ACADEMIC MEDICINE
Journal of the Association of American Medical Colleges

Author: Tavares Walter PhD; Sadowski Alexander; Eva Kevin W. PhD

Title: Asking for Less and Getting More: The Impact of Broadening a Rater's Focus in Formative Assessment

DOI: 10.1097/ACM.0000000000002294

Academic Medicine

DOI: 10.1097/ACM.0000000000002294

Asking for Less and Getting More: The Impact of Broadening a Rater's Focus in Formative Assessment

Walter Tavares, PhD, Alexander Sadowski, and Kevin W. Eva, PhD

W. Tavares is a scientist and assistant professor, The Wilson Centre, Department of Medicine and Post-MD Education, University Health Network / University of Toronto Faculty of Medicine, Toronto, Ontario, Canada; and a clinician scientist, Department of Community and Health Services, Paramedic and Senior Services, Regional Municipality of York, Newmarket, Ontario, Canada.

A. Sadowski is a research associate, The Wilson Centre, University Health Network / University of Toronto Faculty of Medicine, Toronto, Ontario, Canada.

K.W. Eva is senior scientist, Centre for Health Education Scholarship, and professor, Department of Medicine, University of British Columbia Faculty of Medicine, Vancouver, British Colombia, Canada.

Correspondence should be sent to: Dr. Tavares, The Wilson Centre, 200 Elizabeth Street, 1ES-565, Toronto, Ontario, Canada M5G 2C4, Phone: 416-340-3646 / 416-340-3079, Fax: 416-340-3792, email: walter.tavares@utoronto.ca; Twitter: @WalterTava.

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A559.

Acknowledgments: The authors wish to thank Centennial College for supporting this study, as well as Karen McIntyre and Fontana Lim for their assistance in completing the data collection.

Funding/Support: This study was generously supported by the Royal College of Physicians and Surgeons Medical Education Research Grant.

Other disclosures: None reported.

Ethical approval: Ethical approval for this study was provided by the Centennial College Research Ethics Board (REB#191).

Previous presentations: Preliminary results of this study were presented at the Wilson Centre Research Day, University of Toronto, Toronto, Canada; November 4, 2016.

Abstract

Purpose

There may be unintended consequences of broadening the competencies across which health professions trainees are assessed. This study was conducted to determine if such broadening influences the formative guidance assessors provide to trainees and to test whether sequential collection of competency-specific assessment can overcome setbacks of simultaneous collection.

Method

A randomized between-subjects experimental design, conducted in Toronto and Halifax, Canada, in 2016–17 with paramedic educators experienced in observing/rating, in which observers' focus was manipulated. In the simultaneous condition, participants rated four unscripted (i.e., spontaneously generated) clinical performances using a six-dimension global rating scale and provided feedback. In three sequential conditions, participants were asked to rate the same performances and provide feedback but for only two of the six dimensions. Participants from these conditions were randomly merged to create a "full score" and set of feedback statements for each candidate.

Results

Eighty-seven raters completed the study; 23 in the simultaneous condition and 21 or 22 for each pair of dimensions in the sequential conditions. After randomly merging participants, there were 21 "full scores" in the sequential condition. Compared to the sequential condition, participants in simultaneous condition demonstrated reductions in the amount of unique feedback provided, increased likelihood of ignoring some dimensions of performance, lessened variety of feedback, and reduced reliability.

Conclusions

Sequential or distributed assessment strategies in which raters are asked to focus on less may provide more effective assessment by overcoming the unintended consequences of asking raters to spread their attention thinly over many dimensions of competence.

Establishing assessments that promote trustworthy decisions regarding clinical competence but also support learner development is a priority in health professions education.[1] Fulfilling either goal requires the judgment of observers. This in turn requires awareness of the difficulty inherent in attending to, processing, and translating observed performance into ratings that reference a standard and feedback that guides learners.[2,3] This is challenged by modern competence frameworks that lead assessment designers to require raters to simultaneously focus on multiple dimensions of performance.[4–6] Each dimension of competence is important, but cognitive capacity is limited and complexity can negatively affect rating quality, causing raters to apply mental shortcuts that result in suboptimal observation.[7] Determining how to best manage this tension requires an understanding of the factors that influence raters' cognition.

