Generalizability of Competency Assessment Scores Across and Within Clerkships: How Students, Assessors, and Clerkships Matter

Nikki L. Bibler Zaidi, PhD, Clarence D. Kreiter, PhD, Peris R. Castaneda, MA, Jocelyn H. Schiller, MD, Jun Yang, MA, Cyril M. Grum, MD, Maya M. Hammoud, MD, Larry D. Gruppen, PhD, and Sally A. Santen, MD, PhD

Abstract

Purpose
Many factors influence the reliable assessment of medical students' competencies in the clerkships. The purpose of this study was to determine how many clerkship competency assessment scores were necessary to achieve an acceptable threshold of reliability.

Method
Clerkship student assessment data were collected during the 2015–2016 academic year as part of the medical school assessment program at the University of Michigan Medical School. Faculty and residents assigned competency assessment scores for third-year core clerkship students. Generalizability (G) and decision (D) studies were conducted using balanced, stratified, and random samples to examine the extent to which overall assessment scores could reliably differentiate between students' competency levels both within and across clerkships.

Results
In the across-clerkship model, the residual error accounted for the largest proportion of variance (75%), whereas the variance attributed to the student and student–clerkship effects was much smaller (7% and 10.1%, respectively). D studies indicated that generalizability estimates for eight assessors within a clerkship varied across clerkships (G coefficients range = 0.000–0.795). Within clerkships, the number of assessors needed for optimal reliability varied from 4 to 17.

Conclusions
Minimal reliability was found in competency assessment scores for half of clerkships. The variability in reliability estimates across clerkships may be attributable to differences in scoring processes and assessor training. Other medical schools face similar variation in assessments of clerkship students; therefore, the authors hope this study will serve as a model for other institutions that wish to examine the reliability of their clerkship assessment scores.

Accurate assessment of medical students' clinical performance is important for student learning, improvement, and advancement decisions. Medical schools use clinical grades as an indication that students have achieved adequate levels of clinical knowledge and skill.[1] Since many medical schools have transitioned to pass/fail preclinical grading systems, clinical grades are often the only grades available to distinguish among students at a given medical school.[2,3] As a result, residency programs have started to place a greater emphasis on clerkship grades in their selection processes.[4,5] Therefore, it is important to determine whether current models of assessment for core medical school clerkships can reliably reflect students' true level of clinical competency.

Although clearly important, the determination of clerkship grades poses several measurement challenges. Many clerkship grades rely on competency-based assessments from assessors, which are subject to variability,[6] especially because assessors vary in their approaches to performance assessments.[7] Furthermore, the amount of time that each assessor spends with a student varies by length of clerkship and clinical staffing schedules. The clinical setting also introduces variation because some clinical competencies are easier to demonstrate and assess, or even considered more important, in certain clerkships.[8] Finally, factors intrinsic to the assessor or the student, such as gender and assessor bias, can introduce variability in subjective assessments.[9,10] Clerkship assessments contain multiple sources of possible variation. For meaningful summative clerkship assessments to be possible, it is important to conduct a sound reliability analysis to account for sources of error variance in measurement. Generalizability (G) theory is commonly used to estimate reliability in multifaceted behavioral assessment processes,[11] and prior research has shown that G theory can be used to determine the source(s) of error within clerkship assessments.[12–15]

Despite the common use of competency-based assessments as a major component of overall clerkship grades, few studies have examined the reliability of assessments across multiple clerkships. These assessments are intended to reflect students' competencies as they progress through clerkships; therefore, it is important to examine the reliability of competency assessment scores. Furthermore, because the number of assessments per trainee can vary by clerkship, it is important to determine the minimum number of assessors necessary to achieve acceptable reliability. The aim of this study was to examine across-

Please see the end of this article for information about the authors.

Correspondence should be addressed to Nikki L. Bibler Zaidi, University of Michigan Medical School, Office of Medical Student Education, 5310 Taubman Health Sciences Library, Ann Arbor, MI 48109-5726; telephone: (734) 615-3841; e-mail: bibler@med.umich.edu.

