Validity evidence of resident competency ratings and the identification of problem residents

Yoon Soo Park, Janet Riddle & Ara Tekian

OBJECTIVES This study examined validity evidence of end-of-rotation evaluations used to measure progress toward mastery of core competencies in residents. In addition, this study investigated whether end-of-rotation evaluations can be used to detect problem residents during their training.

METHODS Historical data for a 4-year period (2009–2012), containing 4986 observations of 291 internal medicine residents, were examined. Residents were observed and assessed by fellows, faculty members and programme directors on nine domains, including the six Accreditation Council for Graduate Medical Education core competencies, as part of their end-of-rotation evaluations. Descriptive statistics were used to collect evidence of the response process. Correlations between competencies and a generalisability study were used to examine the internal structure of the end-of-rotation evaluations. Hierarchical regression was used to estimate the increase in scores across years of training. Scores on end-of-rotation evaluations were compared with trainees identified as problem residents by programme directors.

RESULTS Compared with fellows, faculty and programme directors had significantly greater variability in assigning scores across different competencies. Correlations between competencies ranged from 0.69 to 0.92. The reliability of end-of-rotation evaluations was adequate (fellows, phi coefficient [φ] = 0.68; faculty [including programme directors], φ = 0.71). Mean scores increased by 0.21 points (95% confidence interval 0.18–0.24) per postgraduate year. Mean scores were significantly correlated with classification as a problem resident ($r = 0.33$, $p < 0.001$); problem residents also had significantly lower ratings across all competencies during PGY-1 compared with all other residents.

CONCLUSIONS End-of-rotation evaluations are a useful method of measuring the growth in resident performance associated with core competencies when sufficient numbers of end-of-rotation evaluation scores are used. Furthermore, end-of-rotation evaluation scores provide preliminary evidence with which to detect and predict problem residents in subsequent postgraduate training years.

Medical Education 2014; 48: 614–622
doi: 10.1111/medu.12408

Discuss ideas arising from the article at "www.mededuc.com discuss"

Department of Medical Education, College of Medicine, University of Illinois at Chicago, Chicago, Illinois, USA

Correspondence: Yoon Soo Park, PhD, Department of Medical Education (MC 591), College of Medicine, University of Illinois at Chicago, 808 South Wood Street, 963 CMET, Chicago, Illinois 60612-7309, USA. Tel: 00 1 312 355 5406;
E-mail: yspark2@uic.edu

INTRODUCTION

Residency programmes in the USA are undergoing major transitions with respect to the training and assessing of residents. With the implementation of the Next Accreditation System (NAS) and the Milestone Project by the Accreditation Council for Graduate Medical Education (ACGME), ongoing data collection and trend analysis of resident performance are mandated on a biannual basis.[1] Given these changes in residency education, a validated assessment system for measuring residents' growth and progress toward the mastery of core competencies at different levels of training is required. A key feature of a valid assessment system is the ability to identify and monitor underperforming residents at early stages of training so that they can be given proper remediation.

Nearly all residency programmes have problem residents.[2] Generally, problem residents are trainees who demonstrate significant difficulties and require intervention by a person of authority.[3,4] Problem residents have also been viewed as learners who fail to acquire the necessary competencies or progress at a slower rate than other residents toward the acquisition of competencies.[5] The literature provides a wide array of remediation plans for problem residents.[6,7] These studies emphasise the early identification of and intervention with problem residents because the needs of these learners can lead to difficulties in their completion of residency and can require substantial resources for remediation.

