Title

Are You Thinking What I'm Thinking? Exploring Response Process Validity Evidence for a Workplace-based Assessment for Operative Feedback

Journal

Journal of Surgical Education, 79(2)

ISSN

1931-7204

Authors

Zhao, Nina W
Haddock, Lindsey M
O'Brien, Bridget C

Publication Date

2022-03-01

DOI

10.1016/j.jsurg.2021.09.007

Peer reviewed

Are You Thinking What I'm Thinking? Exploring Response Process Validity Evidence for a Workplace-based Assessment for Operative Feedback

Nina W. Zhao, MD, MAEd,[*,‡] Lindsey M. Haddock, MD, MAEd,[†] and Bridget C. O'Brien, PhD[†]

[*]Department of Otolaryngology—Head and Neck Surgery, University of California - San Francisco, San Francisco, California; [†]Department of Medicine, University of California - San Francisco, San Francisco, California; and [‡]Department of Otolaryngology—Head and Neck Surgery, University of California - Davis, Sacramento, California

OBJECTIVE: Workplace-based assessments (WBAs) are used in multiple surgical specialties to facilitate feedback to residents as a form of formative assessment. The validity evidence to support this purpose is limited and has yet to include investigations of how users interpret the assessment and make rating decisions (response processes). This study aimed to explore the validity evidence based on response processes for a WBA in surgery.

DESIGN: Semi-structured interviews explored the reasonings and strategies used when answering questions in a surgical WBA, the System for Improving and Measuring Procedural Learning (SIMPL). Interview questions probed the interpretation of the three assessment questions and their respective answer categories (level of autonomy, operative performance, case complexity). Researchers analyzed transcripts using directed qualitative content analysis to generate themes.

SETTING: Single tertiary academic medical center.

PARTICIPANTS: Eight residents and 13 faculty within the Department of Otolaryngology—Head and Neck Surgery participating in a 6-month pilot of SIMPL.

RESULTS: We identified four overarching themes that that characterized faculty and resident response processes while completing SIMPL: (1) Faculty and resident users had similar content-level interpretations of the questions and corresponding answer choices; (2) Users employed a variety of cognitive, behavioral, and emotional processes to make rating decisions; (3) Contextual factors influenced ratings; and (4) Tensions during interpretation contributed to rating uncertainty.

CONCLUSIONS: Response processes are a key source of evidence to support the validity for the formative use of WBAs. Evaluating response process evidence should go beyond basic content-level analysis as contextual factors and tensions that arise during interpretation also play a large role in rating decisions. Additional work and a continued critical lens are needed to ensure that WBAs can truly meet the needs for formative assessment. (J Surg Ed 79:475–484. © 2021 Published by Elsevier Inc. on behalf of Association of Program Directors in Surgery.)

KEY WORDS: workplace-based assessment, assessment, feedback, validity, response process

COMPETENCIES: Interpersonal and Communication Skills, Practice-Based Learning and Improvement, Systems-Based Practice

INTRODUCTION

Feedback is an essential component of surgical education. While faculty and residents agree on the importance of feedback, it remains challenging to integrate into daily clinical practice.[1,2] Workplace-based assessments (WBAs) are increasingly popular tools to aid surgical residents' progression toward operative competence. Although WBAs were designed to provide more frequent and timely feedback as a form of formative assessment, the validity evidence supporting the use of WBAs for this purpose is limited. Formative assessment is characterized by its purpose to support learning with a goal to generate meaningful feedback for learners to shape future performance.[3] Therefore, to claim

Correspondence: Nina W. Zhao, MD, MAEd, Department of Otolaryngology—Head and Neck Surgery, University of California, Davis, 2521 Stockton Boulevard, Suite 6201A, Sacramento, CA 95817. Fax: 916-703-5124; e-mail: nina.zhao.md@gmail.com

WBAs can be used for formative purposes, it is important to understand user response processes, or the actions, strategies, and interpretations of the individuals responding to an assessment.[4-7]

Unfortunately, evidence based on response process is one of the least reported in medical education.[8] Most prior work has focused on the reliability of WBA scores and validity evidence based on the association of assessment scores with other variables, and studies that do report evidence for response process often provide incomplete or inaccurate information.[9,10] In a recent systematic review of technical skills assessments in surgery, the majority of the evidence classified as response process involved methods to control technical and procedural quality such as rater training or blinding assessors to trainees.[10] While these techniques may help with standardization of assessment scores, they do not actually tap into how WBA users interpret the assessment questions (items) and corresponding answer choices (ratings/scores).[7]