DeNisi[8] has studied this issue, as have we,[9] arguing that broadening a rater's focus to the point that cognitive resources are exceeded harms appraisal/assessment processes.[8,9] We have previously suggested that appraisals only have utility to the extent that information (i.e., candidate behaviors, stimuli details, contextual cues and their relevance) is attended to and recalled. Focusing attention sufficiently, however, and recalling information appropriately may be particularly challenging in modern educational practice.[10] Representing the varied competencies expected of modern-day practitioners provides assessment processes with stronger claims to construct validity, but may exceed an observer's ability to adequately consider relevant dimensions of performance.[11] Under high demand conditions people spontaneously engage cognitive behaviors aimed at reducing memory load.[12] Such strategies can include "degraded concurrent processing," where two or more tasks are completed concurrently but one or more suffers relative to when the task is performed in isolation; "strict serial processing," where multiple tasks are performed, but only one is completed at a time, leaving some tasks ignored at

any given moment; engaging heuristics; depending on schemas; and avoidant behaviours.[7,12–15] These tendencies challenge the assumption that observable performance elements are actively and appropriately considered during rater-based assessment.[9,10] They also offer an explanation for commonly reported issues including idiosyncrasy of focus, poor inter-rater reliability, and low quality feedback.

Efforts to improve feedback have focused mainly on the structure and process of how feedback is delivered. These have included targeted post-observation strategies such as guidance to explore learners' understanding of the content and/or inclusion of specific information set against a criterion.[16] We do not, however, fully understand what influences the information faculty have mentally available to them to generate feedback in the first place. It is typically assumed that raters see the same things, but prioritize issues differently or address them inadequately. However, recent research on the influence of broadening raters' focus of assessment casts doubt on the sufficiency of this interpretation by illustrating that greater variability in perception exists when raters' assessment demands are broadened.[7,9]

Such findings demand a strategy for assessment and feedback that considers all constructs included in modern competency frameworks while recognizing that asking raters to do more might reduce their effectiveness. In this study, we sought to test whether asking raters to sequentially assess a subset of a candidate's competencies altered the generation of feedback and performance ratings relative to having raters evaluate more competencies simultaneously. Given the conceptual framework outlined above, and elaborated more fully elsewhere, we hypothesized that distributing the assessment of distinct but inter-related dimensions of competence across raters would result in improved assessment outcomes without compromising the extent to which all dimensions are taken into account.[9] Our goal was not to evaluate the utility of a particular

assessment protocol, but to understand how raters' impressions differed in the two experimental conditions while anticipating that the results could inform the structure of a variety of formative and summative assessment strategies.

Method

Using a randomized between-subjects experimental design, we manipulated whether students were evaluated by an observer directed to attend to six dimensions of clinical competence (the simultaneous condition) or by three observers directed to each consider only two dimensions of competence (the sequential conditions). The dimensions listed on a previously developed global rating scale (GRS) served as the intervention.[17] Participants were asked to rate four recorded and unscripted (i.e., spontaneously generated) clinical performances and to indicate the feedback they would provide to the students in each video. Primary outcomes were indicators of the amount and type of feedback provided and the reliability of the scores observed. We tested the hypothesis that broadening a rater's focus would result in adverse effects on the generation of feedback and reliability. This study took place in Toronto, Ontario, and Halifax, Nova Scotia, Canada, with participants from multiple eligible organizations and colleges, in 2016–17. Research ethics board approval for this study was provided by Centennial College (REB#191).

Participants

We recruited paramedic educators who were functioning, or who had functioned, as an observer or rater in work and/or simulation-based training or assessment using existing email distribution lists (convenience sampling). Participants must have had clinical experience and experience with mannequin or standardized-patient based simulations for the purposes of assessment. Our recruitment letter indicated a time commitment of two hours and an honorarium for participation, which took place independent of work responsibilities.

Procedures

We asked participants randomized to the simultaneous condition to consider six domains of competence (history gathering, patient assessment, decision making, communication, resource utilization, and procedural skill). Participants randomized to the sequential condition performed the same task using one of three unique versions of the GRS that contained only two of the original six dimensions, which we pre-assigned into three pairs (version 1 = history gathering and procedural skill; version 2 = decision making and communication; version 3 = patient assessment and resource utilization). Previous work showed these pairings to have the lowest inter-item correlations and/or greatest conceptual differences, thereby ensuring that raters in both conditions had to focus on multiple dimensions of competence.[17] We provided rater training by providing orientation to the rating tool and clinical scenario, describing performance expectations, and giving frame of reference information with generic guidelines (e.g., treat dimensions independently, review rating label definitions). We used a random number generator to assign participants in a 1:3 ratio, with the latter group also being randomly assigned to one of the three two-dimensional rating tools. All participants observed the same four videos (in random order) and used only one rating form. They were not allowed to pause or review the video at any point, in order to replicate the naturalistic rating demands inherent in work-based assessment or simulation activities. They were permitted to take notes.

