Effect of Rater Training on Reliability and Accuracy of Mini-CEX Scores: A Randomized, Controlled Trial

David A. Cook, MD, MHPE^{1,2}, Denise M. Dupras, MD, PhD³, Thomas J. Beckman, MD^{1,2}, Kris G. Thomas, MD³, and V. Shane Pankratz, PhD⁴

¹Office of Education Research, College of Medicine, Mayo Clinic, Rochester, MN, USA; ²Division of General Internal Medicine, College of Medicine, Mayo Clinic, Rochester, MN, USA; ³Division of Primary Care Internal Medicine, College of Medicine, Mayo Clinic, Rochester, MN, USA; ⁴Division of Biostatistics, College of Medicine, Mayo Clinic, Rochester, MN, USA.

BACKGROUND: Mini-CEX scores assess resident competence. Rater training might improve mini-CEX score interrater reliability, but evidence is lacking.

OBJECTIVE: Evaluate a rater training workshop using interrater reliability and accuracy.

DESIGN: Randomized trial (immediate versus delayed workshop) and single-group pre/post study (randomized groups combined).

SETTING: Academic medical center.

PARTICIPANTS: Fifty-two internal medicine clinic preceptors (31 randomized and 21 additional workshop attendees).

INTERVENTION: The workshop included rater error training, performance dimension training, behavioral observation training, and frame of reference training using lecture, video, and facilitated discussion. Delayed group received no intervention until after posttest.

MEASUREMENTS: Mini-CEX ratings at baseline (just before workshop for workshop group), and four weeks later using videotaped resident-patient encounters; mini-CEX ratings of live resident-patient encounters one year preceding and one year following the workshop; rater confidence using mini-CEX.

RESULTS: Among 31 randomized participants, interrater reliabilities in the delayed group (baseline intraclass correlation coefficient [ICC] 0.43, follow-up 0.53) and workshop group (baseline 0.40, follow-up 0.43) were not significantly different ($p=0.19$). Mean ratings were similar at baseline (delayed 4.9 [95% confidence interval 4.6–5.2], workshop 4.8 [4.5–5.1]) and follow-up (delayed 5.4 [5.0–5.7], workshop 5.3 [5.0–5.6]; $p=0.88$ for interaction). For the entire cohort, rater confidence (1=not confident, 6=very confident) improved from mean (SD) 3.8 (1.4) to 4.4 (1.0), $p=0.018$. Interrater reliability for ratings of live encounters (entire cohort) was higher after the workshop (ICC 0.34) than before (ICC 0.18) but the standard error of measurement was similar for both periods.

Electronic supplementary material The online version of this article (doi:10.1007/s11606-008-0842-3) contains supplementary material, which is available to authorized users.

Received May 7, 2008

Revised September 3, 2008

Accepted October 9, 2008

Published online November 11, 2008

CONCLUSIONS: Rater training did not improve interrater reliability or accuracy of mini-CEX scores.

CLINICAL TRIALS REGISTRATION: clinicaltrials.gov identifier NCT00667940

KEY WORDS: medical education; faculty development; rater training; clinical competence; assessment; randomized trial.

J Gen Intern Med 24(1):74–9

DOI: 10.1007/s11606-008-0842-3

© Society of General Internal Medicine 2008

INTRODUCTION

Residency program directors must certify the competence of graduating residents. Valid assessments of resident performance are needed.^{1–3} However, obtaining valid assessments of clinical skills is challenging.^{1,2,4} Research found scores from the American Board of Internal Medicine (ABIM) “long case” clinical evaluation exercise (CEX) to be unreliable due to large inter-case and inter-rater variance.^{5–9} Consequently, the ABIM now endorses the mini-CEX,^{10,11} in which faculty observe residents during multiple brief clinical encounters and rate performance overall and in six specific domains (interviewing, physical examination, humanistic qualities/professionalism, clinical judgment, counseling, and organization/efficiency) using a nine-point scale.

The reliability of scores from one long case^{5,7,8} and one mini-CEX encounter^{2,3,10,12,13} are similar. However, the mini-CEX’s brevity facilitates multiple observations which collectively demonstrate acceptable reliability^{2,10} (i.e., consistency of scores across measurement replications).¹⁴ Improving interrater reliability could reduce the number of resident-patient encounters required for reliable scores and improve the validity of mini-CEX score interpretations.

