Taking Rater Exposure to Trainees Into Account When Explaining Rater Variability

Christy K. Boscardin, PhD
Marjo Wijnen-Meijer, PhD
Olle ten Cate, PhD

ABSTRACT

Background Rater-based judgments are widely used in graduate medical education to provide more meaningful assessments, despite concerns about rater reliability.

Objective We introduced a statistical modeling technique that corresponds to the new rater reliability framework, and present a case example to provide an illustration of the utility of this new approach to assessing rater reliability.

Methods We used mixed-effects models to simultaneously incorporate random effects for raters and systematic effects of rater role as fixed effects. Study data are clinical performance ratings collected from medical school graduates who were evaluated for their readiness for supervised clinical practice in authentic simulation settings at 2 medical schools in the Netherlands and Germany.

Results The medical schools recruited a maximum of 30 graduates out of 60 (50%) and 180 (17%) eligible candidates, respectively. Clinician raters (n = 25) for the study were selected based on their level of expertise and experience. Graduates were assessed on 7 facets of competence (FOCs) that are considered important in supervisors' entrustment decisions across the 5 cases used. Rater role was significantly associated with 2 FOCs: (1) teamwork and collegiality, and (2) verbal communication with colleagues/supervisors. For another 2 FOCs, rater variability was only partially explained by the role of the rater (a proxy for the amount of direct interaction with the trainee).

Conclusions Consideration of raters as meaningfully idiosyncratic provides a new framework to explore their influence on assessment scores, which goes beyond considering them as random sources of variability.

Introduction

Assessment of clinical performance of learners and physicians in a real practice context is critically important, but issues of reliability and feasibility make it a challenging task. Although the utility and the importance of alternative approaches to assessment, including workplace-based assessment, multisource feedback, and interprofessional teamwork assessments, are widely recognized, outstanding issues surrounding their reliability compromises the potential utility and adoptability of these methods for summative assessment purposes.[1,2]

Recently, normalization of ratings data has been proposed as a way to reduce bias, but this approach still does not address how to best interpret and investigate the sources of these inconsistencies.[3] In a recent review, Gingerich et al[4] introduced 3 perspectives, 1 of which is "the assessor as meaningfully idiosyncratic." Specifically, rater perceptions of a trainee's performance are based on outcomes of complex interplay between the trainee, the rater, and the environment.[5–7] Accordingly, the various observers interacting with residents will be privy to different sets of observations, depending on their role and interaction with the trainee. Recent studies by Govaerts et al[8] and Gingerich et al[4,9] provide greater insight into the underpinnings of rater behavior to help understand that what appears to be idiosyncrasies, on closer examination may reveal more systematic features of rater perception. In alignment with this view, we believe these idiosyncrasies should be accommodated and may reflect important differences.[10–12]

From a traditional psychometric view, idiosyncrasies in ratings reduce reliability, and should be minimized through various methods (ie, rater training, consensus rating). Contrary to this view, we propose that instead of trying to decrease diversity of perspectives, the reliability estimates should allow for this variability, and encompass rater factors (ie, rater characteristics) to help explore and examine the idiosyncrasies. The inclusion of rater factors into the reliability estimate provides several advantages: (1) it allows for variability of raters without arbitrarily forcing the ratings to consensus; (2) it helps explain the sources of variability in the performance ratings; and (3) it may increase overall reliability. We propose using a mixed-effects model to incorporate both random effects for raters and systematic effects of rater characteristics (ie, rater role, rater experience,

DOI: http://dx.doi.org/10.4300/JGME-D-16-00122.1

rater conditions, etc) as fixed effects to gain more detailed information around rater variability.

In this article we introduce a statistical modeling technique that corresponds to this new rater reliability framework using a case example as an illustration of the utility of this new approach.

What was known and gap
Rater-based judgments are widely used in graduate medical education, despite concerns about reliability.

What is new
A modeling technique that uses rater roles as a legitimate component in differences among ratings, creating a new reliability framework.

Limitations
High reliability of ratings reduced the ability to show gains that would result from the new framework.

Bottom line
Considering raters' differences as meaningful sources of variance provides a new framework for analyzing their influence on assessments.

