International Journal of Medical Education. 2012;3:245-251
ISSN: 2042-6372
DOI: 10.5116/ijme.50c3.2f27

Sources of variability in medical student evaluations on the internal medicine clinical rotation

Ryan Spielvogel[1], Zachary Stednick[2], Laurel Beckett[2], Darin Latimore[3]

[1]Department of Family and Community Medicine, University of California Davis Medical Center, USA
[2]Department of Public Health Sciences, University of California Davis, USA
[3]Department of Internal Medicine, University of California Davis Medical Center, USA

Correspondence: Ryan Spielvogel, 4860 Y Street, Suite 1600, Sacramento, CA 95817, USA
Email: Ryan.Spielvogel@gmail.com

Accepted: December 08, 2012

Abstract

Objectives: To explore the sources of variability in evaluator ratings among third year medical students in the Internal Medicine clinical rotation. Also, to examine systematic effects and variability introduced by differences in the various student, evaluator, and evaluation settings.

Methods: A multilevel model was used to estimate the amount of between-student, between-rater and rater-student interaction variability present in the students' clinical evaluations in a third year internal medicine clinical rotation. Within this model, linear regression analysis was used to estimate the effect of variables on the students' numerical evaluation scores and the reliability of those scores.

Results: A total of 2,747 evaluation surveys were collected from 389 evaluators on 373 students over 4.5 years. All surveys used a nine-point grading scale, and therefore all results are reported on this scale. The calculated between-rater, between-student and rater-student interaction variance components were 0.50, 0.27 and 0.62, respectively. African American/Black students had lower scores than Caucasian students by 0.58 points (t=-3.28; P=0.001). No gender effects were noted.

Conclusions: These between-rater and between-student variance components imply that the evaluator plays a larger role in the students' scores than the students themselves. The residual rater-student interaction variance was larger and did not change by accounting for the measured demographic variables. This implies there is significant variability in each rater-student interaction that remains unexplained. This could contribute to unreliability in the system, requiring that students receive between 8 and 17 clinical evaluations to achieve 80% reliability.

Keywords: Medical student evaluations, reliability, variability, race/ethnicity, clinical rotations

Introduction

Medical students are subjected to a variety of tests designed to assess their knowledge and performance.[1] In their clinical years, medical students' final grades in their rotations are often calculated using standardized surveys.[1,2] These surveys typically rate students on various competencies using a Likert or semantic differential scale, which is completed by faculty and residents.[1,2]

While there has been a move towards more standardized ways of evaluating students, these clinical evaluations are still largely based on non-standardized subjective interactions.[2] As each student is only exposed to a small subset of all possible raters, differences in raters' grading could introduce variability into the measurement of a student's performance and thus make the measurement less reliable. Additionally, specific attributes of the student, evaluator, or environment in which the interaction took place might affect the student's scores by introducing systematic differences into the measurement, further decreasing the reliability of the system. While previous research has studied the relationships between various demographic data of those being evaluated and their numerical scores[1,3-11] as well as the reliability of clinical

© 2012 Ryan Spielvogel *et al.* This is an Open Access article distributed under the terms of the Creative Commons Attribution License which permits unrestricted use of work provided the original work is properly cited. http://creativecommons.org/licenses/by/3.0

evaluations,[4] few have investigated the effect these have on the system's overall reliability or further explored the specific sources of the variation. Additionally, many studies on this topic examined evaluations of residents by faculty. However, medical students typically have different roles on the team than residents, and the students are also being evaluated by residents, who are less experienced than faculty. For this reason, the body of research examining evaluations of residents cannot be assumed to be generalizable to medical students.[12-16] One previous study evaluating these effects did not attempt to resolve out the sources of the variability in the system and was possibly limited by the choice of statistical model.[4] Other studies examining for systematic error introduced by demographic data often yielded conflicting results,[3,4,6-8,17,18] or were using more objective structured clinical encounters, or OSCEs.[3,17,19] There is also limited research on the effects of student age, rater gender, or the rater-student gender interaction on clinical evaluation scores. These variables in particular are of increasing importance as the gender and ethnic diversity of medical school classes continues to increase.[20]