Rating stimuli

A different candidate was portrayed in each video. Each was a paramedic candidate responding to a deteriorating cardiac patient, who at a predetermined point, and regardless of intervention, suffered a cardiac arrest. The candidate responded to the case alone but had two first responders available who were trained to portray differences in their abilities and willingness to assist. The

two first responders were instructed to be disruptive by conflicting with one another but not obstructive of the candidate's efforts. We randomly selected four videos (of two male and two female candidates) from a pool of 80 that included performances from students currently enrolled in a paramedic training program, individuals who had just completed their training program, and working, experienced paramedics. We did not attempt to control for specific aspects of performance. Each video, created as part of a larger program of research, was nine minutes in length. Only one camera view was available (from the foot end of the stretcher).

Data collection

Immediately following each video, we instructed participants to provide formative feedback verbally and to assign numerical scores reflecting the candidate's performance on the dimensions assigned. We provided all participants in both groups with the following instructions: "Address specific and/or overall performance areas, specific to the dimensions included on your rating tool, with the intention of improving the candidate's performance in future similar cases or with any patient they might encounter in the future." Otherwise they received no prompting, directing, or leading apart from asking upon completion if there was anything else they would like to add before moving on. Participants completed the study with the help of a research assistant either in person or remotely using on-line video and audio recording technology. Recorded feedback was transcribed verbatim for analysis. We reviewed approximately 20% of the transcriptions for accuracy, observed no concerns, and proceeded with analyses.

Outcome measures and analysis

Feedback. With no agreed-upon method of evaluating feedback, we coded the transcripts in a number of ways that were informed by earlier work exploring factors influencing rater feedback, characteristics of feedback effectiveness, and the concept of content validity.[2,18,19] Our coding

focused on quantity, characteristics (accuracy, false claims, statements of uncertainty, recommendations, subjective evaluations), and content (breadth and depth of dimension coverage, feedback type). Our intention was not to make firm conclusions regarding feedback quality, recognizing that feedback is better conceptualized as a conversation, but to look for indicators and surrogates that could help determine whether the focus of feedback changed as a result of simultaneous versus sequential competency assessment.[16]

Quantity. We identified and segmented feedback into individual statements representing unique ideas (e.g., "the steps in procedure X were properly sequenced"). These were counted and compared between groups.

Characteristics. Each statement that could be confirmed as clearly linked to an observable behavior in the associated video was coded as accurate. Statements that were clearly not observable in the video were labeled as false claims. We also coded statements that were subjective evaluations (e.g., "the time spent on procedure X was appropriate"), those that described uncertainty (e.g., "I am not sure if step 2 in procedure X was done or not"), and as recommendations (e.g., "next time complete procedure X using your dominant hand"). False claims and statements of uncertainty were not included in the subsequent content analyses because they were considered construct-irrelevant data.

Content. We operationalized content validity in three ways. First, as breadth of coverage, by determining whether dimensions of performance included on the rating tool were omitted in the feedback. Second, as depth of coverage, by counting the number of unique statements included within the dimensions of performance included on the rating tool. Third, as feedback type, by coding the focus of feedback statements as describing a specific behavior or task performed, a

dimension of performance, the individual, the context, directions or recommendations, and/or encouragement of reflection.

During all coding researchers were blinded to the group condition. Three research assistants completed the coding, with disagreements resolved by the principal investigator (W.T.).

Scores and reliability. We explored the extent to which the scores assigned consistently differentiated between candidates' performances using generalizability theory. For the simultaneous rating condition, this amounted to a fully crossed design with four videos assessed on six dimensions of performance crossed with a series of raters. Videos were set as the facet of differentiation and three forms of reliability were calculated: internal consistency (the overall correlation between dimensions on the rating form), inter-rater reliability (the extent to which two raters' ratings correlate with one another), and an overall reliability coefficient that took into account both item and rater variance as sources of measurement error.

