


Concordance of Narrative Comments with Supervision Ratings Provided During Entrustable Professional Activity Assessments



Andrew S. Parsons, MD, MPH¹ , Kelley Mark, BA², James R. Martindale, PhD³, Megan J. Bray, MD⁴, Ryan P. Smith, MD⁵, Elizabeth Bradley, PhD³, and Maryellen Gusic, MD^{3,6}

¹Departments of Medicine and Public Health Sciences, University of Virginia School of Medicine, Charlottesville, Virginia, USA; ²University of Virginia School of Medicine, Charlottesville, Virginia, USA; ³Office of Medical Education, Center for Medical Education Research and Scholarly Innovation, University of Virginia School of Medicine, Charlottesville, Virginia, USA; ⁴Department of Obstetrics and Gynecology, University of Virginia School of Medicine, Charlottesville, Virginia, USA; ⁵Department of Urology, University of Virginia School of Medicine, Charlottesville, Virginia, USA; ⁶Departments of Biomedical Education and Data Science and Pediatrics, Lewis Katz School of Medicine, Temple University, Philadelphia, Pennsylvania, USA.

BACKGROUND: Use of EPA-based entrustment-supervision ratings to determine a learner's readiness to assume patient care responsibilities is expanding.

OBJECTIVE: In this study, we investigate the correlation between narrative comments and supervision ratings assigned during ad hoc assessments of medical students' performance of EPA tasks.

DESIGN: Data from assessments completed for students enrolled in the clerkship phase over 2 academic years were used to extract a stratified random sample of 100 narrative comments for review by an expert panel.

PARTICIPANTS: A review panel, comprised of faculty with specific expertise related to their roles within the EPA program, provided a "gold standard" supervision rating using the comments provided by the original assessor.

MAIN MEASURES: Interrater reliability (IRR) between members of review panel and correlation coefficients (CC) between expert ratings and supervision ratings from original assessors.

KEY RESULTS: IRR among members of the expert panel ranged from .536 for comments associated with focused history taking to .833 for complete physical exam. CC (Kendall's correlation coefficient W) between panel members' assignment of supervision ratings and the ratings provided by the original assessors for history taking, physical examination, and oral presentation comments were .668, .697, and .735 respectively. The supervision ratings of the expert panel had the highest degree of correlation with ratings provided during assessments done by master assessors, faculty trained to assess students across clinical contexts. Correlation between supervision ratings provided with the narrative comments at the time of observation and supervision ratings assigned by the expert panel differed by clinical discipline, perhaps reflecting the value placed on, and perhaps the comfort level with, assessment of the task in a given specialty.

CONCLUSIONS: To realize the full educational and catalytic effect of EPA assessments, assessors must apply

established performance expectations and provide high-quality narrative comments aligned with the criteria.

KEYWORDS: Entrustable professional activities; Narrative assessment; Competency-based medical education.

J Gen Intern Med 37(9):2200–7

DOI: 10.1007/s11606-022-07509-1

© The Author(s) under exclusive licence to Society of General Internal Medicine 2022

BACKGROUND

Effective assessment frameworks provide credible results and a coherent set of data that support programmatic goals.^{1–3} In the case of Entrustable Professional Activity (EPA) assessments, narrative comments provided by assessors must align with assigned supervision ratings in order for the information to be meaningful to learners and their advisors, mentors, and coaches, and to inform institutional decisions to grant autonomy to learners.^{4–6} With the expanding use of EPA-based entrustment-supervision scales to document a learner's readiness to assume patient care responsibilities has come the call to evaluate implementation of this framework as a measure of clinical performance.^{7–11}

A core component of clinical performance assessment is the incorporation of narrative comments that not only illuminate and support quantitative ratings but also provide data to guide a learner to improve.^{5,6,12} Previous studies have explored assessors' and learners' interpretation of narrative comments and the importance of providing information to learners to promote their ongoing development.^{18–22} While much has been written about narrative assessment,^{13–16} the literature examining alignment of the qualitative and quantitative aspects of EPA assessment is limited.¹⁷ Narrative comments provide context for supervision ratings assigned by assessors, justify the "number" selected, and illuminate strengths and areas for improvement related to the observed

Received September 8, 2021

Accepted March 24, 2022

Published online June 16, 2022

performance.^{17,23} Alignment of narrative comments and quantitative scores integrates evidence, thereby facilitating and informing learner self-assessment.²⁴ The availability of clinical faculty who can coach learners to make meaning of the data and set goals for continued development lays the foundation to promote assessment as learning.^{25–27}

In an EPA-based program of assessment, summative entrustment decisions require synthesis of data from ad hoc EPA assessments. Those charged to make these decisions must have a shared understanding about the approach used to aggregate, interpret, and synthesize information from assessments.^{4,27–32} To facilitate evidence-based, sound decisions, entrustment committees need data that are coherent, that is, narrative comments that are correlated with assessors' numeric ratings.^{33–37} In addition, the purpose, process, and consequence of decisions must be clear to the members of the committee and to assessors, learners, and their faculty advisors.^{38–40} The ability of the committee members to analyze data within and across assessments explicitly contributes to the program's "fitness" in achieving an educational and catalytic effect.^{2,3,41}