A key challenge is the lack of clarity about what constitutes response process validity evidence, particularly when considering assessments designed for formative purposes in medical education. In the broader assessment literature, current validity theorists argue that evidence based on response process involves identifying the mechanisms underlying what people do, think, or feel when interacting with the assessment item(s) or task, as well as the broader context in which the responses are generated, both of which are essential to forming a deeper understanding of score meaning.[7]

Therefore, investigating the response process validity evidence for WBAs requires a better understanding of not only how users interpret the items in a WBA, but also what factors mediate their decisions to produce the observed scores. Our study aims to answer the following questions: (1) how do users interpret the questions and rating choices when completing a surgical WBA? and (2) what contextual elements influence their rating decisions?

METHODS

Study Design and Participants

We conducted a qualitative interview study during a 6-month pilot of a surgical WBA for operative performance, the System for Improving and Measuring Procedural Learning (SIMPL) in the Department of Otolaryngology—Head and Neck Surgery at the University of California, San Francisco. Residents and faculty completed SIMPL training in January and February 2020. The department launched SIMPL in March 2020 and then again in July 2020 to re-invigorate use after disruptions from COVID-19. The pilot period was extended until December 2020. All 21 resident physicians and 25 of 34 total clinical faculty agreed to participate in the pilot. All pilot participants were contacted via email and invited for an interview. Residents received a $50 Amazon gift certificate for completing the interview; faculty were not offered an incentive. The University of California, San Francisco Institutional Review Board approved this study as exempt.

SIMPL

SIMPL is a smartphone-based tool designed to improve operative feedback by collecting both faculty and resident ratings of resident performance during surgical procedures. Faculty also dictate feedback to share with the resident. An assessment consists of three questions: first, a rating of the level of resident autonomy using the 4-point Zwisch levels from Show & Tell to Supervision Only;[11,12] second, a global rating of resident operative performance on a 5-point set of ordered categories from unprepared to exceptional, similar to prior operative performance rating scales;[13] and third, a rating of the difficulty of the procedure as easiest 1/3, average, or hardest 1/3.[14] For faculty members, the questions are followed by a screen where they can record verbal feedback that will be sent to the resident. According to the developers, the goal of the first question is to differentiate between levels of faculty guidance, the second question is to measure readiness for independent practice, and the third to judge procedural difficulty based on patient-related factors relative to other similar procedures. Prior work has supported that as resident performance improves, the amount of autonomy they receive increases[15] and that increasing case complexity is associated with decreased autonomy.[12] The option for dictated feedback allows faculty to provide specific feedback on what the resident did well and where they can improve.

Either party can initiate an assessment, and the other will then be notified and asked to complete an identical assessment within 72 hours. The resident and faculty ratings as well as the faculty dictations are available for the resident to review in real-time. With the exception of the program director, the faculty do not have access to the resident self-ratings. In this pilot, residents, rather than faculty, were encouraged to initiate the assessment.

Interviews

Three months after the launch of the SIMPL pilot, the first author (NZ) started conducting semi-structured one-on-one video interviews with resident and faculty members in the department. Participants were asked to describe the most recent time they used the app and describe their thought processes while answering the three SIMPL questions. Their responses were then

TABLE 1. Key Interview Questions.

1. Describe the last time that you used the application, what was the procedure and the context?
2. Please read the question out loud. How would you restate the question in your own words?
3. How did you answer this question? How did you come up with your answer?
4. How do you differentiate between the answer choices?

followed-up with verbal probes for further elaboration. Key interview questions appear in Table 1; the full interview protocols are provided in the Appendix.

The interview protocols were piloted with two individuals (one attending and one resident) 3 months after the initial launch, and the questions were revised for clarity as needed. These interviews were included in the final analysis as they yielded data of similar quality to the regular interviews. All interviews were audio recorded and deidentified. They were then transcribed by a commercially available transcription service (Rev.com).

Data Analysis

Data analysis began during data collection and continued in an iterative fashion using a qualitative data management software program (Dedoose v8.3.41). Three researchers first performed directed qualitative content analysis of the data[16] through theoretically informed coding followed by inductive, open coding. The first author developed an initial set of codes through the lens of validity theory and refined the codebook after reading the first two transcripts, categorizing response process evidence by question item. The codes were applied to the subsequent transcripts by all researchers. Researchers met on a regular basis to discuss discrepancies and reconcile codes. The first author then reviewed the coded excerpts and inductively developed categories that further described faculty and resident users' response processes. Then, using a constant comparative approach,[17] the first author compared and contrasted data in and among categories to understand relationships and determine central themes. Data collection continued until researchers felt they had generated substantial insight into response processes and little new information was obtained with subsequent interviews.