For the sake of comparison, we conducted reliability analyses on the sequential condition scores in a way that takes into account the reality that would exist if this assessment strategy were enacted. Given that the goal would be to gather a full set of competency ratings for each candidate, we combined ratings from different raters in the sequential condition to create sets that covered all six dimensions. To do so, we assigned a random number to each rater and rank-ordered them within each two-dimensional GRS version based on that random number. We then combined the dimensional ratings of each rater who possessed the same rank. Doing so resulted in a set of ratings equivalent to those collected in the simultaneous rating condition on which the same reliability analyses were conducted. To minimize the risk of randomization failure, this process was repeated three times to estimate the mean and standard deviation (SD) of the reliabilities that would be observed. Doing so allowed a sample size calculation to determine

how many observations were required to allow a sufficiently powered z-test of whether the reliabilities in the sequential condition were statistically different from the reliability observed in the simultaneous condition. That analysis suggested that 5 replications would yield a power of 0.81. To be conservative, we conducted 15 replications and report the mean and 95% confidence interval (CI) of those permutations as our best estimate of the reliabilities generated through the Sequential rating intervention.

We conducted all comparisons between groups using descriptive and inferential statistics (i.e., ANOVA, chi-square) as appropriate, with $P = .05$ set as our level of statistical significance. These analyses were conducted using IBM SPSS statistical software, version 21 (IBM Corp., Armonk, New York). We used ANOVA to calculate variance components, which were then used to calculate our reliability coefficients using generalizability theory.

Results

Of the educators we invited, 88 participants were enrolled: 23 in the simultaneous condition, and 65 across the three sequential conditions. In the latter group, 21 completed the decision making and communication tool, 22 used the history gathering and procedural skills tool, and 22 used the patient assessment and resource utilization tool. Two participants in the simultaneous condition did not provide a complete set of ratings and were excluded from reliability analyses (resulting in analyses for 21 raters). Because the minimum number of participants in the sequential condition group was 21, the 22nd ranked individual in each of the other two groups was excluded from the data aggregation after each random sort. See Table 1 for rater demographic characteristics.

Feedback

Quantity. Summing across all six dimensions, raters in the simultaneous condition offered 27.7 (95% CI: 22.3, 33.1) pieces of feedback, on average, compared to 42.9 (95% CI: 36.4, 49.4)

pieces of feedback when sets of raters in the sequential condition were created ($P < .05$ for all four videos, with F-values ranging from 7.4 to 20.0). When examined by dimension, raters in the simultaneous condition still offered less feedback than those in the sequential condition (Table 2).

Characteristics. We found no significant differences in the proportion of feedback statements that were accurate, false, subjective, indicative of uncertainty, or recommendations (Table 3).

Content. Participants in the simultaneous rating condition were more likely to give feedback where at least one dimension of performance was not represented (less breadth), dimensions of performance included only one feedback segment (reduced depth), and types of feedback were excluded (Table 4).

Scores and reliability. The means assigned by the simultaneous condition were 5.4, 2.9, 4.3 and 2.5 for videos one through four, respectively; almost identical to the means assigned by the sequential condition, which were 5.2, 2.8, 4.1 and 2.7, respectively.

Generalizability analyses performed on the ratings assigned by the simultaneous condition demonstrated inter-rater reliability = 0.58, internal consistency (Cronbach's alpha) = 0.74, and an overall reliability = 0.56. These served as the point estimates against which the reliability of the aggregated ratings provided by the sequential condition were compared. Following 15 random sorts of raters in that condition, the minimum reliabilities observed were inter-rater = 0.74, internal consistency = 0.78, and overall = 0.70. The 95% CI surrounding these means did not include the point estimates of reliability calculated for the simultaneous condition (Table 5).