According to a national survey conducted in 1999, 82% of programme directors reported to have discovered problem residents through observations; problem residents were reported to have difficulties in medical knowledge and in clinical judgement.[8] More recently, a 10-year review of resident records in Canada (July 1999 to June 2009) found medical expertise and professionalism to be two competencies in which problem residents were reported to have difficulties.[9] Although previous research describes the characteristics of problem residents and provides methods for identifying them, there is a lack of empirically driven studies providing the necessary data and results to support these findings.[10] For example, it is unclear whether end-of-rotation evaluations measuring core competencies can reproduce consistent scores representing growth in a learner's performance over time. End-of-rotation evaluations are global judgements completed by supervising faculty members or fellows and are based on a form on which core competencies during residency training are rated; they are similar to the In-Training Evaluation Reports (ITERs) used in Canada. End-of-rotation evaluations are not based on a specific observation or direct encounter, but, rather, on a prolonged observation of a resident throughout a rotation (typically one month in duration) and can include second-hand reports and case presentations, in addition to direct observations.[11,12] Although guidelines on the number of evaluations are emerging in the multi-source feedback literature,[13] it is unclear how many evaluations are required when measuring core competencies or whether evaluations provided by fellows or faculty, respectively, are more accurate and reliable. There is also a lack of empirical evidence on whether end-of-rotation evaluations can be used to identify problem residents during the early stages of residency training. End-of-rotation evaluations are based on several factors, including direct observations and multi-source feedback (discussions between the rater and the resident's peers, nurses and students in order to obtain their input); direct observations and multi-source feedback are assessment methods considered by ACGME in the Milestone Project. Therefore, there is an increasing and urgent need for studies that provide psychometric evidence to support the theoretically based frameworks cited in the literature.

End-of-rotation evaluations are a type of workplace-based assessment, which has been shown to be effective in changing the behaviour of learners through feedback.[14,15] However, concerns regarding end-of-rotation evaluations have also been raised because evaluation forms are not completed directly following an actual educational experience and thus may not provide meaningful feedback to the learner.[16] End-of-rotation evaluations have also been subject to the 'failure to fail' phenomenon, whereby raters become reluctant to provide accurate feedback on a resident's performance as a result of multiple social factors, such as pressure from peers and administration to comply, the consequences of reporting underperformance, and conflicting demands and time constraints among instructors who have both clinical and educational duties.[17] Kogan and colleagues have also noted variability in observers' rating processes and in the translation of the observation to numeric scores, which presents a challenge in end-of-rotation evaluations that may be conducted by a wide pool of observers who lack rater training.[18] Despite these concerns that call into question the validity of end-of-rotation evaluations, recent studies on ITERs have shown that evaluation scores in post-

graduate years 1 (PGY-1) and 2 predict performance in PGY-3, providing evidence to support their validity.[19]

A valid assessment system may provide useful information for the early detection of problem residents who can benefit from remediation and feedback. Ratings of internal medicine residents have been electronically stored in the New Innovations™ (NI) platform since 2009 based on nine domains, which include six ACGME core competencies (interpersonal and communication skills, medical knowledge, patient care, practice-based learning and improvement, professionalism, systems-based practices) and three domains relevant to the local internal medicine programme (medical interviewing, physical examination, procedural skills). Residents are rated at the end of each monthly rotation and programme directors use their scores, among other sources of information, to assess residents on their biannual performance as mandated by ACGME. As accreditation agencies in the USA and Canada move toward data-driven systems to measure trends in resident performance, the need to evaluate the validity of assessment systems, such as end-of-rotation evaluations, becomes critical.

The purpose of this study is to investigate the validity of end-of-rotation evaluations that are based on ACGME core competencies. Validity evidence will be investigated using Messick's unified validity framework, focusing on response process, internal structure, relationship to other variables, and consequences.[20] Scores from end-of-rotation evaluations will be examined for their association with learners identified as problem residents. Findings from this study will guide residency programmes as they undergo the changes required in the NAS.

METHODS

Historical data for end-of-rotation evaluations were extracted from an electronic database (NI) at a single institution, which contains ratings of 291 internal medicine residents in residency during 2009–2012, resulting in a total of 4986 assessments. Ratings of residents (PGY-1–3) were obtained from 146 fellows, 144 faculty members and 21 programme directors (associate programme directors or programme directors, including subspecialty fellowship programme directors) on nine domains. Faculty and fellows providing summative evaluations for residents serve as supervisors on clinical rotations. Residents were observed by other groups, including peers, but their ratings were not included in the analysis because the focus of this study was on assessments by fellows, faculty and programme directors. Raters used a 9-point scale on which scores of 1–3 represent 'unsatisfactory' performance, scores of 4–6 represent 'satisfactory' performance, and scores of 7–9 represent 'superior' performance in an ordinal scale; qualitative comments were also collected as part of end-of-rotation evaluations, but were not analysed in this study. The unit of analysis was the residents; multiple ratings of competencies by different raters were averaged by year and by resident to create an annual rating index for each competency; a composite mean, which averaged ratings across the nine domains, was also calculated.