Reflexivity

The study's primary author is an otolaryngologist with a master's degree in education. Her position not only allows her to understand the barriers of surgical training in general, but also to reflect on the culture of training within the subspecialty. The other researchers on this project include a geriatrics physician with a master's in education and a PhD-trained education scholar with a faculty appointment in the Department of Medicine. These individuals have experience in qualitative data analysis and understand the nuances of the medical field. At the same time, they are not surgeons, so they bring less context and assumptions about meaning to their interpretations of the data.

RESULTS

Participants

Eight resident trainees and 13 faculty members completed interviews between June 2020 and March 2021. Participant demographics are summarized in Table 2. Residents from all five post-graduate training years (PGYs) participated. Faculty practices encompassed a breadth of subspecialties. Time in practice ranged from 3 to 31 years with a mean of 10.9 years.

Response Process

Using our interview data, we identified four overarching themes that characterized faculty and resident response processes while using the SIMPL WBA: (1) Faculty and residents expressed similar content-level interpretations of the questions and corresponding answer choices; (2) Faculty and residents employed a variety of cognitive, behavioral, and emotional processes to make rating decisions, especially when the frame of reference was ambiguous; (3) Contextual factors not directly related to resident performance influenced ratings; and (4) Tensions in interpretation contributed to rating uncertainty.

Theme 1. Faculty and residents expressed similar content-level interpretations of the questions and corresponding answer choices.

At face value, participants exhibited similar understandings of what each question was asking and the

TABLE 2. Participant Demographics

Residents (n = 8)	n (%)	Faculty (n = 13)	n (%)
Gender		**Gender**	
Men	1 (12.5)	Men	8 (61.5)
Women	7 (87.5)	Women	5 (38.5)
Year in training		**Rank**	
PGY1	1 (12.5)	Assistant Professor	5 (38.5)
PGY2	2 (25.0)	Associate Professor	5 (38.5)
PGY3	1 (12.5)	Full Professor	3 (23.1)
PGY4	1 (12.5)	**Subspecialty**	
PGY5	3 (37.5)	Rhinology	3 (23.1)
		Pediatrics	3 (23.1)
		Laryngology	2 (15.4)
		Head and Neck Oncology	2 (15.4)
		Neurotology	1 (7.7)
		Facial Plastics	1 (7.7)
		General/Sleep	1 (7.7)

general considerations for distinguishing the answer choices. Faculty and residents interpreted the first item, "how much guidance did you provide/receive for the majority of the critical portion of the procedure," as a question about operative supervision and resident independence. Both groups described three main features related to differentiating levels of guidance: (1) who does the dissection, (2) who makes the decisions, and (3) the frequency and quality of faculty comments.

As one faculty explained:

> "Active help is the resident doing the procedure that is kind of on my verbal cues or commands. . . I would be directly verbally guiding every move. . .Passive help is the resident doing the procedure with intermittent verbal cues like, 'You don't probably need to do that. . .Go ahead and elevate that flap.' And then I'd be quiet unless I needed to say something intermittently. . .Supervision only is I'm just watching and then kind of maybe intermittently or very intermittently saying things, but hardly at all." (Faculty 5)

Participants felt the second question, "what was this resident's performance for the majority of the critical portion of this procedure," was meant to reflect how well the resident operated during the surgery. Faculty and residents discussed that performance was related to the trainee's prior experiences with the procedure and considered not only the trainee's *knowledge* of the procedural steps, but also their ability to *execute* the steps. According to a resident:

> "I think there are two main things to me. One is whether or not I know what the steps of the procedure should be. And then the second part is whether or not I'm actually able to do the steps of the procedure. . .like okay, I know that the first thing I need to do [in a tonsillectomy] is make an incision and. . .find the plane between the tonsillar pillar and the tonsil. And so, knowing that I need to do that is one thing, but am I actually able to find the plane and feel comfortable that I [am]. . .able to execute it? I think those are the two main components I think about in terms of performance." (Resident 8)

Finally, both groups discussed complexity of the case based on patient or case-specific factors that were separate from the resident's performance. For example, one faculty member stated:

> "Well it's like size of tumor, if they've been radiated, their scar, some cases are just challenging because of blood vessels, or they're oozy, or the exposure is hard. So those will all dictate into whether something is easy, average, or hard. . . .the resident has nothing to do with those choices. It has to do with the patient's condition, and how the operation went." (Faculty 3)

Theme 2. Faculty and residents employed a variety of cognitive, behavioral, and emotional processes to make rating decisions, especially when the frame of reference was ambiguous.