Assessors in our EPA program attend professional development training sessions during which they learn how to use direct observation and apply performance expectations (Fig. 1) to assign a supervision rating and provide narrative comments that justify the selected rating.^{42,43} Residents, fellows, attendings (faculty), and master assessors (MA; expert assessors who conduct assessments across clinical contexts) use an adapted, prospective supervision scale during ad hoc assessments. Data from these assessments are immediately available to students and their longitudinal faculty coaches. In regularly scheduled meetings, learners and their coaches co-create individualized learning goals and action plans to achieve those goals.²⁶ The Entrustment Committee (EC) is comprised of the MAs and is facilitated by two members of the leadership team for the EPA program.^{43,44} Members of the EC have first-hand knowledge about workplace assessment, the application of performance expectations for EPA tasks, and the use of observation to collect data predictive of a learner's need for supervision the next time they perform the clinical task.^{21,43} EC members review the results of assessments done by residents and fellows, attending faculty, and their peer MAs throughout the academic year. At the end of the clerkship phase, a collective summative entrustment decision is made about each learner's readiness to enroll in an advanced clinical course in which they are expected to assume patient care responsibilities as an acting intern.⁴⁵

In this study, we explore the concordance of narrative comments with supervision ratings provided for two EPA tasks by three different assessor types in ad hoc assessments in three distinct clinical disciplines. Specifically, we seek to examine if the mean supervision ratings assigned by an expert panel using narrative comments provided at the time of assessment correlate with the ratings assigned by the original assessor.

METHODS

EPA assessment data is collected in a web-enabled tool, iCAN, within our institutional learning management system, VMED. The iCAN tool includes general information about the patient encounter that is entered by the student, a drop down menu to select the supervision rating, and two open text boxes for the assessor to describe what the student did well during the performance of the EPA and what were areas for improvement. Data from assessments completed for students enrolled in the clerkship phase over 2 successive academic years, February 2018–February 2020, were used to extract a stratified random sample of 100 comments. The authors determined that one-hundred comments would provide a representative sample and would also be feasible for the expert panel to review.

EPA assessments include a supervision rating using a 4-point modified entrustment-supervision scale categorizing a student's need for direct or indirect supervision. Assessors select one of the following to indicate their recommendation for the next time the student performs the task (i.e., the student is ready to perform the task): jointly with a supervisor (level 1); with a supervisor in the room, ready to step in as needed (level 2); with a supervisor available to double check all elements of the performed activity (level 3); or with a supervisor available to double check key elements of the performed activity (level 4). All assessments contain narrative comments about observed strengths and areas for development.

To extract the sample of data used in the analysis, data for the two EPA tasks with the largest number of completed assessments at the time of the study were used, i.e., EPA 1 (history taking and physical examination) and EPA 6 (performing an oral presentation based on a clinical encounter). In our program, EPA 1 is assessed through observations of students completing four aspects of this task: EPA 1.1—obtaining a complete history; EPA 1.2—gathering a focused history; EPA 1.3—performing a complete physical; and EPA 1.4—conducting a focused physical. A stratified sampling of assessments was selected to include assessments with the following: each level of supervision rating (level 1 through level 4); completed by each type of assessor in the program (residents/fellows, attendings, and MAs); and from distinct types of patient encounters experienced during the clerkship phase (inpatient and outpatient settings, a procedural-based specialty, and with adult and pediatric patients). More specifically, assessments from the Internal Medicine (inpatient setting), Pediatrics (inpatient and outpatient setting), and Surgery clerkships were used. Narrative comments were separated from the supervision rating provided by the original assessor by one of the members of the research team (JM). Narrative comments were further de-identified before review by the expert panel through the removal of information that would specify a student, assessor, or the clerkship during which the assessment was completed.

The expert panel was comprised of a MA, a faculty coach, and a member of the EPA leadership team. Members of the expert panel are all clinical faculty members who complete EPA assessments in their role as a clinical supervisor or MA. Despite their familiarity with the process of completing assessments, the panel met as a group to discuss how they would use the established performance expectations for each EPA task (Fig. 1) during their review of narrative comments to frame their decisions about the level of supervision suggested by the qualitative information provided. The de-identified narrative comments were provided electronically to the expert panel after this conversation. Each member of the expert panel independently reviewed the narrative information related to strengths and areas for improvement in each comment and assigned a corresponding supervision rating. Interrater reliability (IRR) was measured using Kendall's concordance coefficient *W*. Correlation between the mean supervision rating assigned by the expert panel and the rating provided with the narrative comments by the original assessor was analyzed using a Kendall *W* test (Kendall's concordance coefficient *W*). The Kendall *W* test is a measure of concordance and is commonly used to assess agreement among a group of raters. As a non-parametric procedure, it is particularly well-suited for outcome measures that are ordinal in nature, as is the case with this dataset.⁴⁶ A correlation coefficient of 1 represents perfect agreement. A supervision rating assigned by the original assessor and by a member of the expert panel represents a discrete decision related to a specific clinical encounter.⁴

Means were calculated for ratings of all comments related to specific EPA task.

The UVA Institutional Review Board reviewed this project and determined that it met criteria for exempt review (ref no. 3791).

