were evaluating residents on two factors (knowledge and interpersonal skills) rather than the six domains intended, similar to our findings for evaluation of medicine residents using the pre-1999 evaluation form.[11]

In July 2014, the ACGME further modified their evaluation process to a "Next Accreditation System" that retains the six core competencies, but assesses the progression of learners during their years of training as they advance from novice to proficient in the practice of medicine. This is done by measuring attainment of "milestones" along the path to competency and by switching from a numerical to a narrative-based evaluation system, with clear description of the milestones accomplished. Attendings are asked to compare residents not against each other but against a measure of how far they have progressed toward achieving competency. For most programs, it is anticipated that evaluation of resident progression in achieving milestones will continue to rely heavily on attending evaluations.[3] At the core of this evaluation system are the six competencies developed in 1999. Our study's goals are to determine whether (1) faculty can reliably assess six different ACGME competencies and (2) these assessments have predictive validity.

Methods

Study participants were Medical College of Wisconsin's (MCW) internal medicine residents who completed residency training from 2004 to 2012. Residents were evaluated at least monthly by their attendings. These evaluations followed ACGME guidelines and assessed residents in six competency domains (patient care, medical knowledge, interpersonal and communication skills, professionalism, practice-based learning and improvement, systems-based practice). Patient care had three questions (medical interviewing, physical examination, procedures); the remaining competencies were evaluated with single questions. Attending ratings were Likert-type, scaled from 1 to 9, anchored as 1 (unsatisfactory), 5 (satisfactory), and 9 (superior). Each question also included narrative text describing criteria at each end of the scale. For example, the question about medical interviewing was anchored at one end by "Often incomplete, superficial, by rote and not directed" and at the other end by "Always precise, logical, thorough, reliable, purposeful, efficient, suitably focused, specificity and clarity convey sophistication."

Implementation of evaluation in the six competencies was accompanied by faculty development efforts within the department. These included written descriptions, PowerPoint modules, and in-person presentations on principles of evaluation and effective feedback, although these sessions were poorly attended and it is unclear how many faculty actually completed the written or electronic modules. About halfway through the period of this study, faculty began receiving specific feedback on how they were rating residents in the competencies, including how their average ratings where compared with other faculty's and how on average they rated residents across the six competencies.

We also collected annual ITE results and ABIM certification exam scores for residents who allowed the ABIM to release their scores to MCW. We limited ABIM scores to the graduates' initial attempt. To assess construct validity, we assessed internal consistency of attending evaluations with the Cronbach's α, assessed for reliability of evaluations within each resident and within each attending using intraclass correlation coefficients,[12] and examined factor analysis of attending evaluations.[13] The number of factors retained was based on the Kaiser 1 rule and on Scree plots.[13] We assessed predictive validity by comparing faculty evaluations with ITE and ABIM certifying examination scores. To more closely mirror how program directors may use the ratings, we calculated average scores for each 6 months of training since directors conduct formal reviews with feedback to residents semi-annually. We explored the relationship between ratings, ITE scores, and ABIM certifying examination scores using generalized linear mixed models, either with the ABIM score as a continuous or dichotomous measure, with random intercepts for individual residents, and post-graduate year. This study was approved by our institutional review board (IRB) and all calculations were done using Stata (version 13.1; College Station, TX).

Results

Over the 8 years, there were 228 internal medicine resident graduates with 6603 evaluations by 334 attendings. Residents averaged 17.8 attending evaluations (range = 1–37) and the average attending provided 54.4 evaluations (range = 1–273). We had ITE results for all residents for at least 1 year and ABIM board scores on 183 (80.2%). There were no differences in ITE scores between residents who did or did not allow the ABIM to provide results ($p = 0.47$). Overall, 89% of residents passed the board examination, with scores ranging from 241 to 710 (mean = 466). This pass rate and scores were consistent over the study period.

Construct validity

Attending evaluations had good internal consistency (Cronbach's $\alpha = 0.96$). There was a stepwise increase in ratings as residents progressed through training (Table 1). Attendings tended to give similar ratings from resident to resident for each competency (Table 2), although the reliability of scores given between attendings for the same resident was low

Table 1. Attending average scores on monthly resident evaluations by year of training.