Despite their similar understanding of the content of the questions, faculty and residents used a number of cognitive, behavioral, and emotional processes to make rating decisions when implementing the assessment in practice. The variability in participants' decision-making processes was most prominent when the question was unclear about the frame of reference, or standard for judgment or comparison. We summarize these findings from the most stable frame of reference to the least.

Autonomy Question and Ratings

Reflection on faculty behavior: Participants had the clearest sense of the frame of reference in the question about resident autonomy. To select ratings for this question, participants described a process of reflecting on the faculty supervisor's behavior during the case and analyzing their extent of involvement. As a resident described:

> "The most recent case I've done it for was a. . .trach. . .I think about like, how much of it did I need an attending's help with the procedure? I think about. . .the critical portions of dissecting onto the cricoid or dividing the thyroid and where did I enter the airway and thinking about how much help did I need with modifying my plan to decide. . .the amount of guidance." (Resident 3)

Case Complexity Question and Ratings

Comparison with prior procedures: Both faculty and residents framed the third question regarding case complexity as a comparison to other groups of similar procedures. As one attending reported:

> "What I tend to do is to think of it, not in terms of relative to my personal practice, but what that procedure would be. I feel. . .most of the some tympanoplasties that I would do tend to be more complex than your sort of average tympanoplasties across the entire practice of otolaryngology. . .that's how I've been in calibrating it." (Faculty 4)

However, residents were unclear if they were supposed to consider the case based on their own personal experiences or based on the faculty's point of view. As a result, they employed other processes to arrive at their final ratings, such as embodied processes or reliance on external cues.

Embodied processes: Some residents discussed making decisions based on embodied reactions, such as a 'gut feeling' or sense of struggle.

> "...I try to look back and sort of think about, compared to other cases I've done, where does this one fall in the spectrum, but also, I just kind of go with my gut and say, 'Did I feel like I struggled with this?' Just in and of itself that should tell me whether or not I thought it was hard or not. If I don't feel any particular way, I sort of say average." (Resident 3)

External cues: Other residents attempted to see things from the attending perspective using external signals to guide their decisions.

> "I always put myself in the attending's head because obviously they have more experience and they have more cases in their head. And they have a bigger sample size for whether something is easy, hard, or average...if they were very frustrated, it was probably one of the harder ones. If they were in a really good mood...probably easiest...And [if] they had to focus a little bit, but they were pretty chill, then it was average. It's generally me gauging the vibe of the room..." (Resident 4)

Performance Question and Ratings

Faculty and residents exhibited the most uncertainty regarding the frame of reference for the second question about operative performance. As one faculty questioned, "Well, intermediate relative to what? All residents? What a normal practitioner would do? What someone at their PGY level would do? There's a little bit of that ambiguity..." (Faculty 6) The variability in the decision-making processes was also the most pronounced and included considering a resident's prior experience, evaluation of self-comfort, extrapolation, and emotional esponses.

Consideration of resident experience: Resident's prior amount of experience was one of the most basic methods participants used to make decisions about operative performance. As one resident stated, "'Unprepared' means...first time ever [doing the procedure] as opposed to inexperienced because I've done it once or twice." (Resident 5) For faculty, this was often deduced from conversations with the resident about what they have seen or done before:

> "[If] a resident tells me like, 'I've never seen this done before. I've never done this before,' then that's inexperienced with the procedure...where like, 'Well, I've seen this, but I've never done it,' or 'I've only done one,' that kind of puts them into intermediate performance...that they've got some experience with it but not a lot." (Faculty 5)

Some faculty noted this interpretation became problematic when the procedure was completely new to the resident as there was not a clear delineation of how much to weigh prior experiences.