Acad Med. 2018;93:1212–1217.
First published online April 24, 2018
doi: 10.1097/ACM.0000000000002262

clerkship reliability and to determine how many clerkship competency assessment scores were necessary to achieve an acceptable threshold of reliability within each of our core clerkships.

Method

Context

This study used data collected for the purpose of student assessment in the core third-year clinical clerkships at the University of Michigan Medical School during the 2015–2016 academic year. This academic year marked a significant change in our medical school's overall curricular model—a move toward competency-based medical education (CBME) with a corresponding assessment program.[16] Our 2015–2016 competency assessment forms represented a shift from general, norm-referenced behavioral anchors (e.g., less-than-average performance to above-average performance) to competency-based behavioral anchors that aligned with the medical school's newly adopted competencies (e.g., unsatisfactory to exemplary, with corresponding criteria). Clerkships were responsible for facilitating their own assessor training and operationalizing these competency assessment forms.

Data collection

Data were collected as part of our medical school's assessment program. For this study, we included assessment data for any student who completed a clerkship during the 2015–2016 academic year—May 2015 through April 2016 (see Table 1). Clerkships used in these analyses were family medicine, internal medicine, neurology, obstetrics–gynecology, pediatrics, and psychiatry (randomly labeled clerkships A through F). This included some students (fewer than 5% of the sample) who were off-cycle and only partially completed the academic year in the time frame of this study (e.g., left registration because of a leave of absence or entered registration after completion of their PhD). This study was determined exempt from ongoing review by the University of Michigan institutional review board.

Faculty and residents completed competency assessment forms for students in their core third-year clerkships. This form used a scale from 1 (unsatisfactory) to 5–6 (competent to proficient) to 9 (exemplary) to assign scores based on students' clinical performance in the clerkship. As we transitioned to CBME, our assessment tool combined both normative and criterion-referenced approaches to assessment; for example, we used a criterion-based scale but maintained a grading scale that employed norm-referenced standards (e.g., "one of the best 10 students I've worked with"). Assessors were asked to evaluate students on the basis of eight competency domains (communication; patient care; medical knowledge; professionalism; leadership and teamwork; systems-based practice; practice-based learning and improvement; critical thinking and discovery) and assign scores for each of these competencies as well as submit a separate *overall* assessment score per student. Therefore, the competency assessment forms included nine scores—eight domain scores and one overall score. We examined the consistency of these scores across clerkships. Because scoring was consistent across the eight domains and overall scores (Cronbach alpha for mean domain and overall scores = 0.981), we simplified our methods to only include the overall scores in this generalizability analysis.

We found that assessment practices varied considerably across clerkships. Clerkship directors asked faculty and residents to complete competency assessment forms and assign student scores based on the specific services or teams to which they were assigned. This meant that the combinations of assessors (e.g., proportion of residents to faculty), the experience of assessors, and the number of unique assessors varied across clerkships (see Table 1). In addition, each clerkship's general assessment methods, assessor training, and calibration processes varied. Consequently, there was variability in how many competency assessment forms were completed and who completed them. Many medical schools face similar variation in assessments of clerkship students.[5,17,18]

Sampling

We extracted data for all students who completed at least one third-year clerkship during the 2015–2016 academic year. Data subsetting was conducted in SPSS statistical software, version 22 (IBM Corp., Armonk, New York) to create balanced, stratified, and random samples

Table 1

Descriptive Summary of Student and Assessor Clerkship Data, From a Study of Across-Clerkship and Within-Clerkship Assessment Reliability, University of Michigan Medical School, 2015–2016

