Accuracy of Entrustment-Based Assessment: Implications for Programs and Patients

C. Jessica Dine, MD, MSHP
Lisa N. Conforti, MPH
Eric S. Holmboe, MD
Jennifer R. Kogan, MD

ABSTRACT

Background Although entrustment-supervision ratings are more intuitive compared to other rating scales, it is not known whether their use accurately assesses the appropriateness of care provided by a resident.

Objective To determine the frequency of incorrect entrustment ratings assigned by faculty and whether accuracy of an entrustment-supervision scale differed by resident performance when the scripted resident performance level is known.

Methods Faculty participants rated standardized residents in 10 videos using a 4-point entrustment-supervision scale. We calculated the frequency of rating a resident incorrectly. We performed generalizability (G) and decision (D) studies for all 10 cases (768 ratings) and repeated the analysis using only cases with an entrustment score of 2.

Results The mean score by 77 raters for all videos was 2.87 (SD=0.86) with a mean of 2.37 (SD=0.72), 3.11 (SD=0.67) and 3.78 (SD=0.43) for the scripted levels of 2, 3, and 4. Faculty ratings differed from the scripted score for 331of 768 (43%) ratings. Most errors were ratings higher than the scripted score (223, 67%). G studies estimated the variance proportions of rater and case to be 4.99% and 54.29%. D studies estimated that 3 raters would need to watch 10 cases. The variance proportion of rater was 8.5% when the analysis was restricted to level 2 entrustment, requiring 15 raters to watch 5 cases.

Conclusions Participants underestimated residents' potential need for greater supervision. Overall agreement between raters and scripted scores were low.

Introduction

Workplace-based assessment (WBA) is vital for evaluating resident performance in clinical settings.[1,2] However, rating errors, particularly those stemming from inconsistent raters, pose a significant challenge.[3,4] These errors can lead to suboptimal patient care and educational outcomes.[5] This study addresses this issue by emphasizing the critical need to understand and mitigate rating errors in WBAs, providing essential insights for program directors.

While increasing the number of assessments helps mitigate poor interrater reliability, and it is also important to understand other sources of error. This involves estimating how much of the score variation is attributed to residents, raters, or other factors through generalizability (G) and decision (D) studies. This psychometric approach aims to identify sources of variability (G studies) and how to use the results of assessments to make decisions about the learner (D studies). However, this psychometric approach doesn't guarantee accurate assessment in individual patient encounters, potentially leading to competency and care appropriateness concerns. For example, a resident may be deemed competent across several observations but may not have performed well or may not have been accurately assessed in one or more of those encounters. Therefore, aggregating the assessments does not address the competency level or appropriateness of care provided by the resident or the accuracy of the observation in a single patient encounter, potentially impacting the quality of care a patient receives.

Medical educators aim to improve WBA reliability with entrustment-supervision rating scales. These scales, based on decreasing resident supervision needs, are often more intuitive for faculty and residents.[6-9] Early research suggested faculty can more easily identify with the concept of entrustment versus competency (thereby improving interrater reliability).[9] While these scales may reduce the number of needed observations for acceptable reliability, questions about their enhanced effectiveness have emerged.[9-11]

It is not known whether the use of entrustment-supervision ratings improves the accuracy of single observations, therefore addressing the appropriateness of care provided by a resident with a patient in a single encounter. While programmatic determination of the overall competency of a resident is important, it is equally important to ensure each patient encounter provides safe, effective, and patient-centered care under the right amount of supervision.[12]

DOI: http://dx.doi.org/10.4300/JGME-D-23-00275.1

We aimed to measure the accuracy of single-encounter, entrustment-supervision scale WBAs. The main objective of this study was to determine the frequency of entrustment rating errors when the scripted resident performance is known, where we define error as a participant rating differing from the scripted rating. The second objective was to determine whether the accuracy of an entrustment rating differed by resident skill level. To compare the individual observation assessments to a more programmatic view, we also performed G and D studies to understand the performance of the WBA across all observations.

KEY POINTS

What Is Known
Use of entrustment scales is growing yet we still need to understand the psychometric implications of their use.