Descriptive statistics were examined to study characteristics of the data and to collect evidence of the response process. Descriptive statistics are presented for fellows, faculty and programme directors. Pairwise correlations between competencies were calculated to examine their association. A generalisability study (G study) was conducted to examine the reliability and internal structure of the end-of-rotation evaluation using the r (rater) : p (person) $\times$ i (competencies) design, in which raters are nested in residents and crossed with ratings of competencies. The selection of this G study design was based on the 'unbalanced' nature of the data, which included unequal numbers of observations by different raters, following recommendations from Brennan.[21] A similar design was used by Kreiter *et al.* to resolve issues in the unbalanced ratings of observations.[22] Details of variance component estimation for unbalanced random effects are beyond the scope of this study. The G study was conducted using rating data from fellows and faculty members. Variance components from the G study were used to examine sources of error variance and the reliability of end-of-rotation evaluations. A decision study (D study) was conducted to project the number of required observations to reach a sufficient level of reliability (phi coefficient [φ] > 0.70).

To measure the longitudinal progression of competency ratings over time, hierarchical regression was used by specifying cross-classified random effects for residents and raters.[23] This analysis was conducted to identify whether ratings of a particular competency increased at a faster rate than those of other competencies. To examine whether scores from end-of-rotation evaluations can be used in the early

detection of problem residents, as part of consequential validity evidence, mean ratings at PGY-1 were correlated with learners identified as problem residents. Problem residents were identified by programme directors based on their holistic judgement of resident performance documented in a portfolio assessment system, which includes mean competency ratings based on end-of-rotation evaluations, assessment scores and qualitative comments, among other sources of information. Mean competency ratings were compared between problem and all other residents using *t*-tests. Data compilation and analyses were conducted using STATA Version 12 (StataCorp LP, College Station, TX, USA). This study was approved by the institutional review board of the study institution.

RESULTS

Assessments based on end-of-rotation evaluations were recorded, on average, 14.9 (standard deviation [SD] 8.1) times per year for each resident (by fellows, $n = 5$; by faculty members, $n = 8$; by programme directors, $n = 2$). The mean ± SD length of time spent supervising residents was 23.2 ± 7.7 days. Raters logged their ratings of competency scores a mean ± SD of 41.3 ± 51.7 days after the completion of the observation period. Institutional policy requires raters to complete scoring within 6 months of the completion of observation.

Response process

Average ratings across competencies by rater and postgraduate year are shown in Table 1. The composite mean ± SD rating was 7.74 ± 0.40. Compared with fellows, faculty and programme directors showed significantly greater variability in scores assigned across different competencies assessed during the first 2 years of residency training (*F*-test for homogeneity of variance, $p < 0.001$). During PGY-1, the variability of scores assigned across the nine domains by fellows had an SD of 0.27; for faculty members and programme directors, the SD was 0.39. The SDs of scores assigned by fellows, faculty and programme directors during PGY-2 were 0.23, 0.33 and 0.41, respectively. These differences in variability indicate that fellows provided similar ratings across the nine domains (e.g. assigning values of '7' across all nine domains), whereas faculty members and programme directors provided different ratings. Across the 3 years, the mean ± SD composite rating increased from 7.54 ± 0.95 in PGY-1, to 7.83 ± 0.86 in PGY-2, and 7.97 ± 0.83 in PGY-3. Table 1 also shows the increase in mean composite scores by rater.

Internal structure

Pairwise correlations between core competencies ranged from 0.69 to 0.92; however, after adjusting for multiple comparisons, there were no significant

Table 1 Average ratings across competencies by rater and postgraduate year (PGY): descriptive statistics

PGY	Raters	Evaluations, *n*	Composite score, mean ± SD	Min	Max
1 (2092 observations)	Fellows	997	7.57 ± 0.27	1.00	9.00
	Faculty	817	7.49 ± 0.39	2.20	9.00
	PDs	141	7.55 ± 0.39	5.10	9.00
2 (1447 observations)	Fellows	656	7.83 ± 0.23	4.63	9.00
	Faculty	616	7.83 ± 0.33	3.50	9.00
	PDs	93	7.76 ± 0.41	5.00	9.00
3 (1447 observations)	Fellows	700	7.90 ± 0.20	3.88	9.00
	Faculty	582	8.04 ± 0.24	4.78	9.00
	PDs	95	8.01 ± 0.22	4.56	9.00

Missing values were excluded from the calculations. Composite refers to the mean rating across the nine domains. Raters scored on a 9-point scale, on which scores of 1–3 represent 'unsatisfactory', scores of 4–6 represent 'satisfactory' and scores of 7–9 represent 'superior' performance. The median number of evaluations completed by a rater per year was four by fellows, five by faculty, and six by programme directors.
SD = standard deviation; PDs = programme directors.