Although faculty training could improve rater performance, few studies have examined faculty development in clinical performance assessment.¹⁵ One study found that neither 30-minute nor two-hour training sessions improved score reliability, but this study was limited by sample size.¹⁶ Another study found that a 15-minute instructional videotape did not affect rating accuracy.⁹ More recently, a study using the mini-CEX found that following a one-day “direct observation of competence training” course participants’ mini-CEX ratings were lower, but not necessarily more accurate, than nonparticipants’ ratings.¹⁷ However, research suggests that rater

training can improve score reliability in clinical assessments.^{18–20} Moreover, a meta-analysis of non-medical education studies suggested that rater training workshops can improve score accuracy.²¹ Given the potential benefits of rater training, further investigation in medical education appears warranted.

We sought to evaluate the effect of a rater training workshop, in comparison to no intervention, on the interrater reliability and accuracy of mini-CEX scores from resident continuity clinic preceptors. We hypothesized that the workshop would improve both interrater reliability and accuracy.

METHODS

Overview

We conducted a randomized trial comparing a rater training workshop (held in May 2006) to no intervention. Rater reliability and accuracy were measured at baseline and four weeks later using videotaped resident–patient encounters (Fig. 1). A single-group pre/post comparison using ratings of live resident–patient encounters was also performed. Our Institutional Review Board exempted this study.

Setting and Participants

All internal medicine residency continuity clinic preceptors were invited to participate in the workshop and in the study.

Figure 1. Study design and flow of study participants. Each pretest and posttest consisted of 16 videotaped encounters (nine scripted and seven unscripted). Analyses for interrater reliability and halo effect used all 16 encounters. Analyses for accuracy (discrimination, percent agreement, and chance-corrected agreement) used only the nine scripted encounters. Randomized participants completed the baseline survey with the pretest.

Study participation was voluntary, and all participants consented. A subset agreed to participate in a randomized trial. Author DAC used MINIM (version 1.5, London Hospital Medical College, London) to randomly assign these participants to immediate workshop (intervention) or delayed workshop (control), with stratification by continuity clinic firm.

Interventions

The half-day workshop used rater training methods similar to those used by Holmboe et al.¹ and shown to be effective in other settings.^{15,16,21} One of the workshop facilitators (author DMD) had attended the rater training course described by Holmboe¹ and was thus familiar with these methods. Our workshop focused on the need for accurate and reliable resident assessment; potential errors and biases in rating systems (rater error training); facilitated discussion of the domains of the mini-CEX (performance dimension training); methods to improve observation (behavioral observation training); and facilitated discussion in large and small groups of videotaped interactions between residents and standardized patients (frame of reference training). Since one meta-analysis found that frame-of-reference training improves rater accuracy more than other training methods,²¹ we used this method for more than half the workshop.

Delayed workshop participants received no specific intervention until after rating the second set of videotaped encounters (posttest). They then attended the workshop.

Outcomes and Instruments

The primary outcome was interrater reliability for mini-CEX scores of videotaped resident–patient encounters. Secondary outcomes included score accuracy, mean scores and interrater reliability for mini-CEX scores of real resident–patient encounters, halo effect, and preceptor perceptions of the mini-CEX.

Participants in the immediate workshop group rated 16 videotaped encounters between residents and standardized patients just before the workshop (pretest) and a second set of 16 encounters four weeks later (posttest), while those in the delayed workshop group rated the first set of cases four weeks prior to the workshop (pretest) and rated the second set immediately before the workshop (posttest). Raters were instructed to rate whatever competence domain(s) they could appropriately assess and provide an overall rating for each case.

Nine cases came from the Yale Rater Training Program and nine cases from the ABIM-NBME Faculty Development Program. These cases, developed for and used in previous studies of the mini-CEX,^{1,17} were scripted to portray unsatisfactory, satisfactory, and superior performance for interviewing, examination, and counseling skills. To achieve the number of cases required by our power analysis, we developed an additional fourteen cases using standardized patients to portray seven patient scenarios. We coached internal medicine residents to demonstrate varying levels of competence, but performance was not scripted. These 32 cases were divided into two case sets (pretest and posttest), each consisting of nine scripted cases (one at each level of performance for each skill domain) and seven local cases (one from each patient scenario). Case sequence within each set was randomized. All participants completed the same pretest and the same posttest.