What Is New
Entrustment decision accuracy was measured using standardized resident performance, and levels of agreement were not always optimal.

Bottom Line
This adds to the growing body of literature around how entrustment decisions should be used in high-stakes ways.

Methods

Setting and Participants

All program directors from Accreditation Council for Graduate Medical Education-accredited family and internal medicine programs within a 5-hour drive from our study sites in Chicago and Philadelphia (324 programs from 6 Midwest and 5 Mid-Atlantic states) were invited via email to recommend eligible faculty who might have interest in participating.[13] All potential participants, whose email addresses were provided by program directors, were practicing clinicians who trained and assessed residents in the outpatient setting, were on faculty for at least one year, provided care for their own panel of patients in the outpatient setting, had not yet taken the course or participated in one of the studies about direct observation, and were available for a 2-day session. At the time of the trial, a power calculation called for a sample size of 25 per group.[13] We oversampled to account for potential participant attrition. The final 77 participants were asked to independently rate 10 standardized resident-patient video encounters using a modified 4-point prospective entrustment-supervision scale (TABLE 1).[13] Raters were given the scripted level of training (ie, postgraduate year) of the residents depicted in each of the video cases but were blinded to the scripted level of performance (entrustment scale rating). All participants completed a demographic survey.

Development of Trigger Videos and Expert Assessment

The 10 video cases used in this study were developed for a previously published randomized controlled trial depicting a standardized resident obtaining a history from or counselling a standardized patient.[13] As described in that original manuscript, each case was first rigorously scripted using the best available evidence to represent specific supervision-based entrustment levels for residents performing a history or counseling a patient across a variety of diagnoses to ensure a patient receives high quality care in the scenario.[14] Six physicians with expertise in physician-patient communication and trainee assessment, along with study authors, worked together to create a matrix of observable behaviors and skills that would be necessary to display a certain resident skill level. One investigator (J.K.) wrote trigger video scripts using the observable behaviors and skills. The experts and 2 study investigators (E.S.H., L.C.) reviewed the scripts for accuracy before filming. To finalize the entrustment level portrayed by the standardized residents in the videos after filming, the videos were reviewed by one expert who had reviewed the original script and 2 experts who had not seen the script and were blinded to the scripted performance level. Of the 10 videos, 5videos depicted a resident performing at an entrustment level of 2 (learner can practice skill with direct supervision), 3 videos depicted a level of 3 (learner can practice skill with indirect supervision), and 2 videos depicted a level of 4 (unsupervised practice allowed).[15,16]

Data Analysis

To best evaluate the common approach residency programs use to assess residents (combining multiple ratings across raters), we compared both the individual rater's and the group assessments to the scripted score for each case. We first calculated the mean

TABLE 1
Modified 4-Point Prospective Entrustment-Supervision Rating Scale

1	2	3	4
Observer only	Direct supervision	Indirect supervision	Unsupervised practice
Learner cannot practice, can observe only	Learner can practice skill with direct supervision	Learner can practice the skill with indirect supervision	Unsupervised practice allowed

score obtained from raters across all 10 cases and for cases representing each entrustment level. We compared the observed mean score across all cases and for cases at each entrustment level to the scripted score using 2-sided *t* tests. We calculated the frequency of errors, which we defined as an entrustment rating higher or lower than scripted, within and across cases. We then calculated kappa coefficients to determine the level of agreement between raters and experts.

We then performed G and D studies to mirror how a residency program may attempt to overcome poor interrater reliability.[3,4] G studies can estimate the source of variation in scores, that is, how much of the score variation is explained by the rater versus the resident skill level. We performed G studies for all cases, with the cases (ie, standardized residents) as the object of measurement. We used a one-facet crossed design (rater x case model) where raters represent the participants and cases represent the standardized residents in the videos. Since G studies use the difference of the score from the overall mean to estimate variance components, the rater variance component was recalculated using the scripted instead of the population mean to determine if this would impact the score variance attributable to the raters.[3]

Simulated D studies demonstrate how the score precision changes based on changing the number of observations; the results can be used to determine how many observations need to be obtained before a residency program can make a reliable determination of a resident's performance. We performed D studies to estimate the number of raters needed to accurately assign an entrustment rating to a case. Since raters describe more difficulty and discomfort with assessing struggling or poor performing residents,[17] G and D studies were repeated for cases scripted with a level 2 entrustment rating.