Table 2 Variance components by fellows and faculty members: generalisability study

	Fellows		Faculty members	
Facet	**Variance component**	**Variance component, %**	**Variance component**	**Variance component, %**
p	0.117	12.3	0.097	11.5
r : *p*	0.687	72.1	0.538	63.6
i	0.008	0.9	0.017	2.0
p × *i*	0.002	0.2	0.007	0.9
r × *i* : *p*	0.139	14.6	0.187	22.1

The G study was conducted using an *r* (raters) : *p* (residents) × *i* (competencies) design in order to consider unbalanced data structure. Faculty results include data from both faculty members and programme directors.

differences in correlations to indicate greater association between any two competencies.

Variance decomposition for ratings conducted by fellows and faculty members are presented in Table 2. Ratings by programme directors were combined with those of faculty for this analysis in view of the small number of programme directors. Percentage variance components were similar between the two rater types, for which 12.3% and 11.5%, respectively, of total variance represented resident variance. However, the greatest source of error variance derived from raters nested in residents (*r* : *p*), which accounted for 72.1% of variance for fellows and 63.6% of variance for faculty members. This indicates differential levels of severity depending on the fellow or faculty member observing and assessing the resident; greater variability in severity was noted among fellows than among faculty. The reliability of end-of-rotation evaluations had a phi coefficient of about 0.70 (fellows ratings, $\varphi = 0.68$; faculty ratings, $\varphi = 0.71$). To reach sufficient levels of reliability ($\varphi > 0.70$), projections from the D study indicate at least 14 end-of-rotation evaluations.

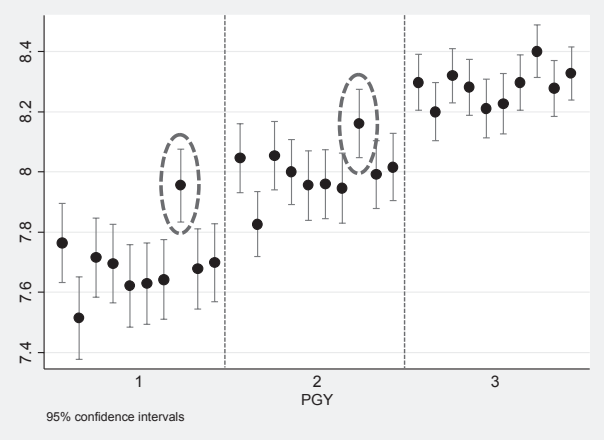

Figure 1 Mean ratings by competency and postgraduate year. The 10 bars per year group represent ratings on (from left to right): interpersonal and communication skills; medical knowledge; patient care; medical interviewing; physical examination; procedural skills; practice-based learning and improvement; professionalism; systems-based practice; and overall summary. Bars represent 95% confidence intervals; •, mean ratings. Y-axis indicates mean rating. X-axis indicates postgraduate year. Average ratings for professionalism are circled for PGY-1 and PGY-2

Relationship to other variables

Scores from end-of-rotation evaluations were examined in relation to postgraduate year of training. Figure 1 presents average ratings by competency and by postgraduate year. The *x*-axis denotes the 3 years of postgraduate residency training. Within each year, 10 plots are presented (nine domains and an overall summary rating). As Figure 1 illustrates, medical knowledge and professionalism, respectively, were awarded relatively lower and higher mean ratings than other competencies during the first 2 years of postgraduate training. The plot shows a gradual increase in the mean ratings of competencies over the 3 years. Results of a hierarchical regression show that on average ratings increased by 0.21 points per postgraduate year (95% confidence interval [CI] 0.18–0.24). Among the nine domains, medical knowledge had the lowest baseline mean rating, but the fastest rate of increase of 0.28 points per postgraduate year (95% CI 0.25–0.31); this rate of increase was significantly

greater than those of the other competencies. Professionalism had the slowest rate of increase of 0.12 points per postgraduate year (95% CI 0.09–0.15), which was significantly lower than those of the other competencies.