We also obtained ratings from all resident–patient encounters observed in residents' outpatient clinic by workshop participants. Faculty rated these encounters using a modified mini-CEX with a five-point scale where 1="needs improvement," 2–4=average, and 5="top 10%."²² Observations for the 12 months preceding and following the workshop were compared.

Prior to the pretest or starting the workshop, each workshop participant completed a questionnaire ascertaining demographic information. Confidence in the use of, and perceived accuracy and usefulness of, the mini-CEX were assessed before and after the course using a six-point Likert scale.

Statistical Analysis

Interrater reliability for scores from videotaped encounters was measured with intraclass correlation coefficients (ICC), calculated using variance components estimated via maximum likelihood. ICC confidence intervals were obtained via a profile likelihood approach, and comparisons among ICCs were made using a likelihood ratio test accounting for repeated assessments made by each rater.²³ A sample of 15 preceptors per group and 16 ratings per preceptor was estimated to provide 80% power to detect a change from an estimated pre-workshop ICC of 0.6 to a post-workshop ICC of 0.8.²⁴

Accuracy analyses used only scripted encounters. Discrimination was evaluated by comparing mean ratings across the three levels of competence using mixed linear models with repeated measures on subjects. The frequency with which actual ratings matched scripted performance was compared between groups using a logistic regression model that included repeated measures across subjects via generalized estimating equations.²⁵ Our experience with the scripted cases indicated that experts had legitimate disagreements in performance ratings, suggesting some degree of error in the scripted performance levels. Hence, ICC (which is similar to weighted kappa²⁶) was calculated as a measure of chance-corrected agreement with scripted performance level. We calculated ICCs for each rater individually and compared these between groups using mixed linear models with repeated measures on subjects.

The halo effect was evaluated by averaging across domains for each set of ratings (i.e., the average of interviewing, exam, efficiency, etc. from one rater on one case) and comparing the resultant standard deviations between groups²⁷ using mixed linear models with repeated measures on subjects. In this analysis, lower standard deviations indicate greater halo (less variability among domains).

Raters occasionally omitted overall ratings for some videotaped cases. When this occurred, the average of all rated domains was used as the overall rating (supplemented scores). To verify the appropriateness of this method, the strength of the association between overall scores and domain scores was computed as the square of Pearson's r (R^2 , or percent of variance explained). Domain scores accounted for >61% of the variance in overall scores (R^2 range 0.61–0.84). Because of this, and because analyses based on the subset of raters recording overall ratings yielded similar results, only analyses with supplemented scores are reported.

For scores from observations of live resident–patient encounters, mean ratings before and after the workshop were compared using mixed linear models with repeated measures on raters. To estimate the reliability of these scores, random effects variance components were used to calculate the standard error of measurement (SEM)^{28,29} and the dependability index (ϕ).²⁹ Residents with fewer than three observed encounters were excluded from reliability analyses.

Raters' beliefs about the mini-CEX were compared pre-workshop and post-workshop and between groups using the Wilcoxon signed rank and rank sum tests.

All analyses were conducted with SAS 9.1 and S-Plus 8.0.1 using a two-sided alpha of 0.05. All participants were analyzed in their assigned group.

RESULTS

Fifty-two of 54 eligible faculty members participated in the workshop and consented to their data being reported (Fig. 1). Mean (SD) age was 45.3 (11.2) years. Fifteen (29%) were women, six (12%) were full professors, 18/50 (36%) had been precepting for more than 12 years, and 7/51 (14%) had prior rater training. There were no significant demographic differences between the 31 who participated in the randomized trial and the 21 who did not ($p \geq 0.18$); see Appendix for details.

Interrater Reliability for Videotaped Encounters

Differences in interrater reliability for overall ratings among the four study conditions (immediate or delayed workshop, and pretest or posttest; $p = 0.19$) were small relative to the variability in the estimates (Table 1). For overall ratings the difference in ICC change of -0.12 favored the control group, with an approximate 95% confidence interval (CI) -0.50 to 0.26 . Variance components are reported in the Appendix.