RESULTS

One-hundred narrative comments were extracted from assessments completed for 305 clerkship phase students (149 in 2018–2019; 156 from 2019 to 2020). These 100 narrative comments represented 100 unique students and 80 unique assessors. The sample included 32 narrative comments for assessments of history taking (EPA 1.1 + EPA 1.2), 32 comments related to physical examination skills (EPA 1.3 + EPA 1.4), and 36 comments for oral presentation (EPA 6); 37 comments were originally provided by residents/fellows, 27 comments by attendings, and 36 comments by an MA; 37 comments were from the internal medicine clerkship, 36 from the pediatrics clerkship, and 27 from the surgery clerkship.

IRR among supervision ratings assigned by members of the expert panel in their independent review ranged from .536 for comments associated with focused history taking to .833 for complete physical exam. Kendall *W* (KW) test correlation coefficients (CC) for panel members' assignment of supervision ratings for history taking (complete + focused), physical

Component:	Observed Performance		
	Direct Supervision		Indirect Supervision
1. Presentation of Hx (includes CC, HPI, PMHx, PSHx, ALL, MEDS, FHx, SHx, ROS):	Student does not organize information systematically according to the usual sequence. Student does not accurately* present the history, or documents an incomplete history or one with extraneous facts.	Student accurately* present relevant information from all components of a patient's history in an organized narrative.	Student contextualizes the presenting problem by incorporating critical elements from all components of the history and from alternate sources of information in a cogent oral narrative.
2. Presentation of physical exam, laboratory and diagnostic studies:	Student does not organize information systematically according to the usual sequence. Student does not accurately* present exam findings and laboratory and diagnostic study results, or presents incomplete or extraneous information.	Student accurately* presents all pertinent information from the physical examination including a statement about the patient's general appearance and vital signs, and from laboratory and diagnostic studies in an organized manner.	Student incorporates relevant information from the physical examination and from laboratory and diagnostic studies related to all of the patient's health conditions that may be contributing to the patient's presentation/current health status.
3. Presentation of assessment:	Student does not articulate a summary statement, offers only one diagnostic possibility, or includes several unlikely diagnoses without prioritizing or justifying the working diagnosis.	Student articulates a summary statement and provides a prioritized differential with several relevant alternate diagnoses. Student states a most likely diagnosis and justifies its likelihood using key features of the patient's presentation and health conditions, as well as relevant epidemiology and findings from laboratory and diagnostic studies. Student also details the status of all other health conditions.	Student incorporates critical elements of the patient's history, physical exam, laboratory, and diagnostic studies into the summary statement. Student states reasoning behind the inclusion of all alternate diagnoses by comparing and contrasting the discriminating features of the alternate diagnoses with those of the working diagnosis, at the same time acknowledging any incongruous elements in the patient's presentation.
4. Presentation of plan:	Student states a treatment plan directed only to the most likely diagnosis.	Student states a plan that includes appropriate laboratory or diagnostic testing to discriminate the diagnoses in the differential, or to affirm the working diagnosis. Student states reasoning to support proposed management and treatment plan and addresses all other health conditions with particular attention to those that are impacting the patient's current health status.	Student includes patient/family preferences, literature/practice guidelines, cost-effectiveness principles in the management and treatment plan.
5. Speaking style:	Student does not speak clearly or audibly or does not make appropriate eye contact.	Student speaks clearly and audibly making appropriate eye contact.	

Figure 1 Oral presentation (EPA 6) performance expectations to apply in assigning supervision ratings

Table 1 Interrater Reliability Among Mean Supervision Ratings Assigned by Each Member of the Expert Panel

EPA task	Expert panel member 1 (mean/SD)	Expert panel member 2 (mean/SD)	Expert panel member 3 (mean/SD)	KWCC
1.1 Complete history	2.50/.73	2.44/.81	2.25/.58	.793
1.2 Focused history	2.63/.62	2.06/.57	2.13/.50	.536
1.1 + 1.2 History taking	2.56/.67	2.25/.72	2.19/.54	.668
1.3 Complete physical exam	2.64/.74	2.57/.94	2.29/.61	.833
1.4 Focused physical exam	2.22/.88	2.11/1.08	1.78/.65	.685
1.3 + 1.4 Physical exam	2.41/.84	2.31/1.03	2.00/.67	.697
6 Oral presentation	2.36/.87	2.22/.83	2.03/.65	.735

examination (complete + focused), and oral presentation comments were .668, .697, and .735 respectively (Table 1).

CC between the mean supervision rating of the expert panel and the mean rating for the task provided at the time of the assessment ranged from .327 for history taking to .697 for physical examination and .735 for oral presentation. The mean supervision rating assigned for each task includes assessments completed by all assessor types (residents/fellows, attendings, and MAs) and in all of the clerkships included in the study (internal medicine, pediatrics, and surgery). The mean supervision rating of the original assessors and the mean supervision rating determined through the expert panel member’s review of the narrative comments are included in Table 2. Representative narrative comments with a high and low level of correlation between the supervision ratings provided by the expert panel and by the original assessor during assessments of oral presentations (EPA 6) are illustrated in Table 3.

Correlation between the mean supervision ratings assigned by the expert panel and the supervision rating provided with the narrative comments at the time of the observation varied by assessor type (Table 4). For history taking, the CC was .525, .540, and .941 respectively for supervision ratings originally determined by residents/fellows, attendings, and MAs respectively; for physical examination, the CC ranged from .403 to .790 for ratings from residents/fellows to ratings from MAs and for oral presentation, the CC spanned from .309 for residents/fellow ratings to .854 for MA ratings.