	Original data (before sampling)			Sampled data (6 clerkships, 3 random scores/clerkship)		
Students (p)	Mean (SD)	No. of cases	No. of students	Mean (SD)	No. of cases	No. of students
All 6 clerkships	6.13 (1.26)	7,236	185	6.15 (1.24)	1,674	93
Clerkship A	6.19 (1.19)	751	178	6.24 (1.13)	279	93
Clerkship B	6.31 (1.34)	1,974	169	6.26 (1.44)	279	93
Clerkship C	5.73 (1.04)	1,008	167	5.76 (1.04)	279	93
Clerkship D	6.30 (1.22)	1,068	173	6.38 (1.21)	279	93
Clerkship E	5.68 (1.31)	1,256	169	5.67 (1.24)	279	93
Clerkship F	6.47 (1.08)	1,179	172	6.58 (1.09)	279	93
Assessors (a)	**No. of scores provided, range**	**No. of unique assessors**		**No. of scores provided, range**	**No. of unique assessors**	
All 6 clerkships	1–55	984		1–16	614	
Clerkship A	1–17	116		1–8	84	
Clerkship B	1–25	424		1–5	186	
Clerkship C	2–46	64		1–13	57	
Clerkship D	1–55	123		1–16	92	
Clerkship E	1–35	155		1–8	114	
Clerkship F	1–41	119		1–10	87	

for the analysis. This data subsetting and sampling allowed us to achieve a balanced design that would maximize accuracy of our estimates.[19] A "balanced sample" means that there was the same number of assessments for all students for all clerkships. This was also necessary to run the generalizability analysis in SPSS using G1 program files provided by Mushquash and O'Connor.[20]

This study largely replicated methods employed by Kreiter and Ferguson[19] for sampling our data. Our criterion for inclusion in the analysis sample was three scores or more per student for each clerkship. We chose three scores per student and clerkship to maximize the overall sample size while maintaining a sufficient number of observations per student. This meant that six of our seven clerkships were included in this analysis; one clerkship (surgery) was excluded because only two overall assessment scores per student were available. Some students had more than three scores per clerkship; therefore, random sampling was used to select three overall scores for each student. The final sample, after subsetting the data and randomly selecting only three scores for each student, was reduced by 50% (n = 93). Therefore, only half of the available students were included in this analysis.

Generalizability study

Generalizability analysis was used to determine the extent to which assessors' scores could reliably differentiate between students' level of performance. First, we estimated variance components attributable to students (p), clerkships (c), the student × clerkship interaction (pc), and the residual assessor (a) nested (:) within student by clerkship interaction (a:pc). This allowed us to generalize across specific clerkship combinations. In each sample, all students (p) completed all clerkships (c), which created a crossed effect of students by clerkship (pc). However, the assessors (a) assigned to assess a student (p) within a clerkship (c) varied. This meant that assessors were nested within a student by clerkship model, a:(p × c). Clerkship (along with assessor) was treated as a random effect because we were interested in generalizing our findings beyond both the assessors and the clerkship combinations comprising each sample.

Next, we used a within-clerkship model to generalize within each clerkship. This assessor-nested-within-student model (a:p) allowed us to determine the number of assessors (i.e., scores) needed for acceptable reliability within each clerkship. In our dataset, assessors only evaluated the *same* student *once* but could assign scores to several students. Similar to the across-clerkship generalization, assessor was considered a random effect.

Using G study estimates, a series of decision (D) studies were conducted. D studies estimate the effects of different measurement designs (e.g., assessor sample sizes) in an attempt to find one that best minimizes unwanted variance and increase reliability.[11] For this study, G = 0.70 was used as the threshold for acceptable reliability in this study; this standard aligns with other studies examining assessment data.[19,21,22]

Results

As presented in Table 1, the sampled data means and standard deviations were similar to the original data; however, the number of cases (scores for students) and overall sample size varied for both students and assessors.

G study

For the across-clerkship analysis, Table 2 shows that variations in students' scores from one clerkship to another (pc) accounted for 10.1% of the variance. Only a small proportion of variance (7%) was attributable to the student (p)—the true score variance. Similarly, a small percentage of variance (7.9%) was attributable to the systematic effect of the clerkship (c), which was a constant effect for all students due to behavioral inconsistencies from one clerkship to another. The largest source of variance in scores (75%) was attributable to the residual error (a:pc). For within-clerkship analysis, the proportion of variance attributable to the student (p) varied by clerkship (Table 3); however, the residual error (a:p) was consistently at least twice as large as the student variance (p).