Table 1. Interrater Reliability of Mini-CEX Ratings Before and After Rater Training

	Delayed workshop (95% CI)* N=16 raters		Immediate workshop (95% CI)* N=15 raters	
	Pretest	Posttest	Pretest	Posttest
Overall	0.43 (0.26 – 0.64)	0.53 (0.35 – 0.70)	0.40 (0.23 – 0.61)	0.43 (0.26 – 0.64)
Interviewing	0.47 (0.26 – 0.71)	0.41 (0.20 – 0.68)	0.64 (0.44 – 0.83)	0.40 (0.21 – 0.65)
Exam	0.26 (0.074 – 0.66)	0.48 (0.19 – 0.84)	0.12 (0.00 – 0.49)	0.30 (0.038 – 0.80)
Counseling	0.36 (0.16 – 0.64)	0.50 (0.30 – 0.74)	0.23 (0.065 – 0.55)	0.45 (0.23 – 0.71)

* Reliability coefficient measured using the intraclass correlation coefficient. Abbreviation: CI—confidence interval

Accuracy for Videotaped Encounters

Accuracy was measured in three ways: mean ratings (discrimination), the frequency with which ratings matched scripted performance (percent agreement), and chance-corrected agreement using ICC. Ratings (Table 2) improved significantly as scripted performance rose (mean [SD] 3.5 [1.4] for unsatisfactory, 5.3 [1.5] for satisfactory, and 6.4 [1.4] for superior performance, $p < 0.0001$ for each pair-wise comparison). Overall ratings for delayed (5.0 [1.8]) and immediate groups (5.1 [1.9], CI for difference -0.2 to 0.4 , $p = 0.68$) were similar when averaged across pre- and posttests. However, posttest ratings were higher (5.3 [1.9]) than pretest ratings (4.8 [1.8], CI for difference 0.2 to 0.8 , $p = 0.002$) when averaging across groups. The interaction between group and testing period was not significant ($p = 0.88$), suggesting that the workshop did not affect mean ratings. The interaction between performance, group, and testing period was also not statistically significant ($p = 0.94$), suggesting that results were similar for each performance level. Similar results were found for analyses of interviewing, exam, and counseling domains (see Appendix).

Percent agreement (Table 2) was similar for delayed and immediate groups when averaging across tests (odds ratio [OR] for overall ratings 1.08 [CI 0.64 – 1.81] for delayed versus immediate [referent], $p = 0.78$) and between pretest and posttest when averaging across groups (OR 1.04 [0.66 – 1.65] for posttest versus pretest, $p = 0.85$). When assessing whether percent agreement for overall ratings changed to a greater extent from pretest to posttest in the immediate group relative to the delayed group, the gains were only 1.03-fold higher (CI 0.57–1.88, $p = 0.92$) in the immediate group. Differences for interviewing, exam, and counseling domains were likewise not statistically significant ($p \geq .08$; see Appendix).

For overall ratings, accuracy ICCs were similar between groups at pretest (0.56 [0.28] delayed group, 0.54 [0.29] immediate group; CI for difference -0.19 to 0.22 , $p = 0.86$) and again following the intervention (0.58 [0.26] delayed, 0.56 [0.21] immediate; CI for difference -0.15 to 0.19 , $p = 0.80$). The interaction between group and testing period was not statistically significant ($p = 0.49$). Similar results were found for analyses of interviewing, exam, and counseling domain ratings.

Halo Effect for Videotaped Encounters

Halo describes the degree to which overall impressions influence domain-specific ratings. In our analysis, lower standard deviations indicate greater halo. However, these values did not

vary significantly between groups or testing periods ($p \geq 0.25$) nor was the interaction significant ($p = 0.85$).

Ratings of Encounters in Residents' Clinic

Over two years, 45 workshop participants assessed 177 different residents in outpatient clinic using the mini-CEX. On average, preceptors rated each of 116 residents 2.6 (1.6) times in the 12 months preceding the workshop, and each of 138 residents 3.7 (2.0) times in the 12 months following. Mean overall scores (using the 1 to 5 scale described above) were similar before (3.77 [0.55]) and after (3.78 [0.56]) the workshop (CI for difference, -0.08 to $+0.07$, $p = 0.80$).