Table 5 contains CC for data from assessments completed by all assessor types in a variety of clinical settings (clerkships). CC between the mean supervision ratings provided at the time of the observation and the mean expert panel rating

for comments from assessments from the internal medicine clerkship ranged from .596 to .942 for history taking and oral presentation; for assessments on the surgery clerkship, the CC ranged from .301 to .873 for history taking and oral presentation respectively. CC for mean ratings of comments provided on the pediatrics clerkship were .738 for history taking, .862 for physical examination, and .663 for oral presentation.

DISCUSSION

In this study, we explored the alignment of narrative comments with the supervision ratings provided by assessors during EPA assessments. The expert panel’s assignment of supervision ratings served as a “gold standard” for comparison to supervision ratings assigned by original assessors. A higher degree of correlation between the “gold standard” supervision rating and the original rating suggests that the narrative comments provided by the original assessor more closely align, are concordant with, support, and justify supervision ratings. We found, however, that supervision ratings given by the expert panel had variable levels of concordance with the ratings given by original assessors at the time of assessment.

Narrative comments contain critical information about learners’ performance not fully captured by a quantitative entrustment-supervision scale score.^{5,6,47} For narrative comments provided in EPA assessment to be useful to learners⁴⁸ and also to summative decision-making committees, all stakeholders who provide the comments must be clear about the importance of meaningful, high-quality, performance-based narrative that substantiates quantitative ratings.¹⁷ Our findings

Table 2 Correlation Between Mean Supervision Rating Provided by the Original Assessor and the Mean Rating Assigned by the Expert Panel

EPA task	Supervision rating original assessors (mean/SD)	Supervision rating expert panel (mean/SD)	KWCC
1.1 Complete history	2.94/.85	2.40/.72	.389
1.2 Focused history	2.81/.91	2.27/.45	.388
1.1 + 1.2 History taking	2.88/.87	2.33/.52	.327
1.3 Complete physical exam	2.00/.68	2.50/.76	.737
1.4 Focused physical exam	2.78/1.00	2.04/.91	.788
1.3 + 1.4 Physical exam	2.44/.95	2.24/.74	.667
6 Oral presentation	2.97/.81	2.20/.68	.775

KWCC Kendall W (KW) correlation coefficient (a correlation coefficient of 1 represents perfect agreement), SD standard deviation

Table 3 Representative Narrative Comments from Assessments of Oral Presentations (EPA 6) Representing High and Low Levels of Correlation Between the Supervision Ratings of the Expert Panel and the Original Assessor

	Narrative comments	Example OA supervision rating	Example EP supervision rating
High correlation	Very well organized and concise, yet thorough presentation. Structured in the SOAP format and accurate throughout, connected the patient’s history of present illness to the objective findings. Provided only the relevant findings needed for clinical reasoning. Provided an accurate and concise representation of the problem including relevant risk factors, tempo of illness, and key signs and symptoms. The assessment led nicely into the problem list—verbalized reasoning for each diagnosis on the differential and exclusion of less likely. Captured the severity and complexity of each problem. Provided a diagnostic and therapeutic plan for each problem that incorporated patient’s goals and value—incorporating cost-effectiveness is the next step in development.	4	4
Low correlation	Clearly presented. Made eye contact. Included all components of the history including CC, HPI, PMHx, PSHx, ALL, MEDS, FHx, SHx, and ROS and described these accurately. Provided an assessment including the correct diagnosis followed by a plan.	3	2

OA original assessor, EP expert panel

support the call not only to evaluate the fidelity of implementation but also to measure outcomes that provide meaningful information to learners, their coaches, and institutional decision-makers.^{7,9,11,47}

Supervision ratings of the expert panel had the highest degree of correlation with ratings provided by MAs. As noted, MAs are experienced clinicians, selected and trained to perform assessments across various clinical settings. In our program, all assessors (residents/fellows/attendings) are required to attend an EPA training session. All sessions are interactive

and structured to promote skill building and hands-on practice in applying performance expectations, translating observations into decisions about the level of supervision a student needs the next time they perform the task, and providing narrative comments to justify the level of supervision selected.⁴³ MAs are “frequent observers” with designated effort for this role. They are not simultaneously supervising students nor providing clinical care for the patient at the time of the assessment and participate in additional professional

Table 4 Correlation Between Mean Supervision Rating Provided by Each Original Assessor Type and the Mean Rating Assigned by the Expert Panel

EPA task	Supervision rating OA (mean/SD)	Deviation supervision rating expert panel (EP) (mean/SD)	KWCC	Supervision rating OA (mean/SD)	Supervision rating EP (mean/SD)	KWCC	Supervision rating OA (mean/SD)	Supervision rating EP (mean/SD)	KWCC
	Resident/Fellow			Attending			MA		
1.1 Complete history	3.00/.93	2.40/.68	.212	3.00/1.00	2.40/.26	.375	2.67/.58	2.40/.69	.929
1.2 Focused history	2.56/1.01	2.27/.46	.690	3.40/.55	2.27/.37	.823	2.50/.71	2.27/.47	1.000
1.1 + 1.2 History taking	2.76/.97	2.33/.61	.525	3.20	2.33	.540	2.60	2.33	.941
1.3 Complete physical exam	2.00/.00	2.50/.72	.500	1.75/.96	2.50/.63	.528	2.17/.75	2.50/.50	.883
1.4 Focused physical exam	3.43/.79	2.04.85	.590	3.00/1.40	2.04/.71	1.000	2.22/.83	2.04/.60	.899
1.3 + 1.4 Physical exam	2.91/.94	2.24/.83	.403	2.17/1.17	2.24/.58	.656	2.20/.78	2.24/.66	.790
6 Oral presentation	3.11/.93	2.20/.74	.948	3.09/.70	2.20/.55	.442	2.81/.83	2.20/.72	.854