Discussion

Given the dependence of health professions education on observer judgment for performance assessment,[20] it is important to understand the factors that influence rater-based appraisals.[10,21]

Models of clinical competence are broadening and greater emphasis is being placed on using assessment practices to improve future performance (assessment for learning), both of which place additional demands on observers. While previous research has demonstrated problematic outcomes when rating demands are high, the influence on feedback provision has been less clear. Further, that prior research has not offered a solution to the problem given that reducing rater burden often involves limiting the scope of practice assessed.[7] The purpose of this study was to explore that tension by examining the influence of asking a sequence of raters to assess a subset of competency domains relative to the norm of asking raters to assess all competencies simultaneously. Consistent with our conceptual framework and hypotheses, our findings suggest that broadening raters' focus has potentially deleterious effects. When asked to consider six dimensions of performance simultaneously, raters offered less feedback (overall and by dimension), were more likely to ignore some dimensions of performance, and limited the variety of feedback provided relative to observers who were asked to consider a subset of dimensions. The intervention, however, did not appear to affect their ability to generate true and false memories, or the rate at which statements were described as subjective, uncertain, or as recommendations. When scores from the sequential condition were aggregated, the reliability of the scores assigned increased as well. These findings suggest that asking raters to pay attention to fewer aspects of performance can lead to formative and summative assessments with greater utility, which has theoretical and practical implications.

Seminal work on the provision of feedback emphasizes basing feedback on "firsthand data" and observable "decisions and actions" that are considered in relation to "performance standards" and goals.[21] This assumes faculty have the capacity to detect and select meaningful information, process it in relation to learner context, ignore irrelevant data, and then translate observations

into coherent feedback. Researchers have subsequently challenged this assumption, arguing that these are complex cognitive activities that are dependent on capacity limited structures (e.g., attention, working memory).[8,12,22–24] Our own research subsequently showed that, when raters broaden their focus, they tend to mentally encode only a portion of leaners' behaviours.[7] Such cognitive limitations, thereby, reduce the opportunity to provide well-rounded feedback and increase the likelihood of disconnects of perception between raters and between raters and candidates. Limitations in the feedback provided in natural circumstances is hard to detect because observers can always provide some feedback and remain unaware of overlooked aspects of performance. This suggests that efforts to improve feedback through observer training will be insufficient because overcoming limitations induced by working memory capacity requires re-structuring the tasks we impose on observers. While problematic or ineffective feedback has been attributed to raters' limited skill, emotions, or poor insight, one additional and potentially causal mechanism may be the complexity of the demands placed on our raters when they are asked to evaluate many aspects of competence simultaneously.[23,24]

That said, there is a need to ensure that candidates satisfactorily achieve all competencies regardless of the cognitive limitations of their assessors. Validity frameworks require that all relevant constructs be adequately represented, thereby creating a new challenge if we must reduce what we ask raters to consider at a point in time.[19,25] It makes no more sense, however, to emphasize construct representation over raters' inherent capacity than it does to suggest that an invalid clinical procedure should be used because it is easier to apply than a more sensitive and specific diagnostic tool. If performance declines by requiring raters to consider all dimensions of performance simultaneously in ways that may diminish the value or utility of the assessment activity, then the fundamental goals of the assessment will not be fulfilled regardless of how

comprehensive the construct representation.

Implications

To explore this issue further, researchers will need to determine how to operationalize assessment activities that distribute focus across raters and faculty without fundamentally undermining the feasibility of assessment practices. Simulation-based settings offer opportunity for raters to consider only segments at a time, but the same is not true in work-based settings. In real clinical contexts, asking raters to consider the variety of competencies, one portion at a time, in a sequential model provides one avenue of exploration through which this tension might be resolved. Whether it is important to have multiple assessors involved in the sequential observations or to pre-specify which competencies observers should focus upon, as we did in this study, remains to be determined. A further area in need of exploration is whether different people need to be involved or if the sampling strategies outlined here can be operationalized by one individual over several points in time.

Limitations

Our findings should be considered in the context of the study's limitations. First, we chose a common construct and case stimulus to experimentally test our hypothesis, but it is possible that other combinations of tools, item pairings, and stimuli may lead to better or worse outcomes. Similarly, using more than six or less than two dimensions may lead to different findings, whereas we presume these two levels to represent points on a continuum. Second, this study was completed using a simulation-based assessment with videos of performances as stimuli. This eliminates many of the factors that would exist in workplace-based assessments where contexts are generally more complex, but also can be more authentic. Third, in this study we asked multiple raters to assess the same performance. That was important to enable exploration of the

focal issue, but may not be practical in simulation- or workplace-based settings. Our intention was not to replicate those environments precisely, but to test the hypotheses we generated based on a program of research and the conceptual framework considered. Whether or not it matters that different raters completed the three versions of the two dimensional rating form should be determined to make decisions about how to design interventions based on our results. Fourth, as is the case in all rater studies, there may have been unidentified rater variables that influenced the outcomes observed. Finally, as we described above, there is no standard for the evaluation of feedback, preventing us from claiming with certainty that the differences observed here necessarily translate into better learning outcomes on the part of the feedback recipients. We considered as many codes of feedback as we could with the intention of determining what changes, rather than proving that the feedback from one condition or the other would be more fitting for that purpose.