> "There's definitely times when it was someone's first time doing this type of case and so then...what's the difference in that scenario between an intermediate performance and an inexperienced with the procedure?... A lot of times, people are inexperienced with the procedure but they still might have an intermediate performance or potentially a practice-ready performance, at least depending on how you assess things." (Faculty 6)

Evaluation of self-comfort: Faculty and residents also relied on their own self-comfort to make rating decisions, especially when they had little prior experience for comparison. For faculty, this experience was often related to how long they had been an attending and working with residents; for residents, this experience was how many times they had previously performed the procedure. As one resident explained:

> "A lot of times I'm filling this out for surgery that I haven't done very many times. And then I feel like, well, I don't know how to compare my performance to other times I've done the same surgery. But I can compare my comfort level. That would be easier for me to compare it. And my comfort level ends up making me feel like, oh, if I was really comfortable, then I feel pretty good about that." (Resident 5)

Extrapolation: Sometimes, participants found themselves making decisions based on extrapolation, or formulating inferences about an unknown situation based on current information. This mechanism occurred most often during decisions about what made a resident "practice-ready."

> "Practice-ready performance...you're trying to sort out if they were on their own and you weren't in the room, would they complete the critical portions of the procedure in an adequate manner?" (Faculty 6)

Emotional responses: Members in both groups also expressed emotional reactions to the valence of specific answer choices that drove their rating decisions. "'Critical deficiency' makes you not want to check that," (Faculty 2), according to one faculty. Another also described a significant undesirable connotation with the same answer choice. "I never pick that for that question...I would really think

carefully before... that's such a dis on somebody to send that out. You are going to battle with that person." (Faculty 3) This faculty member also found an overly positive tone challenging: "I don't think I ever chose exceptional either. Because. . .you don't want people resting on their laurels." (Faculty 3)

Theme 3. Contextual factors not directly related to resident performance influenced ratings.

Contextual factors were situational features surrounding, but not directly related to, the assessment in the moment. These factors included prior faculty-resident interactions, the resident's year of training, and potential patient consequences.

Prior Faculty-Resident Interactions

The nature of previous faculty-resident interactions influenced ratings in ways that were not directly related to a resident's capabilities and operative performance. In particular, participants noted that regardless of a resident's prior experiences, faculty naturally offered more supervision when operating with a resident for the first time on a particular procedure.

> "The feedback that you get and the guidance that you get does not always correlate to your own perception of competency and performance. So, if I felt really comfortable with the maxillary antrostomy, but it was like an attending's first time ever seeing me do it, they might offer more help than I think I need." (Resident 4)

Resident Year of Training

The level of the resident also influenced rating decisions in various ways. Many faculty indicated that the resident's level set specific expectations for the case, and they considered rating decisions through that lens: "I try to think about the score in the context of where the resident is in their training, and whether that meets my expectations." (Faculty 4) For some faculty, this lens was similar to a halo effect, where as a resident's level increased, they not only had higher expectations, but also were primed to believe that the resident could meet these expectations:

> "My expectations are much higher if it's a Chief or a [PGY-]4. . .a Chief should be practice-ready performance. It's concerning if they're intermediate performance. . .I think for a senior resident, I really want to do the practice-ready performance. I'm looking for that. I want to believe that they're ready." (Faculty 10)

Conversely, junior resident level was associated with a ceiling effect. Both faculty and residents reported a tendency to not choose higher rating levels for junior residents despite good performance. In particular, residents described a strong sense of selecting ratings that were socially desirable according to their level of training.

> "I was able to do the tonsillectomy from start to finish on my own, and I felt confident doing it. [But intermediate performance] is the highest of the three options that I could pick within what I'm comfortable picking, because. . .I'm still an R2. . .So, the fact that I was able to do the case on my own, I was like, 'Okay, I think I should pick the highest of the three that I think are available to me.' That's kind of how I thought that through." (Resident 8)

Potential Patient Consequences

Some faculty described how their approach to patient care and the potential consequences of a procedure impacted the way they operated with residents, which affected the ratings in a way that did not reflect true resident performance.

> ". . .I would say there's highly competent people that get active help on a case because I took that active help to mean I was there manipulating tissues with them, not just providing exposure. . . because like an oral cavity or a pharynx tumor, the margin is so important that you can't not be there. It's a total disservice to the patient. . .So, I'm mostly actively helping. Again, it's not a reflection of the resident. It's a reflection of the demands of being a surgeon in my field." (Faculty 3)

Theme 4. Tensions during interpretation contributed to rating uncertainty.

Finally, we identified several conceptual tensions that arose during the rating process. These tensions occurred as participants attempted to interpret the questions and answer choices within their specific context. The resulting interpretations were variable and led to uncertainty in the final rating. The tensions we identified included those related to the definition of the 'critical portion' of the procedure and the balance between resident autonomy and faculty teaching.