KWCC Kendall W (KW) correlation coefficient (a correlation coefficient of 1 represents perfect agreement), OA original assessor, MA master assessor, SD standard deviation

Table 5 Correlation Between the Mean Supervision Ratings Provided by Original Assessors in Each Clinical Discipline/Setting (Clerkship) and the Mean Supervision Rating of the Expert Panel

Clerkship	Internal medicine*			Pediatrics**			Surgery***		
	Supervision rating OA (mean/SD)	Supervision rating EP (mean/SD)	KWCC	Supervision rating OA (mean/SD)	Supervision rating EP (mean/SD)	KWCC	Supervision rating OA (mean/SD)	Supervision rating EP (mean/SD)	KWCC
1.1 Complete history	2.25/.50	2.67.82	.233	3.14/.90	2.33/.54	.683	3.20	2.27	.149
1.2 Focused history	2.75/1.04	2.29/.49	.745	2.33/.58	2.22/.38	1.000	3.20/.84	2.27/.64	.324
1.1 + 1.2 History taking	2.58/.90	2.42/.61	.596	2.90/.88	2.30/.48	.738	3.20/.79	2.27/.49	.301
1.3 Complete physical exam	2.20/.63	2.63/.58	.605	1.67/.58	2.11/1.17	.929	1.00/.00	2.33/.00	****
1.4 Focused physical exam	2.67/.58	2.44/.51	.929	2.14/.90	1.81/.74	.937	3.38/.92	2.08/.79	.747
1.3 + 1.4 Physical exam	2.31/.63	2.59/.55	.626	2.00/.82	1.90/.83	.862	3.11/1.17	2.11/.75	.638
6 Oral presentation	2.83/.72	2.36/.78	.942	2.94/.85	2.08/.59	.663	3.25/.89	2.21/.71	.873

KWCC Kendall W (KW) correlation coefficient (a correlation coefficient of 1 represents perfect agreement), OA original assessor, EP expert panel, SD standard deviation

*All assessments completed in an inpatient setting; Adult patients

**Assessments completed in inpatient and outpatient settings; Pediatric patients

***Procedural-based specialty

****The small number of assessments for this EPA task in this clinical discipline did not allow analysis of correlation

development to enable them to complete assessments outside of their clinical specialty.

While authors⁴⁹ have noted enhanced generalizability of ad hoc entrustment decisions when provided by clinical supervisors who assess students frequently, decisions in the workplace require an assessor to weigh the risk of granting autonomy to a learner.^{4,50,51} The relative lack of concordance between narrative comments and supervision ratings provided by residents/fellows and attendings may be explained by the challenges inherent in serving concurrently as a teacher, assessor, and clinical supervisor.^{28,52,53} The quality and focus on assessment can vary when any one role is emphasized. Assessors may also struggle with assigning a supervision rating indicating what level of supervision a student will need in future clinical encounters.⁵² Prospective decisions about a learner’s adaptive competence based on a discrete observation of clinical performance require a different mindset than traditional end of clerkship/rotation evaluation.^{4,52,455}

MAs constitute the Entrustment Committee (EC) and have participated in additional training to facilitate group decision-making. The Committee meets regularly to review and analyze data from ad hoc assessments further developing their expertise and fortifying their shared mental model about the criteria for assessment, specifically how the performance expectations outline behaviors that can be translated to the assignment of a supervision rating. To make summative entrustment decisions, the members of the committee integrate and synthesize quantitative (supervision ratings) and qualitative data (narrative comments) from assessments completed across

clinical contexts to predict students’ readiness to meet expectations for future performance.^{4,29–31,33}

Our findings suggest that despite efforts to establish a shared understanding and application of established performance expectations, clinical supervisors may define what constitutes a focused history differently based on their clinical discipline and likely the context of the encounter. In contrast to focused history taking, approaches to hypothesis-driven evidence-based physical exam have been well described and with the availability of published resources, assessors are perhaps less likely to rely on personal opinion in judging a learner’s performance. Correlation between supervision ratings provided with the narrative comments at the time of observation and supervision ratings assigned by the expert panel differed by clerkship and may reflect the value placed on the skill, and perhaps the corresponding comfort level with assessment of the task in a given specialty.⁸ Supervisors on procedural-based specialties spend less time with learners in settings in which a history and physical examination would be performed, leading to a reliance on the use of simulation-based assessment for these skills.⁵⁶ In contrast, supervision ratings and corresponding narrative comments for EPA 6 (oral presentation) provided to students in assessments on the internal medicine clerkship were highly correlated with supervision ratings assigned by the expert panel. This likely reflects comfort with this traditional approach used to assess learners on the internal medicine clerkship.⁵⁷ These findings suggest the need to consider the existing teaching and assessment practices of various

clinical disciplines when defining opportunities to incorporate EPA assessments in each setting.^{51,58,59}