Conclusions

Crossley and colleagues have advocated for assessment processes that "reflect cognitive structuring."[26] In doing so they mount an argument that the interaction between the rater and the performance observed has largely been ignored by noting "the most remarkable observation might be in how irrational we have been to date with work based assessment instruments and processes." Asking raters to consider broad constructs without taking into consideration their inherent limitations (which are often masked) may be yet another example of how irrational we have been. As health professions education increasingly advocates for meaningful observer contributions as part of assessment practices, a continued emphasis on understanding faculty behaviors and limitations is needed. In many assessment contexts, raters have been asked to deliberately expand their focus to ensure appropriate assessment coverage of all competencies

expected of modern practitioners while promoting a degree of efficiency. On the surface, such changes might appear to be beneficial or innocuous. However, the results of this study suggest that what raters contribute when having to consider many dimensions of performance simultaneously is feedback that is lessened in quantity, breadth, depth, and diversity. Relative to aggregating the ratings of multiple raters who are asked to consider fewer dimensions of performance, simultaneous assessment of many dimensions generated scores with lower reliability. Therefore, sequential or distributed assessment strategies, where raters are either asked or allowed to limit their focus, may optimize formative and summative assessment efforts. In other words, when planning rater-based assessments of clinical competence, asking for less may get you more.

References

1. Eva KW, Bordage G, Campbell C, et al. Towards a program of assessment for health professionals: From training into practice. Adv Health Sci Educ Theory Pract. 2016;21:897–913.
2. van de Ridder JM, Stokking KM, McGaghie WC, ten Cate OT. What is feedback in clinical education? Med Educ. 2008;42:189–197.
3. Govaerts MJ, van de Wiel MW, Schuwirth LW, van der Vleuten CP, Muijtjens AM. Workplace-based assessment: Raters' performance theories and constructs. Adv Health Sci Educ Theory Pract. 2013;18:375–396.
4. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–1407.
5. Bandiera G, Sherbino J, Frank J. The CanMEDS Assessment Tools Handbook: An Introductory Guide to Assessment Methods for the CanMEDS Competencies. Ottawa, ON: Royal College of Physicians & Surgeons of Canada; 2006.
6. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. JAMA. 2009;302:1316–1326.
7. Tavares W, Ginsburg S, Eva KW. Selecting and simplifying: Rater behavior when considering multiple competencies. Teach Learn Med. 2016;28:41–51.
8. DeNisi A. A Cognitive Approach to Performance Appraisal: A Program of Research. New York, NY: Routledge; 1996.
9. Tavares W, Eva KW. Impact of rating demands on rater-based assessments of clinical competence. Educ Prim Care. 2014;25:308–318.

10. Tavares W, Eva KW. Exploring the impact of mental workload on rater-based assessments. Adv Health Sci Educ Theory Pract. 2013;18:291–303.

11. Kane MT. Validity. In: Brennan BL, ed. Educational measurement. Westport, CT: Praeger Publishers; 2006.

12. Wickens CD. Multiple resources and mental workload. Hum Factors. 2008;50:449–455.

13. Kool W, McGuire JT, Rosen ZB, Botvinick MM. Decision making and the avoidance of cognitive demand. J Exp Psychol Gen. 2010;139:665.

14. Botvinick MM, Rosen ZB. Anticipation of cognitive demand during decision-making. Psychol Res. 2009;73:835–842.

15. Shah AK, Oppenheimer DM. Heuristics made easy: An effort-reduction framework. Psychol Bull. 2008;134:207–222.

16. Sargeant J, Lockyer J, Mann K, et al. Facilitated reflective performance feedback: Developing an evidence- and theory-based model that builds relationship, explores reactions and content, and coaches for performance change (R2C2). Acad Med. 2015;90:1698–1706.

17. Tavares W, Boet S, Theriault R, Mallette T, Eva KW. Global rating scale for the assessment of paramedic clinical competence. Prehosp Emerg Care. 2013;17:57-67.