Given the widespread use of these subjective evaluation surveys in calculating students' grades, the importance these evaluations play in students' prospects for advancement, and the above limitations in previous works in this area, our study was designed to take an in-depth look at the sources of variability within the evaluation system using a more comprehensive array of possible predictors. We hypothesized that we would find a significant amount of between-rater variability and variability introduced by the rater-student interaction. Additionally, we suspected that controlling for the various student, rater and setting characteristics would diminish the observed rater-student interaction variance component and account for a significant amount of the observed variability in the system.

Methods

Study design

We conducted a cross-sectional observational study using performance evaluations of third year medical students at the University of California, Davis School of Medicine in Sacramento, California, USA, collected from January 2005 - April 2009 using the online E-Value system. The UC Davis Institutional Review Board deemed this study exempt from review due to our use of pre-existing de-identified data and study of normal educational practices and approved the design under protocol number 200917414-1.

Participants

All third year medical students completing their required rotation in Internal Medicine during the above time frame, as well as all residents and faculty completing their evaluation of the students, were included in analysis. Clinical evaluations were completed by faculty and residents. First year residents do not receive any formal evaluation training, and second year residents receive a 30 minute overview of the evaluation survey form and process at the beginning of their second year of residency. Faculty do not have regular evaluation training. Students completed the rotation in one of six eight-week rotation slots throughout the year, which was split into two consecutive four-week blocks. Each block was completed at one of four training sites, including a university hospital (University of California Davis Medical Center), a community hospital (North Kaiser Permanente), a veteran hospital (Mather VA) and a military hospital (David Geffen Hospital). Assessments were based on performance during clinical rounds and while carrying out clinical duties. Only students who completed the course during their third year of medical school were included in the study. Also, in the rare instance that a rater evaluated a student more than once, e.g. in different training sites, only the first of such clinical evaluations was included so as to eliminate any effects from repeated observations.

Instrument

Evaluations were submitted using the online E-value system, which is a web-based scoring and tracking platform. For each student, evaluation surveys were automatically sent by the school via email to faculty and second and third year residents, with first year residents only receiving an evaluation form at the request of the student. On the surveys, students were scored on 12 competencies, e.g. "fund of knowledge", "physical exam", "organization", "professionalism," etc. Each competency was graded on a scale of one to nine with examples for what behaviors would constitute a "1-3" (unsatisfactory), "4-6" (satisfactory), or "7-9" (superior) under each competency. Using the online system, no incomplete evaluations can be submitted, and <1% of respondents used the "insufficient contact to judge" answer for any competency.

Data collection

All data were downloaded from the online E-Value system archive. Setting characteristics such as training site, rotation number, and block number, as well as unique rater and student identifiers were embedded in the acquired data set. All data and demographic information were de-identified and coded by UC Davis research assistants prior to release for analysis. Rater demographic information was obtained using public record and student demographic information was self-reported on their medical school entry questionnaires. The breakdown of student and rater demographic information is included in Table 1.

Data analysis

All separate competency scores for each student were included as outcomes in the analyses, rather than computing a mean score across competencies. The initial goal of statistical analysis was to assess whether there were any systematic differences, or "fixed effects," in evaluation scores introduced by known characteristics of the rater (rank/training level, gender), setting (competency, training site), and student (age group, gender, race/ethnicity, academic year), as well as rater-student gender match or difference. All of these characteristics were included as predictors in the final model. The second goal was to assess the amount of unexplained variability, or "random effects," present in the system among how students are graded, how raters grade and the variation within a single student's scores from different raters. This was accomplished by estimating the components of variance between students, between raters and within student after accounting for all of the above calculated fixed effects.