The reproducibility coefficient (ϕ) was somewhat higher after the workshop (0.34) than before (0.18). However, the SEM was similar for both periods (0.51 before, 0.47 after). Variance components are reported in the Appendix.

Rater Confidence and Beliefs

Among 45 respondents, baseline ratings were 3.8 (1.4) for "confidence in your ability to use the mini-CEX," and 4.1 (1.2) for accuracy and 4.6 (1.2) for usefulness of mini-CEX scores, using a scale where 1=Not confident/accurate/useful at all and 6=Very confident/accurate/useful. Following the workshop, confidence among 50 respondents (43 of whom provided pre-workshop responses) increased to 4.4 (1.0) (change 0.5 [0.1 to 0.9], $p = 0.018$). However, accuracy (3.9 [1.3]) and usefulness (4.7 [1.2]) ratings were not significantly changed (accuracy change -0.1 [-0.5 to $+0.3$], $p = 0.53$; usefulness change 0.0 [-0.3 to $+0.3$], $p = 0.83$). Ratings were not significantly different between the two randomized groups ($p \geq 0.058$) or between randomized and nonrandomized participants ($p \geq 0.085$).

DISCUSSION

This half-day rater training workshop did not appear to significantly affect the reliability, accuracy, or rater halo of mini-CEX scores. These findings were consistent in both a randomized comparison using videotaped encounters as the object of assessment, and in pre/post comparisons of scores from actual resident-patient encounters. Although participants' confidence in using the mini-CEX improved, beliefs about the accuracy and usefulness of the mini-CEX changed little.

Several reasons might explain why rater training did not significantly affect accuracy and interrater reliability. First, we

Table 2. Accuracy of Overall Mini-CEX Ratings Before and After Rater Training

Accuracy measure	Scripted performance level	Delayed workshop, N = 16 raters		Immediate workshop, N = 15 raters		p*
		Pretest	Posttest	Pretest	Posttest	
Mean rating (95% CI)	All levels	4.9 (4.6–5.2)	5.4 (5.0–5.7)	4.8 (4.5–5.1)	5.3 (5.0–5.6)	0.88
	Unsatisfactory	3.5 (3.2–3.9)	3.4 (3.0–3.8)	3.6 (3.1–4.1)	3.5 (3.1–3.9)	
	Satisfactory	4.8 (4.4–5.2)	6.1 (5.7–6.4)	4.5 (4.1–5.0)	6.0 (5.6–6.3)	
	Superior	6.3 (5.9–6.7)	6.5 (6.2–6.9)	6.2 (5.8–6.6)	6.5 (6.0–6.9)	
Percent agreement, No./N (%)†	All levels	79/144 (54.9%)	75/136 (55.2%)	70/132 (53.0%)	73/135 (54.1%)	0.92

This analysis is limited to the nine cases at each assessment point (pretest and posttest) that were scripted to portray a given level of performance. See online Appendix for results for other competence domains. Abbreviation: CI—confidence interval.

*p value for interaction between group and test period

†Frequency with which mini-CEX ratings matched scripted performance: No. match/N cases (%)

selected the sample size to detect effects larger than were observed. While the power to detect small effects was inadequate, the point estimate for reliability change favored the control group and the upper confidence limit (0.26) only slightly exceeded the *a priori* level of educational significance (0.2). These findings and the relatively small magnitude of other observed differences suggest that this workshop would be unlikely to produce desired gains in reliability and accuracy. Second, the workshop format might have been ineffective. Although we used instructional methods similar to a previous study of rater training,¹ our workshop was somewhat shorter. Nevertheless, it was longer than the training provided in other studies of rater training.^{9,16} Moreover, these previous workshops did not clearly improve ratings although one workshop resulted in more stringent ratings.¹ Thus a third explanation could be, as Williams et al.¹⁵ speculated, that “physician raters are impervious to training.” A fourth explanation is that frame of reference training is limited by case specificity.³⁰ Workshop participants noted that consensus reached on one case (for example, examining a patient with dyspnea) may have little relevance to another case (examining a patient with knee pain). This failure of cases to generalize suggests that effective training would require detailed discussion of numerous cases, which may be infeasible in a typical workshop. A final explanation is that different raters observe and value different things.³¹ Participants often disagreed regarding what constituted unsatisfactory, satisfactory, or superior performance, and even prolonged discussion did not guarantee consensus.