18. Govaerts MJ, van de Wiel MW, van der Vleuten CP. Quality of feedback following performance assessments: Does assessor expertise matter? Eur J Train Dev. 2013;37:105–125.

19. Kane M. Validating score interpretations and uses. Lang Test. 2012;29:3–17.

20. van der Vleuten C, Schuwirth LW, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: Building blocks for theory development. Best Pract Res Clin Obstet Gynaecol. 2010;24:703–719.

21. Ende J. Feedback in clinical medical education. JAMA. 1983;250:777–781.

22. Wickens CD, Carswell CM. Information processing. In: Salvendy G, ed. Handbook of Human Factors and Ergonomics. Hoboken, NJ: Wiley and Sons Inc; 2012.

23. Kogan JR, Conforti LN, Bernabeo EC, Durning SJ, Hauer KE, Holmboe ES. Faculty staff perceptions of feedback to residents after direct observation of clinical skills. Med Educ. 2012;46:201–215.

24. Telio S, Ajjawi R, Regehr G. The "educational alliance" as a framework for reconceptualizing feedback in medical education. Acad Med. 2015;90:609–614.

25. Messick S. The interplay of evidence and consequences in the validation of performance assessments. Educ Researcher. 1994;23:13–23.

26. Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. Med Educ. 2011;45:560–569.

Table 1

Rater Demographics by Group, From a Multi-Site Study of Simultaneous Versus Sequential Competence Assessment, 2016–2017

Demographic characteristic	Simultaneous raters[a]	Sequential raters[b]		
		DM + CM	HG + PS	PA + RU
Highest level of certification, no. (%)				
CCP	6 (26.0)	3 (15.0)	1 (5.0)	0 (0)
ACP	17 (73.9)	17 (85)	19 (95)	20 (100)
Age, mean (SD)	38.9 (9.0)	37.2 (8.0)	38.3 (8.6)	37.6 (8.5)
Gender, no. (%)				
Male	18 (78.3)	16 (80.0)	12 (60.0)	16 (76.2)
Female	5 (21.7)	4 (20.0)	8 (40.0)	5 (23.8)
Highest education, no. (%)				
Community college	9 (39.1)	7 (35.0)	4 (20.0)	3 (14.3)
University	5 (21.7)	3 (15.0)	7 (35.0)	3 (14.3)
Graduate school	9 (39.1)	10 (50.0)	9 (45.0)	15 (71.4)
Clinically active, no. (%)				
Yes	22 (95.7)	19 (95.0)	18 (90.0)	19 (95.0)
No	1 (4.3)	1 (5.0)	2 (10.0)	1 (5.0)
Years clinically active, mean (SD)	14.6 (8.4)	12.2 (7.2)	13.3 (7.6)	10.6 (6.5)
Years of teaching experience, mean (SD)	10.9 (7.2)	8.9 (7.0)	8.2 (6.1)	6.3 (5.9)
Years of assessment experience, mean (SD)	7.6 (6.4)	8.0 (5.2)	7.2 (5.6)	6.6 (5.6)
Previous familiarity with GRS, mean (SD)[c]	5.9 (2.8)	4.4 (3.3)	4.3 (2.8)	4.4 (2.7)
Prior rater training, no. (%)				
Yes	12 (52.2)	10 (50.0)	10 (50.0)	8 (38.1)
No	11 (47.8)	10 (50.0)	10 (50.0)	13 (61.9)

Abbreviations: DM indicates decision making; CM, communication; HS, history gathering; PS, procedural skills; PA, patient assessment; RU, resource utilization; CCP, critical care paramedic; ACP, advanced care paramedic; GRS, global rating scale.
[a]For the simultaneous condition, there were no missing data.
[b]For sequential condition, data were missing as follows: DM + CM = 1; HS + PS = 2; PA + RU = 2.
[c]The GRS is rated on a scale of 1 (not familiar) to 10 (very familiar).