Table 1. Demographic characteristics of the 389 raters and 373 medical students (N=762)

Variable	Number	%
Rater characteristics		
Rater level of training		
Residents, year 1	57	15
Residents, year 2	109	28
Residents, year 3	32	8
Faculty	191	49
Rater gender (number and % male)	214	55
Student characteristics		
Student gender (number and % male)	164	44
Student ethnicity		
African-American/Black	11	3
Hispanic	31	8
Chinese-American	58	16
Other Asian	90	24
White/Caucasian	176	47
Other	7	2
Student age group		
Up to 25	105	28
26-30	192	52
31-35	58	15
36-40	15	4
Over 40	3	1

Previous studies have focused on the second question, using generalizability theory,[4] which typically uses random effects models to estimate variance components but does not incorporate fixed effects or hypothesis testing.[21] We used hierarchical mixed linear models to allow us to address questions both about potential systematic differences in ratings (fixed effects, such as gender) and unexplained variation (random effects for rater, student, and their interaction).[21-23] We used a likelihood-based estimation approach (REML) to allow for the unbalanced data setting (unequal numbers of ratings, and not all students seen by all raters), and to provide variance estimates and standard error estimates, under the assumptions of normality and homoscedasticity of the residuals from the model, with residuals using the profile approach. Model assumptions were checked graphically and descriptively using residuals. Variance components were used to estimate the proportion of the total variation (sum of student, rater, and within-student between-rater components) attributable to differences across students. A modified Spearman-Brown approach was used to calculate the number of ratings to achieve 80% reliability.

Multilevel models were first utilized to examine subsets of the rater, student and setting characteristics individually as fixed effects while treating between-rater, between-student, and rater-student interaction variability as random components. Following this, more complex models were constructed leading to final models that accounted for all hypothesized fixed and random components. All hypothesis tests were two-sided at level 0.05, and statistical analysis was carried out by author ZS using SAS software, Version 9.3 (SAS Institute, Cary, NC). Variance components reported are assumed to be uniform across student and rater subgroups (homoscedasticity).

Results

A total of 2,747 performance evaluations were collected on 373 students after being evaluated by 389 raters, including 198 residents and 191 faculty (Table 1). Each student received an average of 7 clinical evaluations (range 1 to 16) with 90% of students receiving between 5 and 10 clinical evaluations. Raters submitted between 1 and 68 evaluations over the study period with 50% of all raters submitting only 3 or less.

Results are organized below by systematic and random effects. Table 2 shows the effects of the characteristics of the rater and student as well as academic year on numerical scores as well as the associated p values and t statistics. Table 3 shows the estimated effects of the setting in which the student was rated and the associated p values and t statistics. All effects reported are adjusted for all other rater, student and setting characteristics including competency and are reported relative to a 9-point scale. In all cases, the largest sub-group was chosen as the reference. Residual analysis showed adequate fit to the random effects assumed in developing the multilevel model.

Systematic effects

Scores varied significantly by training level of the rater. Compared to faculty, first year residents scored significantly higher (0.25 points; 95% CI 0.17-0.33; t=6.45; P<0.001), as did second year residents (0.12 points; 95% CI 0.06-0.18; t=3.53; P<0.001) and third year residents (0.20 points; 95% CI 0.12-0.28; t=5.37; P<0.001). Neither rater gender, nor student gender, nor rater-student gender differences were associated with significant differences in mean score. Scores varied significantly by student racial/ethnic group. Compared to White/Caucasian students, African-American /Black students scored 0.58 points lower (95% CI 0.23-0.93 points lower; t=-3.28; P= 0.001) and Chinese-American/Chinese students scored 0.24 points lower on average (95% CI 0.11-0.41 points; t=-2.81; P=0.005). A trend was noted for Hispanic students to score about 0.2

points lower, on average, but did not reach statistical significance. Scores decreased with increasing age of the student. Compared to the 26-30 age group, younger students (up to 25) scored 0.22 points higher (95% CI 0.09-0.35; t=3.31; P=0.0009), students aged 31-35 scored 0.24 points lower (95% CI 0.08-0.40; t=-2.89; P=0.004) and students aged 36-40 scored 0.34 points lower (95% CI 0.04-0.64; t=-2.19; P=0.03). Of the training sites, compared to the university hospital, only scores obtained at the community hospital differed significantly, with students scoring 0.12 points lower (95% CI 0.08-0.16; t=-6.18; P<0.001). Rotation number also affected scores, with students in Rotation 1 scoring 0.27 points lower than those in Rotations 4 and 6 (95% CI 0.10-0.44;t=-3.15; P=0.002).