D study

Using the variance component estimates provided by our G studies, we conducted a series of D studies to estimate both relative error and G coefficients under varying conditions of measurement. On average, our clerkships assigned approximately eight assessors per student. When examining generalizability for eight assessors within a clerkship, we found that G coefficients varied across clerkships (G coefficients range = 0.000–0.795, Table 3). To further illustrate the variation among clerkships' generalizability, we examined the number of assessors needed to achieve G = 0.70, which we considered our standard for acceptable reliability.

Table 2

Generalizability and Decision Study Results for Across-Clerkship Analysis, From a Study of Across-Clerkship and Within-Clerkship Assessment Reliability, University of Michigan Medical School, 2015–2016

	G study			D study			
Effect[a]	*df*	Variance component	Proportion of variance, %	No. of assessors	No. of clerkships	G coefficient	Relative error
p	92	0.110	7.0	1	8	0.397	0.166
c	5	0.123	7.9	2	4	0.371	0.186
pc	160	0.158	10.1	4	2	0.327	0.133
a:pc	1,116	1.171	75.0	8	1	0.265	0.304

Abbreviations: G indicates generalizability; D, decision.
[a]Variation components attributable to students (p), clerkships (c), the student × clerkship interaction (pc), and the residual assessor (a) nested (:) within student by clerkship interaction (a:pc).

Table 3

Generalizability and Decision Study Results for Within-Clerkship Analysis, From a Study of Across-Clerkship and Within-Clerkship Assessment Reliability, University of Michigan Medical School, 2015–2016

	G study			D study	
Effect[a]	*df*	Variance component	Proportion of variance, %	G coefficient (with 8 assessors)	Relative error (with 8 assessors)
Clerkship A					
p	92	0.383	29.8	0.773	0.113
a:p	186	0.903	70.2	—	—
Clerkship B					
p	92	0.444	21.5	0.686	0.203
a:p	186	1.624	78.5	—	—
Clerkship C					
p	92	0.150	13.8	0.561	0.117
a:p	186	0.939	86.2	—	—
Clerkship D					
p	92	0.146	9.9	0.469	0.165
a:p	186	1.323	90.1	—	—
Clerkship E					
p	92	0.503	32.6	0.795	0.130
a:p	186	1.039	67.4	—	—
Clerkship F					
p	92	0.000	0	0.000	0.150
a:p	186	1.201	100	—	—

Abbreviations: G indicates generalizability; D, decision.
[a]Variation components attributable to students (p) and the residual assessor (a), for an assessor-nested-within-student model (a:p).

As shown in Table 4, the number of assessors needed for optimal reliability within a clerkship varied (4–17 assessors per clerkship). Although these findings suggest that some clerkships may need fewer competency assessment scores than others, no single score was enough to achieve adequate reliability. In fact, four scores per assessor was the minimum needed for any clerkship to achieve an acceptable threshold of reliability.

Table 4

Number of Assessors, by Clerkship, Necessary for Acceptable Reliability, From a Study of Across-Clerkship and Within-Clerkship Assessment Reliability, University of Michigan Medical School, 2015–2016[a]

Clerkship	No. of assessors
Clerkship A	5
Clerkship B	7
Clerkship C	12
Clerkship D	17
Clerkship E	4
Clerkship F	—[b]

[a]Standard for acceptable reliability: G = 0.70.
[b]No variance was attributed to the object of measurement—the student—so no reliability can be estimated.

Discussion

We observed substantial variability across assessors and across clerkship settings. Overall, we found that for most clerkships, only a small proportion of variance in competency assessment scores was attributable to the student. Instead, most of the variance in scores was attributable to the residual error, which consisted of undifferentiated error variance that included both assessors and other unmodeled sources of error. This made it impossible to determine whether assessors differed in their levels of stringency or leniency (i.e., an assessor effect) or whether some assessors were just more stringent or lenient for particular students (i.e., a student–assessor effect). Because of the large proportion of error variance, we also found that the number of assessors needed to provide reliable assessments of our clerkship students ranged from 4 to 17.