We used SPSS (Version 28.0.1) for all descriptive and comparisons analysis and urGENOVA (Version 2.1) for the G and D studies.

The institutional review board at the University of Pennsylvania approved this study.

Results

A total of 221 faculty were recommended by program directors. Of these, 31 did not respond, 40 were unable to participate, and 56 were ineligible. Fourteen dropped out post randomization and 3 after baseline data collection (due to scheduling and personal conflicts). Participant demographics are shown in TABLE 2. There were 768 (99.7% of expected) entrustment ratings in the sample. Ratings were missing from 2 participants from 2 cases.

The mean entrustment rating across all 10 cases was 2.87 (SD=0.86), which is statistically significant different from the scripted score (2.70 [SD=0.78; $P<.001$]) (FIGURE). There were statistically significant differences in the observed compared to the scripted score for cases at each entrustment level: 2.37 (SD=0.72) vs 2 ($P=<.001$); 3.11 (SD=0.67) vs 3 ($P=.015$); and 3.78 (SD=0.43) vs 4 ($P<.001$).

Of the total 768 ratings, 331 (43%) were incorrectly rated, with 223 (29%) ratings being higher than the scripted score (TABLE 3). Of the 384 ratings of the 5 cases scripted as level 2 entrustment, half of the ratings (192) were incorrect. Most of these errors (157, 82%) were a higher rating than the scripted score. The overall kappa was -0.19 for all cases (-0.26 for cases scripted to be a level 2 entrustment, -0.18 for cases scripted to be a level 3 and -0.14 for cases scripted to be a level 4).

To conduct G studies, we replaced the missing 2 values with the mean rating from the other 76 raters for the respective cases. The variance component of raters was 0.039 explaining 4.99% of the observed variation, while the cases explained 54.29% (variance component 0.424) with the residual error explaining 40.72% (variance component 0.318) (TABLE 4). D studies demonstrated that 3 raters would be needed to watch all 10 cases for a G coefficient of 0.78 (30 total observations). The rater variance would increase to 9.85% if the scripted score was used to calculate variance components rather than the observed mean. G studies were repeated limited to level 2 scripted cases (TABLE 4). Raters explained 8.5% of the variance observed. D studies estimated that 15 raters were needed to rate all 5 level 2 cases to reach a G coefficient of 0.81 (75 total observations).

Discussion

In 29% of ratings, participants underestimated residents' future supervision needs, as indicated by low agreement with experts (as seen in the low kappa scores). Notably, the error rate was higher (41%) for low-performance cases (entrustment level 2), potentially leading to inadequate supervision for 157 out of 384 patients.

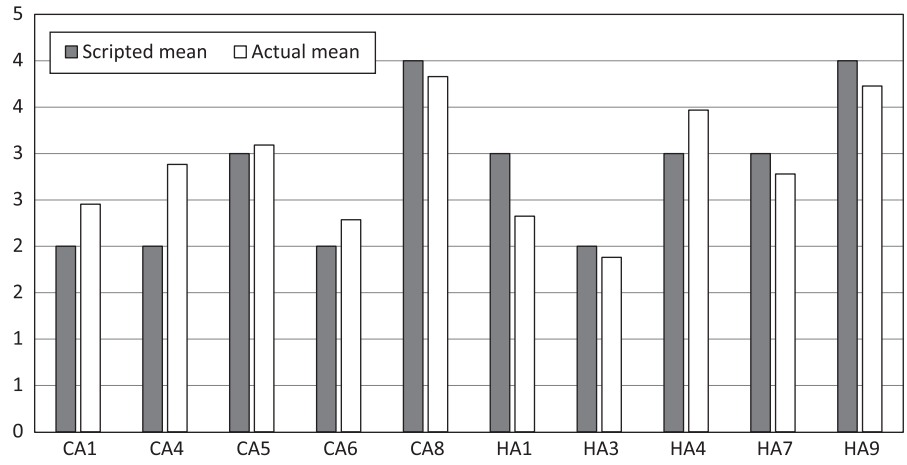
While entrustment rating errors were frequent in individual observations, our findings, supported by G and D studies, affirm the validity of aggregating observations for trainee assessment. Notably, a high-stakes decision regarding supervision levels for patient care can be made with input from just 3 faculty members observing 10 cases, although potentially up to 30 observations may still be needed. We re-evaluated generalizability using only cases scripted at an entrustment level of 2, revealing substantial variation in rater