Methods

To explore the effect of including rater descriptions on the reliability estimates, we used clinical performance rating data collected from recent medical school graduates evaluated for their readiness for supervised clinical practice in authentic simulation settings in the Netherlands and Germany. A brief description of the study setting and assessment procedure is provided in the sections that follow; a more detailed information is provided in Wijnen-Meijer et al.[13]

Setting

The graduates who had just completed undergraduate medical education at medical schools in Utrecht, the Netherlands, and Hamburg, Germany, participated in a simulated environment in the role of a beginning resident on a busy inpatient unit. The assessment consisted of 3 phases. First, graduates encountered 5 standardized patients (SPs) portraying patients with uncommon medical problems who had just been admitted to the hospital. In the second phase, after the patient encounters, graduates were given time to request lab results and gather additional information to determine differential diagnoses along with a management plan for each patient to present to the supervisor at the end of the day. During this phase, graduates also were given the opportunity to call their supervisors by phone if needed, and they also had a brief face-to-face meeting with the supervisor to discuss progress. During the third phase, graduates were given 30 minutes to present their differential diagnoses and management plans for the 5 SPs.

Participants

Each of the 2 schools recruited a maximum of 30 graduates from each institution, representing 50% out of 60 and 17% out of 180 eligible candidates. Clinician raters (n = 25) were selected to participate in the study based on their level of expertise and experience. Each graduate was assessed by 3 raters, and the raters had 3 distinct roles: (1) acting as the graduate's personal supervisor during the assessment; (2) being present for the entire simulation and listening to telephone and face-to-face conversations between the supervisor and the graduate (but without direct contact with the graduate); and (3) observing the graduate only during the final reporting phase. The 3 role categories represent the rater roles typically encountered during graduate training, and were included in the analysis as fixed effects. All 25 raters participated in each of the 3 rater role categories by random rotation.

Assessments

Raters were asked to rate the overall performance on a 5-point scale from 1 (weak) to 5 (very good) on 7 facets of competence (FOCs) considered to be key components in making entrustment decisions by supervisors about the residents across 5 SP cases. The FOCs inform the evaluation of entrustable professional activities in the larger study, which also follows the 5-level entrustable professional activity supervision scale.[13] The 7 FOCs that were rated included (1) scientific and empirical grounded method of working; (2) knowing and maintaining one's own personal bounds and possibilities; (3) teamwork and collegiality; (4) verbal communication with colleagues and supervisors; (5) responsibility; (6) safety and risk management; and (7) active professional development. All 60 trainees were rated on the 7 FOCs by 3 raters representing 3 different levels of interaction with the trainee.

The Netherlands Association for Medical Education Ethical Review Board and the State of Hamburg Physicians Ethics Board provided ethical approval for the study.

Analysis

First, we estimated the variance components for rater effect by using a random-effects model as the baseline model. This represents the traditional approach to estimating reliability.

Step 1: Random-Effects Model

$$Y_{ij} = \beta_0 + \beta_i + \beta_{r(ij)} + e_{ij} \quad (1)$$

TABLE
Variance Components for the 7 Ratings With and Without Rater Characteristics

Competencies	Random-Effects Model (Baseline)			Mixed-Effects Model With Rater Characteristics as Fixed Effects			Change in Variance Associated With Raters
	Rater	Trainee	Error	Rater	Trainee	Error	Change Between Baseline and Mixed Effects
1. Scientific and empirical grounded method of working	0.12	0.54	0.34	0.12	0.54	0.34	0%
2. Knowing and maintaining one's own personal bounds and possibilities	0.13	0.51	0.36	0.13	0.51	0.36	0%
3. Teamwork and collegiality	0.20	0.41	0.39	0.18	0.42	0.40	–2.0%
4. Verbal communication with colleagues and supervisors	0.19	0.46	0.35	0.15	0.48	0.37	–4.0%
5. Responsibility	0.01	0.59	0.40	0.01	0.59	0.39	0%
6. Safety and risk management	0.18	0.42	0.40	0.17	0.43	0.40	–1.0%
7. Active professional development	0.00	0.61	0.39	0.00	0.61	0.39	0%

Where β_0 = the average rating; β_i = the trainee random effect; $\beta_{r(ij)}$ = the rater random effect; and e_{ij} = random errors. This provided information about the variance components associated with rater, trainee, and error. Then, we employed a mixed-effects model to include the rater role (the amount and type of interaction with the trainee as described in the participant section) as fixed effects in addition to the baseline random-effects model. The purpose of adding the rater role as a fixed effect is to help explain the variability in raters and decrease the variance components related to raters (thus increasing reliability).