	PGY1 (SD)	PGY2	PGY3	p
Medical interviewing	7.12	7.50	7.68	<0.0005 for all
Patient care				
Physical examination	6.99	7.33	7.48	
Procedures	7.07	7.38	7.65	
Medical knowledge	6.96	7.33	7.59	
Practice-based learning	7.13	7.45	7.62	
Communication	7.51	7.76	7.87	
Professionalism	7.83	7.96	8.05	
Overall grade	7.15	7.57	7.72	

PGY: post-graduate year; SD: standard deviation.
All questions were Likert-type, scaled from 1 to 9 with 1 = poor and 9 = outstanding.

Table 2. Attending reliability scores.

ACGME question	Within attending[a]	Between attendings[a]
Medical interviewing	0.86	0.71
Physical examination	0.89	0.65
Medical knowledge	0.85	0.71
Practice-based learning	0.85	0.68
Interpersonal communication	0.81	0.74
Professionalism	0.83	0.69
Systems-based practice	0.86	0.67
Overall	0.83	0.74

ACGME: Accreditation Council for Graduate Medical Education.
[a]Intraclass correlation coefficient (>0.8 = very good, 0.5–0.8 = good, 0.2–0.5 = modest, <0.2 = poor).

(Table 2). Factor analysis suggested a single-factor solution rather than a 6-factor solution intended to be measured (Table 3). The eigenvalue for the first factor was 6.3 and for the second factor was 0.5. When reanalyzed by the year of evaluation, all 8 years had single-factor solutions. There was no evidence of improvement over time.

Predictive validity

Attending evaluations of medical knowledge was the only item associated with performance on either the ITE (medical knowledge: $\beta = 0.11$, 95% confidence interval (CI) = 0.02–0.20; overall: $\beta = 0.15$, 95% CI = 0.06–0.24) or ABIM certifying examination performance (medical knowledge: $\beta = 0.63$, 95% CI = 0.09–1.18; overall: $\beta = 0.59$, 95% CI = 0.04–1.14). These results suggest that for every one-point increase in attending rating of the learner's medical knowledge, there was a 0.11 increase in ITE scores and a 0.63 increase in scores on the ABIM certifying examination. This is a weak effect, confirmed by the fact that less than 5% of the variance in test performance was explained by attending ratings. Attending evaluations of a resident's medical knowledge also weakly predicted passing the ABIM certifying examination (odds ratio (OR) = 1.3, 95% CI = 1.1–1.4). When reanalyzed as 6-month average scores, medical knowledge was similarly correlated with both the ITE ($\beta = 5.9$, 95% CI = 5.3–8.5) and ABIM certifying exam performances ($\beta = 83.5$, 95% CI = 55.1–112.0). This suggests that a one-point increase in the 6-month average attending rating of knowledge was correlated with an 84-point increase in ABIM scores. This was a modest effect and 6-month average attending ratings of medical knowledge explained 25% of the variance in board scores. The relationship between attending evaluation of knowledge and ABIM scores was consistent over the 8 years; there was no evidence of improvement over time.

Table 3. Factor analysis of attending ratings of residents using the six ACGME competencies.

Variable	Factor 1 Eigenvalue = 6.3	Uniqueness
Medical interviewing	0.9144	0.1639
Physical examination	0.8891	0.2095
Medical knowledge	0.8737	0.2367
Interpersonal communication	0.8023	0.3564
Professionalism	0.8023	0.3564
Practice-based learning	0.8989	0.1920
Systems-based practice	0.8897	0.2084

ACGME: Accreditation Council for Graduate Medical Education.
Factor 2 had an eigenvalue of 0.55.

Discussion

We found that attending ratings of medicine residents were consistent and that ratings increased as residents progressed from year to year. While attending scores were consistent, there was only modest agreement between attendings for any given resident. We also found that attendings rated residents on a single domain rather than the intended six competencies. These findings suggest that this version of attending evaluations had poor construct validity. Attending evaluations also had poor predictive validity as their ratings of the resident's medical knowledge only weakly correlated with performance on ITE and board certification examinations. The 6-month average of medical knowledge scores did somewhat better, but less than 25% of the total variance in board examination performance was explained by the average rating of resident medical knowledge. Despite considerable local and national efforts[10] to educate attendings about the meaning of these competencies, our results found no improvement over time.