Definition Versus Practical Usage of "Critical Portion"

Users generally understood that the strict definition for "critical portion" meant the portion of the procedure that was essential for getting the case completed. However, in practice, both faculty and residents indicated that their conceptualization of resident operative

competence was much more global than for specific critical portions. According to one faculty:

> "So in sinus surgery, in particular, there are many critical portions that occur with each component of the surgery... I generally think of things more globally unless, for instance, I'm doing a sinus surgery and there's really one part to it...So I tend to think of it, in a normal sinus case, I tend to think of it more globally in terms of how active I was as a participant." (Faculty 11)

Others interpreted the critical portion as what was critical for the current stage of resident learning, which changed over time:

> "The critical portion] varies and probably changes with year... At this point, I'd say facial recess drilling and implant placement I would say are critical portions of the procedure. Probably more so for residency overall and the fact that I will not be an otologist is probably facial recess drilling... So, I think as a junior, it...probably would have been the mastoid...But as a more senior, kind of the further steps of the case." (Resident 7)

Resident Autonomy Versus Faculty Teaching

Both faculty and residents struggled with how to consider the balance between resident autonomy and faculty teaching when making rating decisions. Participants described that as the faculty commentary during the case increased, it often led to a decreased sense of resident autonomy; however, the need to provide instruction was often not because of an actual need for supervision, but more of taking advantage of a teaching moment. As one faculty explained:

> "I felt like there are times when, as the supervisor, I can be quiet, can give them a higher passive help or supervision-only grade. Or as a teacher, I can try to talk more and discuss more about some of the nuances of the procedure, in terms of what sequence or what instrument you're choosing to use...And so, I struggled with rating this question a little bit, in terms of I tended to err on the talking more side. Is that really active help, if I'm suggesting they use a different instrument because I think it's important for them to see how the case is done with that instrument, versus something else?" (Faculty 6)

DISCUSSION

This study sought to explore the validity evidence based on response processes for a WBA designed for resident operative performance assessment. The results not only have implications for the validity of the use of this specific WBA, SIMPL, as a formative assessment for feedback, but also provide a framework for investigating response process validity evidence for WBAs as a whole.

Validity Evidence for SIMPL

For SIMPL, our response process evidence revealed that while faculty and resident users had very similar content-level interpretations of the questions and answer choices, in practice, they employed a wide variety of cognitive, behavioral, and emotional processes when selecting the ratings and were influenced by multiple contextual factors and tensions. However, because formative assessment is meant to generate meaningful feedback to aid learning, faculty and residents should have similar approaches to score interpretations that are also supported by the assessment developers. Without this alignment, it becomes challenging to know how to use the assessment results. In our study, faculty and residents interpreted SIMPL questions and responses in more diverse ways than anticipated by the original creators.[14] As a result, simply receiving a set of scores without an understanding of the rater's cognitive processes and the contextual factors influencing the ratings may not be sufficient to support formative learning experiences. These findings indicate there are opportunities to improve the validity of SIMPL as a formative assessment of operative performance.

Several solutions could be implemented to strengthen the validity evidence for the use of SIMPL as a formative WBA. First, we found more variability in interpretations of the question related to operative performance, which participants indicated had an ambiguous frame of reference. Therefore, the question can be altered to provide a more stable frame or additional frame of reference training may be needed. However, no amount of training can eliminate all variability; in fact, maintaining variability may even be desirable for authentic assessment.[18,19] Therefore, a second solution is to take advantage of SIMPL's option to record verbal dictations of feedback. Although this paper does not report on the dictations, we believe that this feature could be beneficial for facilitating residents' understanding of the ratings they receive from faculty. In addition to providing feedback about what was done well and what could improve, faculty could be prompted to describe their approaches to question interpretation and the factors influencing their rating decisions. Finally, a third solution is to alter implementation; rather than asking faculty and residents to complete these assessments in isolation, the program can use the assessments to encourage feedback conversations. In this way, the various factors involved in decision-making can be shared, promoting a relationship-based and learner-centric feedback culture.[20,21]

Validity Evidence and WBAs

For WBAs as a whole, our work contributes to the overall literature regarding validity evidence for operative performance assessments by examining a previously underreported source of evidence.[5,8,10] We were unable to find any other studies of performance or skill assessments in the medical education literature that explicitly included gathering evidence based on response process *a priori* within their study design. Furthermore, we found only one study that has attempted to construct a complete validity argument for a surgical assessment tool, the Objective Structured Assessment of Technical Skills.[22] While the majority of the available evidence was for the use of the Objective Structured Assessment of Technical Skills for formative assessment purposes, there were still no reports including validity evidence based on response process.