Step 2: Mixed-Effects Model With Rater Role as Fixed Effects

$$Y_{ij} = \beta_0 + \beta_i + \beta_{r(ij)} + e_{ij} \quad (2)$$

$$\beta_{r(ij)} = \gamma_0 * (\text{Rater role}_{r(ij)}) + \gamma_{r(ij)}$$

Where β_0 = the average rating; β_i = the student random effect; e_{ij} = random errors; $\beta_{r(ij)}$ = the rater effect; γ_0 = fixed effect of rater role; and $\gamma_{r(ij)}$ = random effect of rater.

In Step 3, we used the estimates from the regression analysis and mixed-effects models to derive estimates of the variance components analogous to a generalizability study approach. In doing so, we expected an overall increase in reliability. A more detailed description of procedures for estimating variance components and reliability coefficients from regression estimates is provided by Shavelson and Webb.[14] We used Stata version 13.0 (StataCorp LP, College Station, TX) for all statistical analyses.

Results

All trainees were rated by 3 raters representing 3 different levels of interaction with the trainee on a 5-point scale on the 7 FOCs, including: (1) scientific and empirical grounded method of working; (2) knowing and maintaining one's own personal bounds and possibilities; (3) teamwork and collegiality; (4) verbal communication with colleagues and supervisors; (5) responsibility; (6) safety and risk management; and (7) active professional development.

Variance Components

The variance components associated with raters for the 7 FOCs ranged from 0% to 20%, indicating overall high rater reliability. As shown in the TABLE, the 2 competency domains with the highest rater variability (low reliability) were teamwork and collegiality (domain 3, 20% of the total variance) and verbal communication with colleagues and supervisors (domain 4, 19% of the total variance).

Rater Category Effect

In the mixed-effects model, with the inclusion of the 3 rater role categories as a fixed effect, we examined the effect of rater characteristics on rater variability. Rater role was significantly associated with the same 2 FOCs that had highest rater variability (teamwork and collegiality; verbal communication with colleagues and supervisors). In both instances, the third rater role (only observing the reporting phase) was associated with significantly lower ratings on both teamwork and collegiality ($\beta = -0.38$, $P = .049$) and verbal communication with colleagues ($\beta = -0.48$, $P = .021$).

Differences in the Variance Components

For the 2 FOCs with the significant rater role effect, the variance components representing the rater variability decreased. The variance component representing the rater effect for the teamwork and collegiality domain was reduced from 20% to 18%, and for the verbal communication domain, the variance component for raters decreased from 19% to 15% after including the rater role in the model. This translated into an increase of about 0.20 in overall reliability from 0.78 to 0.80 for the verbal communication FOC.

Discussion

Conceptualizing raters as meaningfully idiosyncratic provides an alternative framework to explore the role of raters in the interpretation of the assessment scores, and goes beyond just considering them as a random source of variability. With this framework, the focus shifts from consistency to understanding the source of variability and the attributes of the raters. In this study, the overall reliability of the assessment was increased slightly by taking into account the rater role (eg, amount of exposure to trainee). This finding may suggest additional evidence toward requirement for a minimum amount of exposure prior to feedback to increase overall meaningfulness of the rating.

Despite the relatively limited finding in the current study due to limitations of the data, using the mixed-effects model may help explain some of the rater variability by taking systematic characteristics of the raters into account. Including these characteristics (eg, rater role, amount of contact with trainee) in the analysis may increase the overall rater reliability.

Second, the perspective that raters are meaningfully idiosyncratic suggests allowing for examination of variability rather than arbitrarily standardizing the ratings. If we find that nurses and physicians provide consistently different ratings, then, due to the differences in their environmental roles, the reliability estimates need to be able to represent these differences appropriately. By including these characteristics or differences as part of the rater reliability analysis, we could provide various subscores representing these different perspectives. Identification of the key rater background and factors attributing to the differences in rater perception will be critical to the application of this new reliability framework. Recent studies should spark further discussion and development in this area.[8,9,15]

This study has several limitations. First, the data used for illustration purposes may not be representative of the typical observational ratings encountered in the workplace. However, despite being a simulated case, our novel assessment format was developed to represent the complexity and the unpredictability of the typical clinical setting. This was done to maximize authenticity of the experience, as well as to simulate typical expectations of raters with often limited opportunities for direct observations. Second, the rater characteristic data available were limited to raters' specific role, which was a proxy for the amount of contact with the trainee. Additional background information about the raters would have provided a richer example and probably more significant results. Also, given the high reliability of the ratings, it was difficult to illustrate the maximum potential utility of the method using the current data. Lastly, for illustration purposes, the example was kept purposely simple by excluding the other facets in the model, such as cases and items. In future studies, the issue of case specificity and the relationship between cases and rater characteristics should be explored more in detail.