Table 2. Estimated effects of rater and student characteristics on student score in mixed effects linear model including all rater, student, and setting predictors (N=762)

Variable	Estimated effect size*	Standard error of effect size	t value	P value
Rater Characteristics				
Rater level of training (overall F)				<0.001
Residents, year 1	0.25	0.04	6.45	<0.001
Residents, year 2	0.12	0.03	3.53	<0.001
Residents, year 3	0.20	0.04	5.37	<0.001
Faculty	--	--	--	REF
Rater gender (male compared to female)	0.04	0.07	0.55	0.59
Student characteristics				
Student gender (male compared to female)	-0.09	0.06	-1.66	0.10
Student ethnicity (overall F test)				0.01
African American/Black	-0.58	0.18	-3.28	0.001
Hispanic	-0.19	0.10	-1.84	0.07
Chinese-American	-0.24	0.08	-2.81	0.005
Other Asian	-0.06	0.07	-0.85	0.59
White/Caucasian	--	--	--	REF
Other	0.11	0.19	0.54	0.58
Student age group (overall F test)				<0.001
Up to 25	0.22	0.07	3.31	0.0009
26-30	--	--	--	REF
31-35	-0.24	0.08	-2.89	0.004
36-40	-0.34	0.15	-2.19	0.03
Over 40	-0.53	0.31	-1.68	0.09
Academic year (overall F test)				<0.001
2004-2005	0.13	0.13	1.08	0.28
2005-2006	0.21	0.08	2.47	0.01
2006-2007	0.395	0.08	4.95	<0.001
2007-2008	0.42	0.08	5.18	<0.001
2008-2009	--	--	--	REF

*Estimated differences from mean rating for the reference group for that predictor, after adjusting for other variables and for unequal sample sizes.

Random effects

The multilevel model was fitted using a hierarchical variance structure to more accurately estimate the variation between raters, between students, and in the rater-student interaction across multiple evaluations.

Table 3. Estimated effects of setting characteristics on student score after adjusting for rater and student characteristics (N=762)

Variable	Estimated effect size*	Standard error of effect size	t value	P value
Setting characteristics				
Training site (overall F test)				<0.001
University hospital (UCDMC)	--	--	--	REF
Community hospital (North Kaiser)	-0.12	0.02	-6.18	<0.001
Veteran hospital (Mather VA)	-0.02	0.02	-0.84	0.34
Military hospital (Travis AFB)	0.16	0.14	1.18	0.24
Rotation (overall F)				0.02
Rotation 1	-0.27	0.09	-3.15	0.002
Rotation 2	-0.12	0.10	-1.23	0.21
Rotation 3	-0.16	0.08	-1.94	0.05
Rotation 4	-0.01	0.08	-0.10	0.92
Rotation 5	-0.15	0.09	-1.74	0.08
Rotation 6	--	--	--	REF
Block (second compared to first)	0.06	0.01	-5.38	<0.001

*Estimated differences from mean rating for the reference group for that predictor, after adjusting for other variables and for unequal sample sizes.