The ACGME introduced the "Next Accreditation System" in July 2014, in hopes to improve and strengthen attendings' evaluations of residents. Our results suggest that attending ratings show progressive improvement as residents go from internship to their senior year, consistent with attaining milestones. However, attendings are not discriminating among the six assessment competencies. Rather, it appears that the attendings are rating residents globally. Particularly, disturbing is a

weak correlation between attending evaluation of a resident's medical knowledge and their performance on either the annual ITE or the ABIM certifying exam. The ACGME has identified both ITE and ABIM certifying examination performances as measures against which they intend to assess the validity of resident assessment. Unfortunately, most residency programs rely heavily on attending evaluations to assess residents' growth and attainment of competence. There are a number of reasons for this, including cost and ease of administration. There is also residual belief in the face validity of evaluations from attendings who have worked closely with residents on rotations. Unfortunately, the literature to date and our data suggest that this belief is misplaced. It is not yet known whether adding narrative descriptions can improve predictive validity.

There are several limitations to our study including being a single site, a single type of residency, and a relatively small number of residents. Firm conclusions about the reliability and validity of resident assessment will need to be based on a broad spectrum of programs and a larger sample. It may be that attending evaluations are more accurate in some specialties than others. Second, the only outcome we measured was performance on ITE and ABIM certifying examinations. These are measures of knowledge and do not capture other important aspects of clinical practice, such as professionalism, communication, and patient outcomes. Third, 20% of our sample did not provide ABIM board scores. It is likely that these are not missing at random since learners anticipating doing poorly may be less likely to release their scores. However, the ITE scores were the same among residents who allowed scores to be provided was the same as for those who did not, and ITE scores have been shown to accurately predict residents at risk of failing their boards.[14] Finally, as of July 2014, the "Next Accreditation System" is being used to evaluate medicine residents. We have no data yet to show whether this third major revision in the evaluation system performs better than the previous two. Whether adding more narrative and more questions can help attendings discriminate between the six competencies remains to be seen.

Residency programs are charged with the important task of training the next generation of clinicians. Residencies have a professional obligation to track and to certify when trainees are sufficiently competent to practice independently. Graduate medical education programs receive considerable public funding. Based on these and previous results, if the ACGME is hoping to develop reliable and valid measures of resident accomplishment of milestones on the path to independence, measures other than attending evaluations will likely be necessary to achieve this goal. Our study suggests that it may be time to develop a different evaluation system than relying on a global-, end-of-rotation-, and competency-based evaluation form. The next step in resident evaluation is to develop ideas and studies that would help the medical education community evolve from resident ratings with poor reliability and validity evidence to ones better support the "Next Accreditation System."

Acknowledgements

All opinions expressed in this manuscript represent those of the authors and should not be construed to reflect, in any way, those of the Department of Veteran Health Affairs or the US Government.

Declaration of conflicting interests

The authors have no conflict of interest with this article.

Funding

The authors received no funding for this project.

References

1. Holmboe ES and Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. *Ann Intern Med* 1998; 129: 42–48.
2. Van Rosendaal GM and Jennett PA. Comparing peer and faculty evaluations in an internal medicine residency. *Acad Med* 1994; 69: 299–303.
3. Chaudhry SI, Holmboe E and Beasley BW. The state of evaluation in internal medicine residency. *J Gen Intern Med* 2008; 23: 1010–1015.
4. Brailovsky CA, Grand'Maison P and Lescop J. Residency directors' predictions of candidates' performances on a licensing examination. *Acad Med* 1995; 70: 410–414.
5. Haber RJ and 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: 140–145.
6. Norcini JJ, Webster GD, Grosso LJ, et al. Ratings of residents' clinical competence and performance on certification examination. *J Med Educ* 1987; 62: 457–462.
7. Thompson WG, Lipkin M Jr, Gilbert DA, et al. Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. *J Gen Intern Med* 1990; 5: 214–217.
8. Durning SJ, Cation LJ and Jackson JL. The reliability and validity of the American Board of Internal Medicine Monthly Evaluation Form. *Acad Med* 2003; 78: 1175–1182.
9. Batalden P, Leach D, Swing S, et al. General competencies and accreditation in graduate medical education. *Health Aff* 2002; 21: 103–111.
10. Lurie SJ, Mooney CJ and Lyness JM. Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: a systematic review. *Acad Med* 2009; 84: 301–309.
11. Silber CG, Nasca TJ, Paskin DL, et al. Do global rating forms enable program directors to assess the ACGME competencies? *Acad Med* 2004; 79: 549–556.
12. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res* 2005; 19: 231–240.
13. Kim JO and Mueller CW. *Factor analysis: statistical methods and practical issues*. Newbury Park, CA: SAGE, 1978.
14. Kay C, Jackson JL and Frank M. The relationship between internal medicine residency graduate performance on the ABIM certifying examination, yearly in-service training examinations, and the USMLE Step 1 examination. *Acad Med* 2015; 90: 100–104.