TABLE 2
Baseline Characteristics of the 77 Participants

Characteristics	n (%)
Gender	
Woman	49 (64)
Man	28 (36)
Age in years, mean (SD)	45 (10)
Years post residency, mean (SD)	13 (11)
Completed fellowship	22 (29)
Primary specialty	
Internal medicine	48 (62)
Family medicine	29 (38)
Academic rank	
Instructor	9 (12)
Assistant professor	30 (39)
Associate professor	15 (19)
Professor	5 (7)
Other/not applicable	18 (23)
Institution type	
University-based	27 (35)
Community-based, university affiliated	29 (38)
Community-based program, non-university affiliated	20 (26)
Other[a]	1 (1)
Educational leadership roles[b]	
Program, associate, or assistant director	36 (47)
Core faculty	42 (55)
Clinical Competency Committee chair or member	28 (36)
Other (resident clinic site director, assistant/associate fellowship director, medical school clinical rotation course director, etc)	22 (29)
None	10 (13)
Outpatient ½ days per week seeing own patients, mean (SD)	3 (2)
Years precepting residents in outpatient setting, mean (SD)	11 (9)
½ days per week precepting residents in outpatient, mean (SD)	3 (2)
Precepting supervision practices: Typically see all patients when precepting …	
… interns in clinic the first 6 months of internship	70 (91)
… interns in clinic the second 6 months of internship	25 (32)
… 2nd and 3rd year residents in clinic	14 (18)
Participated in a faculty development workshop in last 5 years focused on	
Observing and assessing residents in a clinical setting	34 (44)
Competency based assessment of medical trainees	42 (55)
Giving feedback to medical trainees	53 (69)
Communication skills with patients	34 (44)

[a] Hospital but not residency has university affiliation.
[b] Participants were able to select more than one role.

influence and D study results based on resident performance level. The accuracy of entrustment scores required an increase in raters from 3 to 15 when considering all 10 cases versus only the lowest performing 5 cases. This underscores the need for caution in using entrustment scales for assessing history taking and counseling, as generalizability varies widely by performance level. These findings reinforce the challenge faculty face in assessing and providing feedback to struggling residents, compared to those performing at a higher level.[18]



FIGURE

Comparison of Scripted versus Actual Mean for each of the 10 Videos

Note: The actual mean of scores assigned by 77 raters of 10 standardized residents interacting with a standardized patient is compared to the scripted score of each of the cases (5 counselling [CA1, 4, 5, 6, and 8], and 5 history-taking [HA1, 3, 4, 7, and 9] cases).

Interestingly, the variation attributed to raters in our study was significantly lower than previous G studies using an entrustment-supervision WBA scale where raters typically explained 40% to 60% of the observed variation.[19] The factors underlying this unexpected finding are unclear. Possibilities include (1) ratings in our study occurred in a controlled setting without typical contextual factors[20,21]; (2) each scripted case level displayed relatively consistent patterns in ratings by the participating faculty: for the high performing videos (level 4) almost all faculty correctly rated the performance, but for the lowest performing videos (level 2) the majority of faculty got the rating incorrect; or (3) study participants first narratively assessed what the resident did well and what needed improvement before completing the entrustment scale. The low variation attributed to raters, however, suggests that the incorrect assignments of future entrustment may be higher in clinical learning environments where rater variation is higher.