Our study not only underscores the importance of evaluating response processes in assessment validation but also highlights the need to explore different facets of response processes beyond the understanding of question content to include respondent decision-making processes for ratings and the influence of context. Traditionally, when response process is discussed, it is equated to purely cognitive models of responding (i.e. focusing on mental operations).[7,23] However, more contemporary views on response process expands the definition to include emotions, motivations, and behaviors as well as the situational, cultural, or ecological aspects of testing.[7] In our interviews, we identified multiple contextual elements beyond the individual user that mediated item interpretation and score selection. Future work may focus on novel methods and frameworks to further develop explanatory models for both response process validity evidence and assessment validity as a whole.

Our findings regarding the multiple frames of reference parallel findings in the literature on rater cognition.[24,25] This literature has similarly identified multiple frames of reference and approaches to assessment interpretation that contribute to variability in rater decisions.[25] However, this work has largely examined the issue of high interrater variability from the lens of poor interrater reliability of scores. While rater cognition also explores how individuals make assessment and rating decisions, it has rarely, if ever, been framed as response processes. There may be an opportunity to bridge these lines of inquiry to strengthen the overall validity evidence for WBAs. Our study identifies specific ways in which the clarity of the frame of reference in each question impacts the variability of the processes or approaches the raters use to generate the score. Considering rater cognition from the lens of response process validity can help us reflect on the issue of rater variability as it relates to assessment purposes as well as guide future directions for response process validation research.

Response processes are only one piece of the validity argument. Yet the evidence we gathered in this study may also have implications for other sources of validity evidence. For example, we found that many faculty used resident training level as a proxy for competence, leading to *a priori* beliefs that a senior resident would be practice-ready for certain procedures. This finding calls into question the strength of correlating rating scores to resident training level as validity evidence based on relationship to other variables as the ratings are not independent from resident level. In addition, our response process evidence serves as a foundation for future work examining the validity evidence based on consequences of assessment, which would help us further understand if and how faculty and residents use the assessment to support their teaching and learning practices.

Limitations

Limitations to our study include the retrospective nature of our interviews, as it was not feasible to conduct the interviews at the time the participants were generally completing the assessment (e.g. in between surgical cases, at home in the evenings). The participants were volunteers and not everyone who completed the pilot agreed to an interview. The interviews were also completed with users within a single department; additional complexity may likely be uncovered when comparing across diverse institutional cultures. However, our work offers new insights into the evaluation of response process validity evidence and the transferability of these findings can be investigated in other settings.

CONCLUSION

Response processes are a key source of evidence in formulating robust validity arguments for the formative use of WBAs. Evaluating response process evidence should go beyond basic content-level analysis as user decision-making processes and contextual factors also play a large role in influencing rating decisions. Additional work and a continued critical lens are needed to ensure that WBAs can truly meet the needs for formative assessment.

FUNDING

This work was supported by a 2020 Innovations Funding for Education grant by Academy of Medical Educators at the University of California, San Francisco.

APPENDIX: INTERVIEW PROTOCOL

You have been asked to speak with us today to learn more about your perceptions of feedback, teaching and learning, and your experiences with the SIMPL application. There are no right or wrong answers; our study does not aim to evaluate your techniques or judge your experiences. Rather, we are trying to learn more about teaching, learning, and feedback practices and to gain information that will help improve these areas in our department in the future.

To facilitate our data analysis, we would like to audio record our conversation today. All information you provide today will be held confidential and any transcriptions will be anonymized. Your participation is voluntary, and you may stop at any time if you feel uncomfortable, just let me know. Is it okay if we proceed?

- What is your understanding of the purpose of the SIMPL app?
  - Follow-up: How has it borne out for you in practice?
- How frequently do you use the app? How do you decide to use it?
  - Possible probes: What barriers do you experience? What makes it easier?
- How frequently do you provide/receive dictated verbal feedback?
- Describe the last time that you used the application, what was the procedure and the context?
- Please read the question out loud. (Questions shown for reference) How would you restate the question in your own words?
- How did you answer this question? How did you come up with your answer?
  - Possible probes: What does this term mean to you? Why did you select this answer?
- How do you differentiate between the answer choices?
  - Possible probes: What aspects are confusing to you?
- How could this question be improved?