Stakes, whether they be low or high, influence all stakeholders.^{50,27,51,53} And dual purposing of data from assessment for both formative feedback and summative decisions may raise concerns for ad hoc assessors.^{5,17} In our program, the data from ad hoc assessments does not contribute to the student's evaluation/grade on the clinical clerkship and the results are visible only to the student, their faculty coach, and their student affairs dean. The student and faculty coach use the data to co-create individualized learning plans to promote continued clinical development.²⁶ Supervisors may be particularly concerned about disadvantaging learners through assessment, highlighting the importance of training for assessors to ensure they understand the goal of the program and how the data from assessment are used.^{28,43,51,54} A dedicated group of "external" assessors, who do not contribute to a student's formal end of clerkship evaluation, does not experience this tension. For narrative comments provided in EPA assessment to be useful to learners⁵³ and also to summative decision-making committees, all stakeholders who provide the comments must be clear about the importance of meaningful, high-quality, performance-based narrative that substantiates quantitative ratings and expands the information provided through the supervision rating.¹⁷

This study has limitations. First, narrative comments sampled for analysis represent a random sampling of the total assessments completed during the study period. The assessments were done during observation of a subset of students by a subset of the total assessors in the program and, so, may not be representative. Second, this study did not examine the accuracy of data (supervision ratings or narrative comments). Although our web-enabled assessment tool allows assessors to use voice dictation to capture verbal feedback, it is not known if the narrative comments were consistent with verbal feedback given to learners at the time of the assessment. Assessment data must be submitted within a specified period after the time of the observation but if not done immediately may be subject to limited recollection and/or recall bias.

CONCLUSIONS

EPA assessments communicate information about a learner through both entrustment-supervision ratings and narrative comments about observed performance. Concordance between both components is critical in making the data meaningful to learners and to those who help them use this information for their continued development.^{2,3,53,60} Committees charged with analyzing and integrating data from EPA assessments to make high stakes decisions must be able to use the information to support summative entrustment.^{6,29,30,36,61} Our findings underscore the need for high-quality narrative comments aligned with performance criteria so that the educational and catalytic effect of an EPA-based program of assessment can be fully realized.^{2,3,43,62,63}

Corresponding Author: Andrew S. Parsons, MD, MPH; Departments of Medicine and Public Health Sciences, University of Virginia School of Medicine, Charlottesville, Virginia, USA (e-mail: Asp5c@virginia.edu).

REFERENCES

1. **Norcini J.** The power of feedback. *Med Educ.* 2010;44(1):16-17.
2. **Dijkstra J, Galbraith R, Hodges BD, et al.** Expert validation of fit-for-purpose guidelines for designing programmes of assessment. *BMC Med Educ.* 2012;12:20.
3. **Norcini J, Anderson MB, Bollela V, Burch V, Costa MJ, Duvivier R, Hays R, Palacios Mackay MF, Roberts T, Swanson D.** 2018 Consensus framework for good assessment. *Med Teach.* 2018 Nov;40(11):1102-1109.
4. **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.
5. **Govaerts MJB, Vleuten CPM van der, Holmboe ES.** Managing tensions in assessment: moving beyond either-or thinking. *Med Educ.* 2019;53(1):64-75.
6. **Carraccio C, Martini A, Van Melle E, Schumacher DJ.** Identifying core components of EPA implementation: a path to knowing if a complex intervention is being implemented as intended. *Acad Med.* 2021 Sep 1;96(9):1332-1336.
7. **Ten Cate O, Balmer DF, Caretta-Weyer H, Hatala R, Hennis MP, West DC.** Entrustable professional activities and entrustment decision making: a development and research agenda for the next decade. *Acad Med.* 2021 Jul 1;96(7S):S96-S104.
8. **Pinilla, S., Lenouvel, E., Cantisani, A. et al.** Working with entrustable professional activities in clinical education in undergraduate medical education: a scoping review. *BMC Med Educ* 21, 172 (2021).
9. **Meyer EG, Chen HC, Uijtdehaage S, Durning SJ, Maggio LA.** Scoping review of entrustable professional activities in undergraduate medical education. *Acad Med.* 2019 Jul;94(7):1040-1049.
10. **Shorey S, Lau TC, Lau ST, Ang E.** Entrustable Professional Activities in Health Care Education: a Scoping Review. *Med Educ.* 2019;53(8):766-777. doi:<https://doi.org/10.1111/medu.13879>
11. **Van Melle E, Hall AK, Schumacher DJ, et al.** Capturing Outcomes of Competency-Based Medical Education: The Call and the Challenge. *Med Teach.* Published online June 12, 2021;1-7. doi:<https://doi.org/10.1080/0142159X.2021.1925640>
12. **Holmboe ES, Yamazaki K, Hamstra SJ.** The evolution of assessment: thinking longitudinally and developmentally. *Acad Med.* 2020 Nov;95(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 59th Annual Research in Medical Education Presentations):S7-S9.
13. **Tekian A, Park YS, Tilton S, Prunty PF, Abasolo E, Zar F, Cook DA.** Competencies and feedback on internal medicine residents' end-of-rotation assessments over time: qualitative and quantitative analyses. *Acad Med.* 2019 Dec;94(12):1961-1969.
14. **Tekian A, Borhani M, Tilton S, Abasolo E, Park YS.** What do quantitative ratings and qualitative comments tell us about general surgery residents' progress toward independent practice? Evidence from a 5-year longitudinal cohort. *Am J Surg.* 2019;217(2):288-295.
15. **Sebok-Syer SS, Klinger DA, Sherbino J, Chan TM.** Mixed messages or miscommunication? Investigating the relationship between assessors' workplace-based assessment scores and written comments. *Acad Med.* 2017 Dec;92(12):1774-1779.
16. **Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA.** Using In-Training Evaluation Report (ITER) qualitative comments to assess medical students and residents: a systematic review. *Acad Med.* 2017 Jun;92(6):868-879.
17. **Ginsburg S, Watling CJ, Schumacher DJ, Gingerich A, Hatala R.** Numbers encapsulate, words elaborate: toward the best use of comments for assessment and feedback on entrustment ratings. *Acad Med.* 2021 Jul 1;96(7S):S81-S86.
18. **Martin L, Sibbald M, Brandt Vegas D, Russell D, Govaerts M.** The impact of entrustment assessments on feedback and learning: trainee perspectives. *Med Educ.* 2020 Apr;54(4):328-336.
19. **Ginsburg S, Regehr G, Lingard L, Eva KW.** Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ.* 2015 Mar;49(3):296-306.
20. **Duijn CCMA, Welink LS, Mandoki M, Ten Cate OTJ, Kremer WDJ, Bok HGJ.** Am I ready for it? Students' perceptions of meaningful