Programs often rely on G and D studies to determine how many observations are needed to determine resident competence.[3] The calculations use the idea of dispersion or deviation from the population mean to help make these estimations. Our study is unique since we know the scripted score of each video case. Therefore, we were able to correct the deviation from the mean by using the scripted rather than calculated population mean. When we used the scripted score to recalculate the G studies, the raters explained more variation (8.50% vs 4.99%) compared to the calculated or population mean—suggesting that the rate of errors in supervision decisions is even higher.

There are several limitations. Clinical Competency Committees (CCCs) and program directors often use multiple types of evaluations to determine residents'

TABLE 3

Number and Percent of Incorrect and Correct Ratings by Faculty Participants

Scripted Entrustment Score[a]	2 (N=5 Cases)	3 (N=3 Cases)	4 (N=2 Cases)	All (N=10 Cases)
Total Ratings	384	231	153	768
Total correct ratings, n (%)	192 (50)	125 (54)	120 (78)	437 (57)
Total incorrect ratings, n (%)	192 (50)	106 (46)	33 (22)	331 (43)
Of incorrect, total, n (%) ratings higher than scripted	157 (82)	66 (62)	N/A	223 (67)
Of incorrect, total, n (%) ratings lower than scripted	35 (18)	40 (38)	33 (100)	108 (33)

[a] An entrustment level of 2=learner can practice skill with direct supervision; 3=learner can practice skill with indirect supervision; 4=unsupervised practice allowed.

TABLE 4
The Percentage of the Contribution of Raters as Estimated by Using Generalizability Studies and Results of Decision Studies

Mean Rating	All Cases Using Population Mean (N=10)	All Cases Using Scripted Score (N=10)	Cases With Scripted Entrustment Score of 2 (N=5 cases)
Mean (SD)	2.87 (0.86)	2.70 (0.78)	2.36 (0.72)
Rater variance, %	4.99	9.85	8.50
Case variance, %	54.29	89.07	N/A
Residual error variance, %	40.72	2.38	91.50
Generalizability coefficient	0.78	-	0.81
Number of necessary raters	3	-	15

performance level. Nevertheless, CCC decisions typically rely heavily on faculty members' direct observations of residents caring for patients. CCCs also use multiple observations over time to make decisions about trainees. This may increase the accuracy of the pooled information. Our study was limited to internal medicine and family medicine physicians observing videos of standardized residents in outpatient encounters. As such, our findings may not be generalizable to other specialties, other care contexts, or evaluations with actual patients. It is possible that the video entrustment levels were not accurate. In addition, the video creation focused on content validity and response process as opposed to the other metrics of validity.

Conclusions

Entrustment scale ratings varied significantly by performance level of the resident, with more errors occurring with lower performance of the resident. Residents who perform well are more likely to be accurately evaluated.

References

1. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE guide no. 31. *Med Teach*. 2007;29(9-10):855-871. doi:10.1080/01421590701775453
2. Prentice S, Benson J, Kirkpatrick E, Schuwirth L. Workplace-based assessments in postgraduate medical education: a hermeneutic review. *Med Educ*. 2020;54(11):981-992. doi:10.1111/medu.14221
3. Brennan R. *Generalizability Theory*. Springer-Verlag; 2001.
4. Monteiro S, Sullivan GM, Chan TM. Generalizability theory made simple(r): an introductory primer to G-studies. *J Grad Med Educ*. 2019;11(4):365-370. doi:10.4300/JGME-D-19-00464.1
5. Holmboe ES, Kogan JR. Will any road get you there? Examining warranted and unwarranted variation in medical education. *Acad Med*. 2022;97(8):1128-1136. doi:10.1097/ACM.0000000000004667
6. ten Cate O, Chen HC. The ingredients of a rich entrustment decision. *Med Teach*. 2020;42(12):1413-1420. doi:10.1080/0142159X.2020.1817348
7. Weller JM, Coomber T, Chen Y, Castanelli DJ. Key dimensions of innovations in workplace-based assessment for postgraduate medical education: a scoping review. *Br J Anaesth*. 2021;127(5):689-703. doi:10.1016/j.bja.2021.06.038
8. Dudek N, Gofton W, Rekman J, McDougall A. Faculty and resident perspectives on using entrustment anchors for workplace-based assessment. *J Grad Med Educ*. 2019;11(3):287-294. doi:10.4300/JGME-D-18-01003.1
9. Eltayar AN, Aref SR, Khalifa HM, Hammad AS. Do entrustment scales make a difference in the inter-rater reliability of the workplace-based assessment? *Med Educ Online*. 2022;27(1):2053401. doi:10.1080/10872981.2022.2053401
10. Robinson TJG, Wagner N, Szulewski A, Dudek N, Cheung WJ, Hall AK. Exploring the use of rating scales with entrustment anchors in workplace-based assessment. *Med Educ*. 2021;55(9):1047-1055. doi:10.1111/medu.14573
11. ten Cate O. When I say … entrustability. *Med Educ*. 2020;54(2):103-104. doi:10.1111/medu.14005
12. Kogan JR, Conforti LN, Iobst WF, Holmboe ES. Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Acad Med*. 2014;89(5):721-727. doi:10.1097/ACM.0000000000000221
13. Kogan JR, Dine CJ, Conforti LN, Holmboe ES. Can rater training improve the quality and accuracy of workplace-based assessment narrative comments and entrustment ratings? A randomized controlled trial. *Acad Med*. 2023;98(2):237-247. doi:10.1097/ACM.0000000000004819