Implications

These findings provide an opportunity to improve not only the reliability of our competency assessment scores but also the validity of these scores. Given the dramatic differences in the level of reliability across clerkships, further investigation into *what* contributed to high and low reliability within clerkships can help us better understand what might be an effective intervention to improve the accuracy of these assessment scores. Greater reliability and validity in our clerkship assessments is important because these scores largely determine grades, which affect graduation and residency placements for medical students.

Over the last few years, our understanding about the etiology of rater error in the clinic has advanced considerably. The large student–assessor effects found in this study and others[23] demonstrate that expert clinical raters may be using gestalt-based impressions to gauge clinical skills. Such global impressions from the clinical experts are likely mediated by subconscious processes and formed early in the student–preceptor interaction; yet, they may still be valid. Although these ratings may not be based on observable actions that can be recorded on a checklist, they are likely grounded in expert assessors' deep insight regarding competence.[24] On the contrary, less experienced assessors focus more on specific, discrete aspects of student performance, thereby employing more literal descriptions of student behavior.[24] We do not know or account for the experience of assessors in our analysis, although some assessors, such as new residents, are likely to be inexperienced. Consequently, it is quite possible that our assessors approached their assessments differently simply according to their level of experience.

We also recognize that it may not be possible to measure medical competencies as stable and homogenous traits that can be assessed independently of each other.[25] The observed variability may simply be a function of an assessment form that attempts to measure

broad and complex constructs. Moreover, the model used to estimate the variance may not fully define the scoring process. For example, when we assume that students' competencies are a stable trait, any deviation in observations of these competencies will be treated as error variance.[25]

Yet another possibility is that the variance actually *does* represent true score differences among assessors' assessment of students. That is, assessors may be accurately differentiating among students' competencies and the differences are therefore due to context, which could explain the large proportion of variance attributed to the assessor effects. If this is true, this could be an important observation about these learners and the development of competency. Similar to Kreiter and Ferguson's[19] study, we found variability in our clerkships' assessment of students; yet, we found more than twice as much variation attributable to the student–clerkship effect (pc)—10.1% compared with Kreiter and Ferguson's 4.5%. As we consider the results of our study and the variability among clerkships, we have made some observations regarding our internal processes that may support these findings. In our family medicine clerkship, students often spend up to four weeks with faculty in clinic. Thus, there is prolonged contact time in a uniform context, which may contribute to more reliable assessments. In contrast, students on obstetrics–gynecology clerkships are assigned to multiple services with limited regular contact with assessors. This may result in less reliable assessment of students. These differences in context and overall assessor contact hours may affect the reliability of assessment scores. As we review our clerkship assessment practices, we plan to use our findings as impetus for change. First, we will continue to broaden our perspectives on clinical performance assessments—shifting from reliance on numbers to more narrative and constructivist approaches to assessments.[26–28] This will help support robust programmatic assessment and ensure that no single data point should be overemphasized.[29] Second, more standardized scoring processes across clerkships that include rigorous and uniform assessor training may help increase consistency. By working with individual assessors, we can better control for effects of assessor leniency or stringency. Therefore, we will also consider using a limited number of master assessors who are highly trained. Further, as we work to provide feedback to our clerkships and individual assessors, generating a shared mental model within and across clerkships will be key. This shared mental model should help ensure consistency regarding the knowledge and skills we expect of our students as they work to achieve competency across all domains.[30] Lastly, we plan to closely examine our assessment items and item-level variance as part of our ongoing quality improvement. This includes better understanding how assessors use our current competency assessment form—including whether they are able to distinguish among the competencies or whether they view the form as a global rating assessment.