These findings are summarized in Table 4 along with associated p values, which were calculated using a maximum likelihood approach. These estimates were remarkably consistent across simple and more complex models, i.e. when considering demographic and setting characteristics individually and combined. We estimated that the between-rater variance component, the estimated variance corresponding to how differently raters grade from one another on average across a nine-point scale, was 0.50 (95% CI 0.46-0.54; z=13.05; P<0.001).The between-student variance component, a measure of how differently students score from one another, was 0.27 (95% CI 0.23-0.31; z=12.47; P<0.001). After controlling for both the between-rater and between-student variance components, the additional rater-student interaction variance component was estimated at 0.62 (95% CI 0.61-0.63; z=123.74; P<0.001). Based on these variance estimates, the reliability of a single rating of a student, by one rater, in one interaction, was estimated to be 19%. In order to achieve 80% reliability, which is the reliability cutoff used in previous studies,[4] a student would need to obtain 8-17 clinical evaluations, depending on how much of the rater-student interaction variance is due to true differences in a student's performance versus factors introduced by the rater or limitations in the system. In our sample set, only 39% of students received eight or more evaluations.

Table 4.Variance components between raters, between students, and within students not explained by known characteristics of the rater, student or setting (N=762)

Source of variation	Estimated variance component	Standard error	z value	P values*
Between students	0.273	0.022	12.47	<0.001
Between raters	0.499	0.038	13.05	<0.001
Within students	0.620	0.005	123.74	<0.001

*P values based on mixed model estimates using normal approximations for variance components.

Discussion

Our study was designed to estimate the sources of variability in the medical student evaluation process and to identify characteristics of the rater, student and interaction setting that affect the overall reliability. Better understanding of these relationships could have a significant impact on the evaluation process as similar processes are used at medical schools around the world to determine students' final grades in rotations, which in turn greatly affects their prospects for residency. Below is a discussion of our findings for both the systematic and random effects.

Systematic effects

We found that multiple characteristics of the student, rater, and setting have a systematic effect on the students' numerical scores. While most such variables contribute little, certain characteristics, such as ethnicity, were found to contribute more. This was in contrast to one previous study, which showed no significant difference between "minority" and "majority" students in performance evaluations,[3] yet agreed with other studies that showed being "non-white" or African-American was an independent risk factor for lower evaluation scores in the clinical years.[8, 11]

The fact that older students scored lower agrees with one previous study.[6] However, that study used an academic performance scale that was based on licensing exam scores, and was therefore not as relevant to our study's purpose. While there is no previous literature on the role of the rater's gender on performance evaluations, there is a general consensus in the literature that female medical students score higher than male medical students in their clinical rotations.[6,17] However, no correlation was noted in this study, either when looking at rater and student gender alone or at rater-by-student gender interactions.

The difference in grading among first, second, and third year residents and faculty is well documented and the finding of this study that residents, in general, grade more leniently than faculty agrees with the majority of the literature[7,18] except for one previous study.[4] First year residents were selected by the students, potentially introducing a selection bias. Nevertheless, these data were still included in the analysis since, like all of the data subsets, they were considered separately as well as combined when estimating systematic effects. Also, while our finding that rotation order and training site have an effect on evaluation scores is in contrast to previous studies,[9,24] the effect was very modest.

Random effects

While the between-student variability was found to be relatively small (0.27), the between-rater variability was estimated to be almost twice as large (0.50), which implies that the evaluator plays a larger role in the students' numerical scores than the students. In addition, there was a very large amount of variability within an individual student's evaluations, i.e. the rater-student interaction variability (0.62) – much more than could be accounted for by the between-rater variability on its own. Building models of varying complexity using different student-level, rater-level and setting-level characteristics did not appreciably decrease this variation and, in fact, it remained very stable across multiple model iterations.