Conclusion

As we move toward competency-based education with increased emphasis on work-based and interprofessional assessments, we will need a new framework for considering rater reliability. Our approach to rater reliability may provide ways to maximize the information derived from the variability in raters.

References

1. Driessen EW, van Tartwijk J, Govaerts M, et al. The use of programmatic assessment in the clinical workplace: a Maastricht case report. *Med Teach*. 2012;34(3):226–231.
2. McGill DA, Van der Vleuten CP, Clarke MJ. Supervisor assessment of clinical and professional competence of medical trainees: a reliability study using workplace data and a focused analytical literature review. *Adv Health Sci Educ Theory Pract*. 2011;16(3):405–425.
3. Baker K. Determining resident clinical performance: getting beyond the noise. *Anesthesiology*. 2011;115(4):862–878.
4. Gingerich A, Kogan J, Yeates P, et al. Seeing the "black box" differently: assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–1068.
5. Lance CE, Butts MM, Michels LC. The sources of four commonly reported cutoff criteria: what did they really say? *Org Res Methods*. 2006;9(2):202–220.
6. Murphy KR, Cleveland JN, Skattebo AL, et al. Raters who pursue different goals give different ratings. *J Appl Psychol*. 2004;89(1):158–164.
7. Ginsburg S, Regehr G, Lingard L, et al. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ*. 2015;49(3):296–306.
8. Govaerts MJ, Schuwirth LW, Van der Vleuten CP, et al. Workplace-based assessment: effects of rater expertise.

Adv Health Sci Educ Theory Pract. 2011;16(2):151–165.
9. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Acad Med*. 2011;86(suppl 10):1–7.
10. Cianciolo AT, Kegg JA. Behavioral specification of the entrustment process. *J Grad Med Educ*. 2013;5(1):10–12.
11. ten Cate O. Entrustability of professional activities and competency-based training. *Med Educ*. 2005;39(12):1176–1177.
12. Hoffman B, Lance CE, Bynum B, Gentry WA. Rater source effects are alive and well after all. *Pers Psychol*. 2010;63(1):119–151. http://psychology.uga.edu/sites/default/files/CVs/Rater%20Source%20Effects%20Are%20Alive%20and%20Well%20After%20All-1%20(1).docx. Accessed August 31, 2016.
13. Wijnen-Meijer M, Kilminster S, Van Der Schaaf M, et al. The impact of various transitions in the medical education continuum on perceived readiness of trainees to be entrusted with professional tasks. *Med Teach*. 2012;34(11):929–935.
14. Shavelson RJ, Webb NM. *Generalizability Theory: A Primer*. Vol 1. Newbury Park, CA: Sage Publications; 1991.
15. Hauer KE, ten Cate O, Boscardin C, et al. Understanding trust as an essential element of trainee supervision and learning in the workplace. *Adv Health Sci Educ Theory Pract*. 2014;19(3):435–456.

Christy K. Boscardin, PhD, is Associate Professor, Department of Medicine, School of Medicine, University of California, San Francisco; **Marjo Wijnen-Meijer, PhD,** is Assistant Professor of Medical Education, and Director, Center for Research & Development of Education, University Medical Center Utrecht, the Netherlands; and **Olle ten Cate, PhD,** is Professor of Medical Education and Director, Center for Research & Development of Education, University Medical Center Utrecht, the Netherlands.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

The authors would like to thank Pat O'Sullivan, EdD, for the helpful suggestions and feedback on the earlier versions of this paper.

Corresponding author: Christy K. Boscardin, PhD, UCSF School of Medicine, Department of Medicine, Office of Medical Education, 533 Parnassus Avenue, Suite U-80, San Francisco, CA 94143-3202, 415.519.3570, christy.boscardin@ucsf.edu

Received February 22, 2016; revision received July 6, 2016; accepted July 27, 2016.