Consequence

Of the 210 PGY-1 residents referred to in the data, nine (4.3%) were identified as problem residents by programme directors. Although the prevalence of problem residents was small, correlation (point-biserial) between mean composite ratings and a dichotomous indicator of problem residents and all other residents was 0.33 ($p < 0.001$) (correlations ranged between 0.28 and 0.34 by competency; all $p < 0.001$). However, there were no significant differences in the magnitude of correlation for a specific competency; moreover, there were no significant differences in correlations by fellow or faculty raters. A comparison of mean ratings by ACGME core competencies between problem and all other residents is presented in Table 3. Results indicate significant differences in mean ratings, even after adjusting for the type I error rate using Bonferroni correction, for all competencies with differences ranging between 0.70 and 0.85 ($p < 0.001$). Although there were no significant differences in the magnitude of mean ratings between competencies, these results show a meaningful difference in ratings between problem and all other residents. The average ratings of problem residents were generally 1.5 SD below those of all other residents; a cross-classified random-effects model accounting for resident and rater effects indicated that a resident with a mean composite rating below 1.5 SD had significantly higher odds of being classified as a problem resident (odds ratio: 4.48; $p = 0.039$).

DISCUSSION

The general trends found in this study are consistent with results from the workplace-based assessment literature.[24] Findings from this study also support recent results of ITERs that produced reasonable reliability indices and demonstrated predictive validity.[19] This study provides several meaningful implications for end-of-rotation evaluations used to measure core competencies in residents at various levels of training and for identifying problem residents. Results show that compared with fellows, faculty and programme directors showed significantly greater variability in their ratings of different competencies. Given the 'failure to fail' concept raised in clinical skills assessments, the greater variability in scores assigned by faculty and programme directors may contribute to the detection of underperforming residents.[17] Although this does not suggest that faculty and programme directors necessarily assign more accurate scores, this finding provides empirical evidence that they may have greater ability to detect and discriminate differences between competencies when observing and assessing residents. To date, there has not been any response process validity evidence suggesting differences in the quality of ratings provided by fellows or by faculty observers. Furthermore, when sufficient numbers of end-of-

Table 3 Comparison of ratings of problem and other residents: descriptive statistics and t-test results

		Problem residents	All other residents
ACGME core competency	**Difference***	**Rating, mean ± SD**	**Rating, mean ± SD**
Interpersonal skills and communication	0.79	6.86 ± 0.76	7.65 ± 0.49
Medical knowledge	0.75	6.56 ± 0.47	7.31 ± 0.52
Patient care	0.85	6.76 ± 0.53	7.61 ± 0.47
Practice-based learning and improvement	0.75	6.79 ± 0.48	7.54 ± 0.48
Professionalism	0.70	7.19 ± 0.55	7.89 ± 0.47
System-based practices	0.80	6.79 ± 0.46	7.59 ± 0.48

* $p < 0.001$; all comparisons of mean ratings between problem and all other residents are based on *t-tests*. Results remain significant, even with Bonferroni adjustment for multiple comparisons. Comparisons are based on nine problem residents and 201 other residents during postgraduate year 1.
ACGME = Accreditation Council for Graduate Medical Education; SD = standard deviation.

rotation evaluations are used (14 or more evaluations per year), the reliability of end-of-rotation evaluations achieve phi coefficients of an acceptable level, supporting internal structure validity evidence within Messick's[20] unified validity framework. However, the greatest source of error was rater differences in severity, which were larger among fellows than among faculty raters. The variability between raters may also reflect differences in frame of reference among groups of raters who process ratings differently when translating the observation into numeric values.[18] These findings provide information on the value of using faculty, rather than fellows, to assess residents' core competencies. These findings also indicate the need to further train faculty and fellows to become better evaluators of resident performance.

The plot of average ratings by competency and by postgraduate year in Figure 1 presents an empirical validation of theoretical trends illustrated in the NAS. Nasca[25] emphasised that standards in professionalism need to be higher regardless of the training level of the resident, including in PGY-1 residents; this is empirically replicated in this study, in which ratings of professionalism were relatively higher than ratings of other competencies during PGY-1 and PGY-2. By PGY-3, mean ratings of all competencies converged to relatively similar scores. The increase in competency ratings over the 3 years of postgraduate training provides additional evidence of the validity of end-of-rotation evaluations.