14. Calaman S, Hepps JH, Bismilla Z, et al. The creation of standard-setting videos to support faculty observations of learner performance and entrustment decisions. *Acad Med*. 2016;91(2):204-209. doi:10.1097/ACM.0000000000000853
15. Chen HC, van den Broek WES, ten Cate O. The case for use of entrustable professional activities in undergraduate medical education. *Acad Med*. 2015;90(4):431-436. doi:10.1097/ACM.0000000000000586
16. ten Cate O, Schwartz A, Chen HC. Assessing trainees and making entrustment decisions: on the nature and use of entrustment-supervision scales. *Acad Med*. 2020;95(11):1662-1669. doi:10.1097/ACM.0000000000003427
17. Boileau E, St-Onge C, Audétat MC. Is there a way for clinical teachers to assist struggling learners? A synthetic review of the literature. *Adv Med Educ Pract*. 2017;8:89-97. doi:10.2147/AMEP.S123410
18. Colletti LM. Difficulty with negative feedback: face-to-face evaluation of junior medical student clinical performance results in grade inflation. *J Surg Res*. 2000;90(1):82-87. doi:10.1006/jsre.2000.5848
19. Wang XM, Wong KFE, Kwong JYY. The roles of rater goals and ratee performance levels in the distortion of performance ratings. *J Appl Psychol*. 2010;95(3):546-561. doi:10.1037/a0018866
20. Yeates P, Moult A, Cope N, et al. Measuring the effect of examiner variability in a multiple-circuit objective structured clinical examination (OSCE). *Acad Med*. 2021;96(8):1189-1196. doi:10.1097/ACM.0000000000004028
21. Park YS, Hyderi A, Heine N, et al. Validity evidence and scoring guidelines for standardized patient encounters and patient notes from a multisite study of clinical performance examinations in seven medical schools. *Acad Med*. 2017;92(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 56th Annual Research in Medical Education Sessions):12-20. doi:10.1097/ACM.0000000000001918

C. Jessica Dine, MD, MSHP, is Associate Dean, Evaluation and Assessment, and Associate Professor of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA; **Lisa N. Conforti, MPH,** is Senior Research Analyst, Accreditation Council for Graduate Medical Education (ACGME), Chicago, Illinois, USA; **Eric S. Holmboe, MD,** is Chief, Research, Milestones Development and Evaluation, ACGME, Chicago, Illinois, USA; and **Jennifer R. Kogan, MD,** is Associate Dean, Student Success and Professional Development, and Professor of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

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

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

The preliminary findings of this study were presented as an abstract at the Association for Medical Education in Europe conference, August 26-30, 2023, Glasgow, Scotland.

Corresponding author: C. Jessica Dine, MD, MSHP, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA, constance.dine@pennmedicine.upenn.edu, X @JessicaDine

Received April 18, 2023; revisions received August 1, 2023, and October 31, 2023; accepted November 9, 2023.