The rater-student interaction variance component must arise from a combination of three sources in any given rater-student pair: rater factor (e.g. similarities or differences in style between the student and rater that may change perceptions of performance), system factor (e.g. internal inconsistencies within the rating process itself), or student factor (e.g. real differences in a student's performance across different interactions). If we assume that this entire variance component is from student factors, than it would not contribute to the unreliability in the system, and the number or evaluations needed to achieve 80% reliability would be 8, using the Spearman-Brown calculation. However, if the opposite were true, and this entire variance component arises from only rater and system factors, it would contribute to unreliability in the system, requiring 17 evaluations to achieve reliability. While this study cannot better resolve the source of the rater-student variance component, we were able to estimate that the current evaluation system requires between 8 and 17 clinical evaluations from distinct raters in order to attain 80% reliability. This finding agrees well with one previous article looking at evaluation reproducibility that found a student needs between 7 and 27 clinical evaluations to achieve 80% reliability depending on the competency measured.[4]

Study limitations and strengths

Our study has several limitations. This study examined only one clinical rotation, at one university, which could limit its generalizability. All evaluations were done via the internet, which could introduce a selection bias in responses received. We were also unable to examine the percentage of evaluations completed since evaluations can be erroneously sent to raters who have not observed the student and these evaluations are, presumably, ignored for that reason. Additionally, an attempt was made to examine the amount of time spent between rater and student; however, the scales used to classify the extent of the interaction were changed during the study period and the two groups were not found to be internally consistent, so analysis was not included.

However, our study had several strengths. Few prior studies have specifically attempted to examine both the sources of variability in medical student evaluations and the effect of rater and student demographic data on this variability, making our study distinctive. For our analysis we used a multilevel model for our calculations, which is an ideal model for examining data sets with multiple sources of variation[21,22] and which made our model and calculations very robust. The fact that the variance calculations were extraordinarily stable under multiple model manipulations gave us great confidence in our results. In addition, we were able to examine a wide array of variables at once, making our composite analysis stronger.

Conclusions

This study set out to determine the sources of variability in medical student evaluations on their clinical rotations. While we were able to estimate that there is more variability between raters than between the students, the sources of the rater-student interaction variability remain more elusive. Attempts to account for this residual variability by controlling for all of the student, rater and setting characteristics individually and together failed to reduce the rater-student interaction variance component, implying that other unmeasured factors must account for this variability. As a more precise estimate of the number of evaluations needed to achieve reliability of the system depends on elucidating the source of this residual variability, future research should focus on the rater-student interaction in particular to study this phenomenon in more depth. Also, in order to achieve 80% reliability, students will need to obtain more evaluations than the current average number of evaluations obtained. Schools could potentially increase the number of returned evaluations by linking number of evaluations returned to resident educational credit or faculty promotions. One way to possibly improve reliability without increasing the number of evaluations would be to have more direct and consistent observation by the evaluators,[2, 25] as many evaluators only see their students in a limited number of situations. Another possible method would be to increase efforts at faculty development aimed at standardizing scoring across raters. One previous study found no significant effect on rater reliability from attending an intensive workshop on evaluations of residents.[26] Therefore, more research is needed in this area, possibly with the use of standardized students.

We could not account for the observation that African-American, Chinese, and older students scored significantly lower than their Caucasian and younger counterparts respectively. This study, unfortunately, cannot determine whether this was due to unconscious biases on the part of the raters, true deficiencies in the students' performances or other factors. Also, while the small number of African-American students in the study (n=11) make this finding suspect, the result is congruent with prior studies[8,11] and highlights a need for further study in this area. Additionally, future research should examine a broader range of rotations, as most research in this area has focused only on the internal medicine clinical rotation.

As discussed by Kassebaum in Academic Medicine in 1999,[2] the medical education system currently relies heavily on subjective encounters to grade its students, despite mounting evidence of its unreliability. This study, as well as others, highlights the need for more widely-used objective alternatives.

Conflicts of Interest

The authors declare that they have no conflict of interest.