Limitations

Although these findings are important, our study has some limitations. First, assessors often evaluated more than one student in each sample; however, we were unable to estimate the systematic assessor effects because of the nested nature of these data. Therefore, all assessor-related effects in both the across-clerkship and within-clerkship analyses were ultimately confounded with the residual error due to nesting. Also, our G study models treated the clerkship facet as random. This suggests that one clerkship is interchangeable with another clerkship, an assumption that may be debatable. Therefore, these methods could be replicated by treating clerkship as a fixed facet. Furthermore, we chose to simplify our analysis by using only the overall assessment scores. Although the high Cronbach alpha (0.981) we report suggests that the eight competency scores and overall score are correlated and that variance attributable to student-by-item variance is likely very small, this was not confirmed in our current study. We believe that students were assessed in a gestalt fashion; however, if this is true, disparate domains such as medical knowledge, professionalism, and communication are assumed to be the same construct—which is not a valid assumption. Thus, it is important to remember that consistency, reliability, and accuracy are not always connected.

In the context of grading, our study does not take into account potential weighting of assessment scores, with assessors' scores weighted proportionally based on contact with students. By using raw assessment scores, we also did not account for the leniency and stringency of our assessors, which varies across clerkships. For example, the mean overall scores for clerkship E suggest greater overall assessor stringency (mean = 5.66) compared with the mean scores for clerkship F (mean = 6.45). Clerkship directors also rely heavily on assessors' comments when making grading decisions. Therefore, although it is important to generate reliable clerkship assessment scores, the raw scores used in this analysis do not fully represent the data used to inform decisions regarding students' competency. Instead, these scores represent part of our larger programmatic assessment.

Conclusion

The literature suggests that other medical schools face similar challenges in generating reliable assessments of clerkship students.[5,17,18] Therefore, we recognize that the low reliability estimates for our clerkships' competency assessment scores are not unique to our institution. We believe that these findings suggest that developing a reliable assessment program to align with a competency framework may prove difficult to implement. Nonetheless, the results of this study can be used to initiate changes to some of our medical community's common assessment processes. In sum, we hope this study will serve as a model for other institutions that wish to examine the reliability of their clerkship assessment scores.

Funding/Support: None reported.

Other disclosures: The University of Michigan Medical School has an Accelerating Change in Medical Education Grant from the American Medical Association.

Ethical approval: The authors received Notice of Exemption from the University of Michigan Institutional Review Board on June 21, 2017 (HUM00131288).

N.L.B. Zaidi is associate director, Evaluation and Assessment, Office of Medical Student Education, University of Michigan Medical School, Ann Arbor, Michigan.

C.D. Kreiter is professor, Office of Consultation and Research in Medical Education, University of Iowa Carver College of Medicine, Iowa City, Iowa.

P.R. Castaneda is first-year medical student, University of Michigan Medical School, Ann Arbor, Michigan.

J.H. Schiller is associate professor of pediatrics and director of pediatric student education, University of Michigan Medical School, Ann Arbor, Michigan.

J. Yang is statistician in evaluation and assessment, Office of Medical Student Education, University of Michigan Medical School, Ann Arbor, Michigan.

C.M. Grum is professor and senior associate chair, Undergraduate Medical Education, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan.

M.M. Hammoud is professor of obstetrics and gynecology and of medical education, University of Michigan Medical School, Ann Arbor, Michigan.

L.D. Gruppen is professor, Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, Michigan.

S.A. Santen is senior associate dean of assessment, evaluation, and scholarship, Virginia Commonwealth University School of Medicine, Richmond, Virginia. At the time this study was conducted, she was assistant dean, Educational Research and Quality Improvement, Office of Medical Student Education, and associate professor and chair of education, Department of Emergency Medicine, University of Michigan Medical School, Ann Arbor, Michigan.