While inconsistent ratings will lower interrater reliability and accuracy, differences of opinion among faculty may actually enhance formative feedback by illustrating alternative approaches to specific tasks. For summative assessments mini-CEX scores are limited by modest reliability.² Yet the mini-CEX's more important role may be to encourage faculty to “get in the room” with residents, observe specific skills, and provide formative feedback. Despite its importance in professional development, feedback is insubstantial and infrequently given.^{7,32-34} Perhaps feedback should be the focus of future rater training.³⁵

The workshop did enhance confidence in using the mini-CEX. The post-workshop increase in mini-CEX observations of working residents may also reflect a benefit from the workshop.

This study has several limitations. We felt that a 0.2-point increase in reliability would be educationally significant since Cohen classified correlation coefficient changes <0.3 as small³⁶ and Landis and Koch's classification of rater agreement³⁷ used blocks 0.2 points wide, but smaller changes could also be significant. The randomized trial used videotaped cases of standardized patients as the outcome, and these may not accurately reflect performance in practice. However, pre-post data from actual encounters demonstrated concordant results. Despite admirable validation efforts,¹⁷ the scripted competence classifications are undoubtedly imprecise, which may lead to underestimates of accuracy. The study took place at a single institution and all participants were general internists.

This study also has several strengths. We examined both reliability and accuracy, evaluated workshop outcomes using both videotaped and real-life applications of the mini-CEX, and achieved the target sample size for the randomized trial.

Future research will need to clarify the purpose of the mini-CEX (formative versus summative assessment) and then develop effective methods for training raters to achieve this purpose. Behavioral observation training and performance

dimension training might be more realistic than trying to improve accuracy and reliability across domains. Alternatively, extending the training period using methods such as Web-based learning may help.³⁸

Acknowledgments: We thank E.S. Holmboe for use of scripted cases, G.R. Norman for advice on psychometric analyses, and Mayo Internal Medicine faculty for their participation in this study. Funding was provided by internal sources (the Mayo Education Innovation Program). Study sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; or preparation, review, or approval of the manuscript. Dr. Cook had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Conflict of Interest: None disclosed.

Corresponding Author: David A. Cook, MD, MHPE; Division of General Internal Medicine, College of Medicine, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA (e-mail: cook.david33@mayo.edu).

REFERENCES

1. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med.* 2004;140:874-81.
2. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med.* 2003;138:476-81.
3. Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Acad Med.* 2003;78(10 Suppl):S33-5.
4. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. *Ann Intern Med.* 1998;129:42-8.
5. Woolliscroft JO, Stross JK, Silva J Jr. Clinical competence certification: a critical appraisal. *J Med Educ.* 1984;59:799-805.
6. Kroboth FJ, Kapoor W, Brown FH, Karpf M, Levey GS. A comparative trial of the clinical evaluation exercise. *Arch Intern Med.* 1985;145:1121-3.
7. Herbers JE Jr., Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ. How accurate are faculty evaluations of clinical competence. *J Gen Intern Med.* 1989;4:202-8.
8. Kroboth FJ, Hanusa BH, Parker S, et al. The inter-rater reliability and internal consistency of a clinical evaluation exercise. *J Gen Intern Med.* 1992;7:174-9.
9. Noel GL, Herbers JE Jr., Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents. *Ann Intern Med.* 1992;117:757-65.
10. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Ann Intern Med.* 1995;123:795-9.
11. Schroter S, Plowman R, Hutchings A, Gonzalez A. Reporting of Ethical Committee Approval and Patient Consent by Study Design in 5 General Medical Journals. Paper presented at the Fifth International Congress on Peer Review and Biomedical Publication, Chicago, Illinois, September, 2005.
12. Margolis MJ, Clauser BE, Cuddy MM, et al. Use of the Mini-Clinical Evaluation Exercise to Rate Examinee Performance on a Multiple-Station Clinical Skills Examination: A Validity Study. *Acad Med.* 2006;81(10 Suppl):S56-S60.
13. Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Med Educ.* 2006;40:950-6.
14. Brennan RL. An essay on the history and future of reliability from the perspective of replications. *J Educ Meas.* 2001;38:295-317.
15. Williams RG, Klamen DA, McGaghie WC. Cognitive, Social, and Environmental Sources of Bias in Clinical Performance Ratings. *Teach Learn Med.* 2003;15:270-92.