- feedback on entrustable professional activities. *Perspect Med Educ*. 2017 Aug;6(4):256-264.
21. **Bradley EB, Waselewski EA, Gusic ME.** How do clerkship students use EPA DATA? Illuminating students' perspectives as partners in programs of assessment. *Med Sci Educ*. 2021;31(4):1419-1428.
 22. **Ginsburg S, van der Vleuten CP, Eva KW, Lingard L.** Cracking the code: residents' interpretations of written assessment comments. *Med Educ*. 2017;51(4):401-410.
 23. **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.
 24. **Sargeant J, Armson H, Chesluk B, Dornan T, Eva K, Holmboe E, Lockyer J, Loney E, Mann K, van der Vleuten C.** The processes and dimensions of informed self-assessment: a conceptual model. *Acad Med*. 2010 Jul;85(7):1212-20.
 25. **Torre DM, Schuwirth LWT, Van der Vleuten CPM.** Theoretical considerations on programmatic assessment. *Med Teach*. 2020;42(2):213-220.
 26. **Parsons AS, Kon RH, Plews-Ogan M, Gusic ME.** You can have both: coaching to promote clinical competency and professional identity formation. *Perspect Med Educ*. 2021 Jan;10(1):57-63.
 27. **Ross S, Hauer KE, Wycliffe-Jones K, et al.** ICBME Collaborators. Key considerations in planning and designing programmatic assessment in competency-based medical education. *Med Teach*. 2021 Jul;43(7):758-764.
 28. **Schut S, Maggio LA, Heeneman S, van Tartwijk J, van der Vleuten C, Driessen E.** Where the rubber meets the road - an integrative review of programmatic assessment in health care professions education. *Perspect Med Educ*. 2021 Jan;10(1):6-13.
 29. **Edgar L, Jones MD, Harsy B, Passiment M, Hauer KE.** Better decision-making: shared mental models and the Clinical Competency Committee. *J Grad Med Educ*. 2021;13(2 Suppl):51-58.
 30. **Hauer KE, Edgar L, Hogan SO, Kinnear B, Warm E.** The science of effective group process: lessons for Clinical Competency Committees. *J Grad Med Educ*. 2021;13(2s):59-64.
 31. **Carraccio C, Englander R, Holmboe ES, Kogan JR.** Driving care quality: aligning trainee assessment and supervision through practical application of entrustable professional activities, competencies, and milestones. *Acad Med*. 2016 Feb;91(2):199-203.
 32. **ten Cate O, Carraccio C, Damodaran A, et al.** Entrustment decision making: extending Miller's pyramid. *Acad Med*. 2021;96(2):199-204.
 33. **Pack R, Lingard L, Watling CJ, Chahine S, Cristancho SM.** Some assembly required: tracing the interpretative work of Clinical Competency Committees. *Med Educ*. 2019 Jul;53(7):723-734.
 34. **Pack R, Lingard L, Watling C, Cristancho S.** Beyond summative decision making: illuminating the broader roles of competence committees. *Med Educ*. 2020;54(6):517-527.
 35. **Kinnear B, Kelleher M, May B, Sall D, Schauer DP, Schumacher DJ, Warm EJ.** Constructing a validity map for a workplace-based assessment system: cross-walking Messick and Kane. *Acad Med*. 2021 Jul 1;96(7S):S64-S69.
 36. **Cook DA, Brydges R, Ginsburg S, Hatala R.** A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49(6):560-575.
 37. **Rich JV, Young SF, Donnelly C, et al.** Competency-based education calls for programmatic assessment: but what does this look like in practice? *J Eval Clin Pract*. 2020;26(4):1087-1095.
 38. **Misra S, Iobst WF, Hauer KE, Holmboe ES.** The importance of competency-based programmatic assessment in graduate medical education. *J Grad Med Educ*. 2021;13(2 Suppl):113-119.
 39. **Lupi CS, Ownby AR, Jokela JA, et al.** Association of American Medical Colleges Core Entrustable Professional Activities for Entering Residency Faculty Development Concept Group. Faculty development revisited: a systems-based view of stakeholder development to meet the demands of Entrustable Professional Activity implementation. *Acad Med*. 2018 Oct;93(10):1472-1479.
 40. **Hauer KE, O'Sullivan PS, Fitzhenry K, Boscardin C.** Translating theory into practice: implementing a program of assessment. *Acad Med*. 2018 Mar;93(3):444-450.
 41. **van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al.** A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205-214.
 42. **Chen HC, van den Broek WE, ten Cate O.** The case for use of entrustable professional activities in undergraduate medical education. *Acad Med*. 2015 Apr;90(4):431-6.