References

1. Magarian GJ, Mazur DJ. Evaluation of students in medicine clerkships. Academic Medicine. 1990;65:341-5.
2. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behaviors in medical school. Academic Medicine. 1999;74:841-9.
3. Campos-Outcalt D, Rutala PJ, Witzke DB, Fulginiti JV. Performances of underrepresented-minority students at the University of Arizona College of Medicine, 1987-1991. Academic Medicine. 1994;69(7):577-82.
4. Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. Journal of General Internal Medicine. 1992;7(5):506-10.
5. Griffith III CH, Wilson JF. The association of student examination performance with faculty and resident ratings using a modified RIME process. Journal of General Internal Medicine. 2008;23(7):1020-3.
6. Haist SA, Wilson JF, Elam CL, Blue AV. The effect of gender and age on medical school performance: an important interaction. Advances in Health Sciences Education. 2000;5:197-205.
7. Hull AL. Medical student performance: a comparison of house officer and attending staff as evaluators. Evaluation & the Health Professions. 1982;5:87-94.
8. Reteguiz J, Davidow AL, Miller M, Johanson WG. Clerkship timing and disparity in performance of racial-ethnic minorities in the medicine clerkship. Journal of the National Medical Association. 2002(94):779-88.
9. Whalen JP, Moses VK. The effect on grades of the timing and site of third-year internal medicine clerkships. Academic Medicine. 1990;65:708-9.
10. White CB, Dey EL, Fantone JC. Analysis of factors that predict clinical performance in medical school. Advances in Health Sciences Education. 2009;14:455-64.
11. Yates J, James D. Risk factors for poor performance on the undergraduate medical course: cohort study at Nottingham University. Medical Education. 2007;41(1):65-73.
12. Durning SJ, Cation LJ, R.J. M, Pangaro L. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. Academic Medicine. 2002;77(9):900-4
13. Herbers JE, Jr., Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ. How accurate are faculty evaluations of clinical competence? Journal of General Internal Medicine. 1989;4(3):202-8.
14. Holmboe ES, Fiebach NH, Galaty LA, Hout S. Effectiveness of a focused educational intervention on resident evaluations from faculty. Journal of General Internal Medicine. 2001;16(7):427-34.
15. Kroboth FJ, Hanusa BH, Parker S, Coulehan JL. The inter-rater reliability and internal consistency of a clinical evaluation exercise. Journal of General Internal Medicine. 1990;7(2):174-9.
16. Williams RG, Verhulst S, Colliver JA, Dunnington GL. Assuring the reliability of resident performance appraisals: more items or more observations? Surgery. 2005;137(2):141-7.
17. Haist SA, Witzke DB, Quinlivan S, Murphy-Spencer A. Clinical skills as demonstrated by a comprehensive clinical performance examination: who performs better – men or women? Advances in Health Sciences Education. 2003;8:189-99.

18. Maxim BR DT. Dimensionality, internal consistency and interrater reliability of clinical performance ratings. Medical Education. 1987;21:130-7.

19. Volkan K, Simon SR, Baker H, Todres ID. Psychometric structure of a comprehensive objective structured clinical examination: a factor analytic approach. Advances in Health Sciences Education. 2004;9:83-92.

20. Cohen JJ, Gabriel BA, Terrell C. The case for diversity in the health care workforce. Health Affairs. 2002;21(5):90-102.

21. Brennan R. Generalizability theory. New York: Springer Verlag; 2001.

22. Goldstein H. Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika. 1986;73(1):43-56.

23. Diggle PJ HP, Liang KY, Zeger SL. The analysis of longitudinal data. 2nd Edition. New York: Oxford University Press; 2002.

24. Durning SJ, Pangaro LN, Denton GD, Hemmer PA. Intersite consistency as a measurement of programmatic evaluation in a medicine clerkship with multiple, geographically separated sites. Academic Medicine. 2003;78(10):S36-S8.

25. Hasnain M, Connell KJ, Downing SM, Olthoff A. Toward meaningful evaluation of clinical competence: the role of direct observation in clerkship ratings. Academic Medicine. 2004;79(10):S21-S4.

26. Cook DA, Dupras DM, Beckman TJ, Thomas KG. Effect of rater training on reliability and accuracy of mini-CEX Scores: a randomized, controlled trial. Journal of General Internal Medicine. 2009;24(1):74-9.