REFERENCES

1. Nathwani JN, Glarner CE, Law KE, et al. Integrating post-operative feedback into workflow: perceived practices and barriers. *J Surg Educ*. 2017;74:406–414. https://doi.org/10.1016/j.jsurg.2016.11.001.
2. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. *Perspect Med Educ*. 2015;4:284–299. https://doi.org/10.1007/s40037-015-0231-7.
3. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. *Med Educ*. 2019;53:76–85. https://doi.org/10.1111/medu.13645.
4. American Educational Research Association. American Psychological Association, National Council on Measurement in Education, eds. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 2014.
5. Cook DA, Lineberry M. Consequences validity evidence: evaluating the impact of educational assessments. *Academic Medicine: Journal of the Association of American Medical Colleges*. 2016;91:785–795. https://doi.org/10.1097/ACM.0000000000001114.
6. Padilla J-L, Benítez I. Validity evidence based on response processes. *Psicothema*. 2014;26:136–144. https://doi.org/10.7334/psicothema2013.259.
7. Hubley AM, Zumbo BD, eds. *Understanding and Investigating Response Processes in Validation Research*. 1st ed. 2017. Cham, Switzerland: Springer International Publishing: Imprint: Springer; 2017. https://doi.org/10.1007/978-3-319-56129-5.
8. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20:1159–1164. https://doi.org/10.1111/j.1525-1497.2005.0258.x.
9. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ*. 2010: 341. https://doi.org/10.1136/bmj.c5064.
10. Vaidya A, Aydin A, Ridgley J, Raison N, Dasgupta P, Ahmed K. Current status of technical skills assessment tools in surgery: a systematic review. *J Surg Res*. 2020;246:342–378. https://doi.org/10.1016/j.jss.2019.09.006.
11. DaRosa DA, Zwischenberger JB, Meyerson SL, et al. A theory-based model for teaching and assessing residents in the operating room. *J Surg Educ*. 2013;70:24–30. https://doi.org/10.1016/j.jsurg.2012.07.007.
12. George BC, Teitelbaum EN, Meyerson SL, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. *J Surg Educ*. 2014;71:e90–e96. https://doi.org/10.1016/j.jsurg.2014.06.018.
13. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery*. 2005;138:640–649. https://doi.org/10.1016/j.surg.2005.07.017.

14. Bohnen JD, George BC, Williams RG, et al. The feasibility of real-time intraoperative performance assessment with SIMPL (System for Improving and Measuring Procedural Learning): early experience from a multi-institutional trial. *J Surg Educ*. 2016;73:e118–e130. https://doi.org/10.1016/j.jsurg.2016.08.010.

15. Chen XP, Williams RG, Sanfey HA, Dunnington GL. How do supervising surgeons evaluate guidance provided in the operating room? *Am J Surg*. 2012;203:44–48. https://doi.org/10.1016/j.amjsurg.2011.09.003.

16. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res*. 2005;15:1277–1288. https://doi.org/10.1177/1049732305276687.

17. Mathison S. Constant Comparative Method. Encyclopedia of Evaluation. Thousand Oaks, CA: Sage Publications, Inc.; 2005. https://doi.org/10.4135/9781412950558.n101.

18. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Acad Med*. 2011;86(10 Suppl):S1–S7. https://doi.org/10.1097/ACM.0b013e31822a6cf8.

19. Govaerts M, Vleuten CP van der. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47:1164–1174. https://doi.org/10.1111/medu.12289.

20. Ramani S, Könings KD, Ginsburg S, Vleuten CPM van der. Twelve tips to promote a feedback culture with a growth mind-set: swinging the feedback pendulum from recipes to relationships. *Med Teach*. 2019;41:625–631. https://doi.org/10.1080/0142159X.2018.1432850.

21. Telio S, Ajjawi R, Regehr G. The "educational alliance" as a framework for reconceptualizing feedback in medical education. *Acad Med*. 2015;90:609–614. https://doi.org/10.1097/ACM.0000000000000560.

22. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. *Adv Health Sci Educ Theory Pract*. 2015;20:1149–1175. https://doi.org/10.1007/s10459-015-9593-1.

23. Embretson SE. Understanding examinees' responses to items: implications for measurement. *Educational Measurement: Issues and Practice*. 2016;35:6–22. https://doi.org/10.1111/emip.12117.

24. Gauthier G, St-Onge C, Tavares W. Rater cognition: review and integration of research findings. *Med Educ*. 2016;50:511–522. https://doi.org/10.1111/medu.12973.

25. Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E. Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 2011;45:1048–1060. https://doi.org/10.1111/j.1365-2923.2011.04025.x.