References

1 Alexander EK, Osman NY, Walling JL, Mitchell VG. Variation and imprecision of clerkship grading in U.S. medical schools. Acad Med. 2012;87:1070–1076.
2 McDuff SG, McDuff D, Farace JA, Kelly CJ, Savoia MC, Mandel J. Evaluating a grading change at UCSD School of Medicine: Pass/fail grading is associated with decreased performance on preclinical exams but unchanged performance on USMLE Step 1 scores. BMC Med Educ. 2014;14:127.
3 Spring L, Robillard D, Gehlbach L, Simas TA. Impact of pass/fail grading on medical students' well-being and academic outcomes. Med Educ. 2011;45:867–877.
4 Lipman JM, Schenarts KD. Defining honors in the surgery clerkship. J Am Coll Surg. 2016;223:665–669.
5 Green M, Jones P, Thomas JX Jr. Selection criteria for residency: Results of a national program directors survey. Acad Med. 2009;84:362–367.
6 Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med. 1998;73:1294–1298.
7 Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract. 2013;18:325–341.
8 Wimmers PF, Kanter SL, Splinter TA, Schmidt HG. Is clinical competence perceived differently for student daily performance on the wards versus clerkship grading? Adv Health Sci Educ Theory Pract. 2008;13:693–707.
9 Lee V, Brain K, Martin J. Factors influencing mini-CEX rater judgments and their practical implications: A systematic literature review. Acad Med. 2017;92:880–887.
10 Riese A, Rappaport L, Alverson B, Park S, Rockney RM. Clinical performance evaluations of third-year medical students and association with student and evaluator gender. Acad Med. 2017;92:835–840.
11 Brennan RL. Generalizability Theory. New York, NY: Springer-Verlag; 2001.
12 Richter RA, Lagha MA, Boscardin CK, May W, Fung CC. A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. Acad Med. 2012;87:1077–1082.
13 Al-Mahroos F. Construct validity and generalizability of pediatrics clerkship evaluation at a problem-based medical school, Bahrain. Eval Health Prof. 2009;32:165–183.
14 Blood AD, Park YS, Lukas RV, Brorson JR. Neurology objective structured clinical examination reliability using generalizability theory. Neurology. 2015;85:1623–1629.
15 Boodoo GM, O'Sullivan PS. Assessing pediatric clerkship evaluations using generalizability theory. Eval Health Prof. 1986;9:467–486.
16 Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32:676–682.
17 Plymale MA, French J, Donnelly MB, Iocono J, Pulito AR. Variation in faculty evaluations of clerkship students attributable to surgical service. J Surg Educ. 2010;67:179–183.
18 Fay EE, Schiff MA, Mendiratta V, Benedetti TJ, Debiec K. Beyond the ivory tower: A comparison of grades across academic and community OB/GYN clerkship sites. Teach Learn Med. 2016;28:146–151.
19 Kreiter CD, Ferguson KJ. Examining the generalizability of ratings across clerkships using a clinical evaluation form. Eval Health Prof. 2001;24:36–46.
20 Mushquash C, O'Connor BP. SPSS and SAS programs for generalizability theory analyses. Behav Res Methods. 2006;38:542–547.
21 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: A key to unlock professional assessment. Med Educ. 2002;36:972–978.
22 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. JAMA. 1993;269:1655–1660.
23 Kreiter CD, Wilson AB, Humbert AJ, Wade PA. Examining rater and occasion influences in observational assessments obtained from within the clinical environment. Med Educ Online. 2016;21:29279.
24 Govaerts MJ, Schuwirth LW, van der Vleuten CP, Muijtjens AM. Workplace-based assessment: Effects of rater expertise. Adv Health Sci Educ Theory Pract. 2011;16:151–165.
25 Schuwirth LW, van der Vleuten CP. A plea for new psychometric models in educational assessment. Med Educ. 2006;40:296–300.
26 Moonen-van Loon JM, Overeem K, Govaerts MJ, Verhoeven BH, van der Vleuten CP, Driessen EW. The reliability of multisource feedback in competency-based assessment programs: The effects of multiple occasions and assessor groups. Acad Med. 2015;90:1093–1099.
27 Hanson JL, Rosenberg AA, Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. Front Psychol. 2013;4:668.
28 Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. Acad Med. 1999;74:1203–1207.
29 van der Vleuten CPM, Schuwirth LWT, Driessen EW, Govaerts MJB, Heeneman S. Twelve tips for programmatic assessment. Med Teach. 2015;37:641–646.
30 Lomis KD, Russell RG, Davidson MA, Fleming AE, Pettepher CC. Competency milestones for medical students: Design, implementation, and analysis at one medical school. Med Teach. 2017;39:494–504.