As part of consequential validity evidence, mean ratings of competencies were significantly associated with whether or not a learner was classified as a problem resident. Although the number of problem residents in this study represented only 4.3% of the total number of residents during PGY-1, the relatively moderate, but significant correlation of 0.33 shows promise that end-of-rotation evaluations can be used to predict problem residents. Comparisons of mean ratings by competencies also showed significant differences. Although this study does not provide absolute evidence that end-of-rotation evaluation scores can be used to identify problem residents, ratings based on this system send a signal to programme directors who base their judgement of global resident performance on various sources of information; moreover, end-of-rotation evaluations can be used to inform programme directors of specific residents who may warrant further investigation. Despite the score inflation by raters, within the score ranges assigned, there was variability that supported the detection of underperforming residents. For example, based on the results of this study, programme directors may define end-of-rotation evaluation ratings of 1.5 SD below the mean ratings of all 'non-problem' residents as criteria for concern. Replications of this study should be conducted with a larger sample of problem residents to increase the precision of the results.

Institutional support and programme coordination are important for effective measurement of resident performance. A database that tracks these ratings and stores data for analysis is a basic necessity for an effective assessment system based on end-of-rotation evaluations. Even with electronic reminders and follow-up by the programme coordinator, the average number of evaluations recorded per resident per year amounted to only 14.9, and raters completed their ratings 41.3 days after the end of the observation period. Although it is unclear whether significant delays in logging ratings in NI have an impact on the accuracy of the recording of observations, this study indicates the need to investigate how this might affect the validity of scores; the methods and resources required to minimise delays should also be examined.

End-of-rotation evaluations included in this study can be viewed as components of multi-source feedback. As studies in multi-source feedback indicate the need to identify learners who are receptive to feedback and who are facing difficulties, the validity of end-of-rotation evaluations by fellows, faculty and other observers demands greater attention.[26,27] Although guidelines for using the evaluation form and ongoing rater training sessions were provided as part of faculty development, some raters may need additional training. Rater training sessions consisted of periodic meetings among core internal medicine faculty (programme directors, associate programme directors and chief residents), who discussed and recalibrated on the rating form, using criterion-based frame-of-reference training. However, a rigorous rater training session for all raters was not provided. As such, fellows may benefit from rater training in order to increase their discrimination among different levels of resident performance because efforts to improve evaluations could lead to a higher quality of resident assessment and feedback.[28] This study provides empirical evidence for the support and further development of effective assessment systems based on end-of-rotation evaluations, which includes direct observations and multi-source feedback, for postgraduate medical education.

There are some limitations to this study. This study was conducted in only one residency programme at a single institution. However, given the exploratory nature of this study and the timing of this work in relation to the NAS, these results provide meaningful implications for the field. Additional studies replicating the procedures in this study using larger and more heterogeneous samples across multiple institutions should be conducted to increase the generalisability of the results presented. Analyses pertaining to problem residents are based on programme directors' overall review of resident performance, which includes end-of-rotation evaluation scores. However, the programme director also bases the classification of the problem resident on other assessment scores, including test scores, conference attendance, and direct observation performance.

When sufficient numbers of evaluation scores are used, end-of-rotation evaluations are a useful method for measuring the growth of resident performance associated with core competencies. Furthermore, end-of-rotation evaluation scores provide preliminary evidence with which to detect and predict problem residents in subsequent postgraduate training years. Additional studies are underway to examine the link between qualitative comments and the quantitative scores reported, including whether feedback differs among different groups consisting of peers, fellows, faculty members and programme directors.

Contributors: YSP contributed to the study design, acquisition of data, analysis and interpretation of results, and wrote the primary draft of the manuscript. JR contributed to the study design and the interpretation of results. AT contributed to the study conceptualisation, the plan for data analysis, and the interpretation of results. All authors contributed to the revision of the paper and approved the final manuscript for publication.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: this study was approved by the Institutional Review Board of the University of Illinois at Chicago.