 43. **Bray MJ, Bradley EB, Martindale JR, Gusic ME.** Implementing systematic faculty development to support an EPA-based program of assessment: strategies, outcomes, and lessons learned. *Teach Learn Med*. 2021 Aug-Sep;33(4):434-444.
 44. **Burm S, Sebok-Syer SS, Lingard L, VanHooren T, Chahine S, Goldszmidt M, Watling CJ.** "You want me to assess what?": faculty perceptions of assessing residents from outside their specialty. *Acad Med*. 2019 Oct;94(10):1478-1482.
 45. **Keeley MG, Gusic ME, Morgan HK, Aagaard EM, Santen SA.** Moving toward summative competency assessment to individualize the postclerkship phase. *Acad Med*. 2019;94(12):1858-1864.
 46. **Kendall, M. G., & Gibbons, J. D.** (1990). Rank correlation methods. New York, NY : Oxford University Press.
 47. **Cook DA, Kuper A, Hatala R, Ginsburg S.** When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med*. 2016 Oct;91(10):1359-1369.
 48. **Schut S, Heeneman S, Bierer B, Driessen E, Tartwijk J van, Vleuten C van der.** Between trust and control: teachers' assessment conceptualisations within programmatic assessment. *Med Educ*. 2020;54(6):528-537.
 49. **Ryan MS, Khamishon R, Richards A, Perera R, Garber A, Santen S.** A question of scale? Generalizability of the Ottawa and Chen scales to render entrustment decisions for the core EPAs in the workplace. *Acad Med*. Published online December 21, 2021.
 50. **Hall AK, Schumacher DJ, Thoma B, et al.** Outcomes of competency-based medical education: a taxonomy for shared language. *Med Teach*. 2021 Jul;43(7):788-793.
 51. **Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ.** Entrustability scales: outlining their usefulness for competency-based clinical assessment. *Acad Med*. 2016 Feb;91(2):186-90.
 52. **Postmes L, Tammer F, Posthumus I, Wijnen-Meijer M, van der Schaaf M, ten Cate O.** EPA-based assessment: clinical teachers' challenges when transitioning to a prospective entrustment-supervision scale. *Med Teach*. 2021;43(4):404-410.
 53. **Schut S, Driessen E, Tartwijk J van, Vleuten C van der, Heeneman S.** Stakes in the eye of the beholder: an international study of learners' perceptions within programmatic assessment. *Med Educ*. 2018;52(6):654-663.
 54. **Gomez-Garibello C, Young M.** Emotions and assessment: considerations for rater-based judgements of entrustment. *Med Educ*. 2018;52(3):254-262.
 55. **Cate O ten, Hart D, Ankel F, et al.** Entrustment decision making in clinical training. *Acad Med*. 2016;91(2):191-198.
 56. **Butler KL, Hirsh DA, Petrusa ER, et al.** Surgery clerkship evaluations are insufficient for clinical skills appraisal: the value of a medical student surgical objective structured clinical examination. *J Surg Educ*. 2017;74(2):286-294.
 57. **Fazio SB, Ledford CH, Aronowitz PB, et al.** Competency-based medical education in the internal medicine clerkship: a report from the alliance for Academic Internal Medicine Undergraduate Medical Education Task Force. *Acad Med*. 2018 Mar;93(3):421-427.
 58. **Amiel JM, Andriole DA, Biskobing DM, et al.** Revisiting the core entrustable professional activities for entering residency. *Acad Med*. 2021;96(7S):S14.
 59. **Hauer KE, Boscardin C, Fulton TB, Lucey C, Oza S, Teherani A.** Using a curricular vision to define entrustable professional activities for medical student assessment. *J Gen Intern Med*. 2015;30(9):1344-1348.
 60. **Caro Monroig AM, Chen HC, Carraccio C, Richards BF, Ten Cate O, Balmer DF, and the EPAC Study Group.** Medical students' perspectives on entrustment decision-making in an EPA assessment framework: a secondary data analysis. *Acad Med*. 2020 Nov 24.
 61. **Chahine S, Cristancho S, Padgett J, Lingard L.** How do small groups make decisions? *Perspect Med Educ*. 2017;6(3):192-198.
 62. **Moreau KA.** Exploring the connections between programmatic assessment and program evaluation within competency-based medical education programs. *Med Teach*. 2021;43(3):250-252.
 63. **Carraccio C, Martini A, Van Melle E, Schumacher DJ.** Identifying core components of EPA implementation: a path to knowing if a complex intervention is being implemented as intended. *Acad Med*. 2021 Sep 1;96(9):1332-1336.