Table 2

The Average Number of Unique Feedback Statements Provided Per Rater and Per Dimension for Each Candidate (Video), From a Multi-Site Study of Simultaneous Versus Sequential Competence Assessment, 2016–2017[a]

	Simultaneous condition		Sequential condition								
			DM + CM			HG + PS			PA + RU		
Video	Mean (SD)	95% CI	Mean (SD)	95% CI	F(1,43), *P* value	Mean (SD)	95% CI	F(1,43), *P* value	Mean (SD)	95% CI	F(1,43), *P* value
Video A	4.2 (2.0)	3.4 – 5.1	6.7 (4.1)	4.8 – 8.5	6.7, .01	8.1 (4.6)	6.1 – 10.2	14.0, <.01	6.3 (3.5)	4.7 – 7.9	5.9, .02
Video B	4.7 (2.5)	3.6 – 5.8	6.2 (3.0)	4.9 – 7.6	3.3, .08	7.3 (3.8)	5.6 – 9.0	7.3, .01	6.7 (4.4)	4.8 – 8.7	3.6, .06
Video C	5.1 (2.0)	4.2 – 6.0	6.9 (3.2)	5.4 – 8.3	4.9, .03	9.2 (5.2)	6.9 – 11.5	12.2, <.01	7.6 (5.9)	4.9 – 10.2	3.4, .07
Video D	4.4 (1.8)	3.6 – 5.2	6.0 (2.1)	5.0 – 7.0	7.0, .01	8.1 (4.2)	6.2 – 9.9	15.1, <.01	6.6 (4.1)	4.8 – 8.5	5.5, .02

Abbreviations: DM indicates decision making; CM, communication; HG, history gathering; PS, procedural skill; PA, patient assessment; RU, resource utilization.

[a]For each participant, the total number of unique feedback statements they generated were divided by six or two as appropriate to calculate a mean number of statements by dimension. Those were then averaged by group and are reported here.

Table 3

Proportion of All Feedback Statements That Were Identified as Accurate, Inaccurate (i.e., Not Observable in the Video), Subjective, Indicative of Uncertainty, or a Recommendation, From a Multi-Site Study of Simultaneous Versus Sequential Competence Assessment, 2016–2017

	Simultaneous condition		Sequential condition		
Statement type	Mean % (SD)	95% CI	Mean % (SD)	95% CI	F(1,43), *P* value
Accurate	23.1 (13)	17.5 – 28.7	24.9 (11.6)	19.6 – 30.2	.237, .63
False	3.2 (3.2)	1.8 – 4.6	3.2 (0.5)	1.7 – 4.6	.009, .93
Subjective	64.1 (14.1)	58.0 – 70.1	62.7 (11.4)	57.5 – 67.9	.123, .73
Indicative of uncertainty	2.0 (3.1)	.59 – 3.3	1.5 (.54)	.39 – 2.6	.249, .620
Recommendation	7.7 (2.6)	2.3 – 13.1	7.7 (1.7)	4.2 – 11.3	.000, .98

Table 4

Comparing Feedback Breadth, Depth, and Feedback Type Included Between the Simultaneous and Sequential Conditions, From a Multi-Site Study of Simultaneous Versus Sequential Competence Assessment, 2016–2017

Category	Description	No. (%) simultaneous condition (n = 23)	No. (%) sequential condition (n = 21)	X^2, df	*P* value
Breadth	At least one dimension not included on at least one video	16 (69.5)	5 (23.8)	9.2, 1	.002
Depth	At least one dimension with only one feedback segment	21 (91.0)	9 (42.9)	10.9, 1	.001
Types included[a]	Three or more types omitted on one or more video	13 (56.5)	4 (19.1)	6.5, 1	.01

[a]Types were (1) a specific behavior or task performed, (2) a dimension of performance, (3) the individual (e.g., judgments), (4) the context, (5) directions or recommendations, and/or (6) encouragement of reflection. Individual data by type are provided in Supplemental Digital Appendix 1, available at http://links.lww.com/ACADMED/A559.

Table 5

Reliability Analysis Outlining the Consistency With Which the Ratings Assigned by Both Groups Differentiated Between the Performances Observed, From a Multi-Site Study of Simultaneous Versus Sequential Competence Assessment, 2016–2017

Condition	Overall reliability	Inter-rater	Internal consistency
Simultaneous rating group	0.56	0.58	0.74
Sequential rating group[a]			
Minimum	0.70	0.74	0.78
Mean (SD)	0.74 (0.02)	0.79 (0.02)	0.82 (0.02)
95% CI	0.73 – .76	0.78 – .81	0.80 – .83

[a]Calculated based on 15 replications of a random aggregation of raters from each of the sequential rating condition groups.