REFERENCES

1 Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system. *N Engl J Med* 2012;**366**:1051–6.
2 Yao DC, Wright SM. The challenge of problem residents. *J Gen Intern Med* 2001;**16**:486–92.
3 American Board of Internal Medicine. Materials from Association of Program Directors in Internal Medicine (APDIM) Chief Residents' Workshop on Problem Residents, 19 April 1999, New Orleans, LA.
4 Steinert Y, Levitt C. Working with the 'problem' resident: guidelines for definition and intervention. *Fam Med* 1993;**25**:627–32.
5 Lucey CR, Boote RM. Working with problem residents: a systematic approach. In: Holmboe ES, Hawkins RE, eds. *Practical Guide to the Evaluation of Clinical Competence*. Philadelphia, PA: Mosby 2008;201–15.
6 Roberts NK, Williams RG, Klingensmith M *et al*. The case of the entitled resident: a composite case study of a resident performance problem syndrome with interdisciplinary commentary. *Med Teach* 2012;**34**: 1024–32.
7 Samenow CP, Worley LL, Neufeld R, Fishel T, Swiggart WH. Transformative learning in a professional development course aimed at addressing disruptive physician behaviour: a composite case study. *Acad Med* 2013;**88**:117–23.
8 Yao DC, Wright SM. National survey of internal medicine residency programme directors regarding problem residents. *JAMA* 2000;**284**:1099–104.
9 Zbieranowski I, Takahashi SG, Verma S, Spadafora SM. Remediation of residents in difficulty: a retrospective 10-year review of the experience of a postgraduate board of examiners. *Acad Med* 2013; **88**:111–6.
10 Hauer KE, Ciccone A, Henzel TR, Katsufrakis P, Miller SH, Norcross WA, Papadakis MA, Irby DM. Remediation of the deficiencies of physicians across the continuum from medical school to practice: a thematic review of the literature. *Acad Med* 2009;**84**: 1822–32.
11 Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees. *JAMA* 2009;**302**:1316–26.
12 Epstein RM. Assessment in medical education. *N Engl J Med* 2007;**356**:387–96.
13 Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts KM. Evaluation of physicians' professional performance: an iterative development and validation study of multi-source feedback instruments. *BMC Health Serv Res* 2012;**12**: 1–11.
14 Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE guide no. 31. *Med Teach* 2007;**29**:855–71.
15 Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? *Acad Med* 2004;**79**:453–7.
16 Iobst WF, Sherbino J, ten Cate O, Richardson DL, Dath D, Swing SR, Harris P, Mungroo R, Holmboe ES, Frank JR. Competency-based medical education in postgraduate medical education. *Med Teach* 2010; **32**:651–6.

17 Cleland JA, Knight LV, Rees CE, Tracey S, Bond CM. Is it me or is it them? Factors that influence the passing of underperforming students. *Med Educ* 2008;**42**:800–9.
18 Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ* 2011;**45**:1048–60.
19 Ginsburg S, Eva K, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med* 2013;**88**:1539–44.
20 Messick S. Standards of validity and the validity of standards in performance assessment. *Educ Meas* 1995;**14**:5–8.
21 Brennan RL. *Generalizability Theory*. New York, NY: Springer-Verlag 2001.
22 Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalisability study of a new standardised rating form used to evaluate students' clinical clerkship performances. *Acad Med* 1998;**73**:1294–8.
23 Hox J. *Multilevel Analysis: Techniques and Applications*. New York, NY: Routledge 2010.
24 Davies H, Archer J, Southgate L, Norcini J. Initial evaluation of the first year of the Foundation Assessment Programme. *Med Educ* 2009;**43**:74–81.
25 Nasca T. Graduate Medical Education in the United States: Vision and General Directions for the Next 10 Years. Association of American Medical Colleges, 7 November 2010, Washington, DC.
26 Smither JW, London M, Reilly RR. Does performance improve following multi-source feedback? A theoretical model, meta-analysis and review of empirical findings. *Pers Psychol* 2005;**58**:33–66.
27 Archer J, Swanwick T, Smith D, O'Keefe C, Cater N. Developing a multi-source feedback tool for postgraduate medical educational supervisors. *Med Teach* 2013;**35**:145–54.
28 Holmboe ES, Fiebach NH, Galaty LA, Huot S. Effectiveness of a focused educational intervention on resident evaluations from faculty: a randomised controlled trial. *J Gen Intern Med* 2001;**16**:427–34.

Received 19 June 2013; editorial comments to author 13 September 2013, 8 November 2013; accepted for publication 14 November 2013