16. **Newble DI, Hoare J, Sheldrake PF.** The selection and training of examiners for clinical examinations. *Med Educ.* 1980;14:345-9.
17. **Holmboe ES, Huot S, Chung J, Norcini J, Hawkins RE.** Construct validity of the miniclinical evaluation exercise (miniCEX). *Acad Med.* 2003;78:826-30.
18. **Müller MJ, Rossbach W, Dannigkeit P, Müller-Siecheneder F, Szegedi A, Wetzel H.** Evaluation of standardized rater training for the Positive and Negative Syndrome Scale (PANSS). *Schizophr Res.* 1998;32:151-60.
19. **Müller MJ, Dragicevic A.** Standardized rater training for the Hamilton Depression Rating Scale (HAMD-17) in psychiatric novices. *J Affect Disord.* 2003;77:65-9.
20. **Angkaw AC, Tran GQ, Haaga DAF.** Effects of training intensity on observers' rating of anxiety, social skills, and alcohol-specific coping skills. *Behav Res Ther.* 2006;44:533-44.
21. **Woehr DJ, Huffcutt AI.** Rater training for performance appraisal: A quantitative review. *J Occup Organ Psychol.* 1994;67(3):189-205.
22. **Cook DA, Beckman TJ.** Psychometric properties of mini-clinical evaluation exercise (mini-CEX) scores: Accuracy, reliability, and effect of scale length. Paper presented at the 2008 meeting of the American Educational Research Association, New York, March, 2008.
23. **Casella G, Berger RL.** *Statistical Inference.* 2New York: Duxbury Press; 2001.
24. **Walter SD, Eliasziw M, Donner A.** Sample Size and Optimal Study Designs for Reliability Studies. *Stat Med.* 1998;17:101-10.
25. **Zeger SL, Liang K-Y.** Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics.* 1986;42:121-30.
26. **Fleiss JL, Cohen J.** The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas.* 1973;33:613-9.
27. **Jacobs R, Kozlowski SW.** A closer look at halo error in performance ratings. *Acad Manage J.* 1985;28:201-12.
28. **Harvill LM.** NCME Instructional Module: Standard Error of Measurement. *Educ Meas: Issues Pract.* 1991;10(2):33-41.
29. **Brennan RL.** *Generalizability Theory.* New York: Springer-Verlag; 2001.
30. **Norman G.** Research in clinical reasoning: past history and current trends. *Med Educ.* 2005;39:418-27.
31. **Murphy KR, Cleveland JN, Skattebo AL, Kinney TB.** Raters who pursue different goals give different ratings. *J Appl Psychol.* 2004;89:158-64.
32. **Kroboth FJ, Hamusa BH, Parker SC.** Didactic value of the clinical evaluation exercise. Missed opportunities. *J Gen Intern Med.* 1996;11:551-3.
33. **Srinivasan M, Hauer KE, Der-Martirosian C, Wilkes M, Gesundheit N.** Does feedback matter? Practice-based learning for medical students after a multi-institutional clinical performance examination. *Med Educ.* 2007;41:857-65.
34. **Fernando N, Cleland J, McKenzie H, Cassar K.** Identifying the factors that determine feedback given to undergraduate medical students following formative mini-CEX assessments. *Med Educ.* 2008;42:89-95.
35. **Holmboe E, Fiebach N, Galaty L, Huot S.** Effectiveness of a focused educational intervention on resident evaluations from faculty. *J Gen Intern Med.* 2001;16:427-34.
36. **Cohen J.** *Statistical Power Analysis for the Behavioral Sciences.* 2Hillsdale, NJ: Lawrence Erlbaum; 1988.
37. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-74.
38. **Kobak KA, Engelhardt N, Lipsitz JD.** Enriched rater training using Internet based technologies: a comparison to traditional rater training in a multi-site depression trial. *J Psychiatr Res.* 2006;40:192-9.