Global Surgical Education - Journal of the Association for Surgical Education (2024) 3:44
https://doi.org/10.1007/s44186-024-00232-x

ORIGINAL ARTICLE

Not so simple: evaluating consequences validity evidence for a workplace-based assessment in surgery

Nina W. Zhao[1,2] · Lindsey M. Haddock[3,4] · Bridget C. O'Brien[3]

Received: 2 June 2023 / Revised: 10 January 2024 / Accepted: 2 February 2024

Abstract

Purpose Workplace-based assessments (WBAs) of trainee operative skills are widely used in surgical education as formative assessments to facilitate feedback for learning, but the evidence to support this purpose is mixed. Further evaluation of the consequences of assessment use and score interpretation is needed to understand if there is alignment between the intended and actual impacts of assessment. This study examines consequences validity evidence for an operative WBA, exploring whether WBA use is consistent with the goals of formative assessment for learning.

Methods Eight residents and 9 faculty within the Department of Otolaryngology—Head and Neck Surgery at a tertiary institution completed semi-structured interviews after participating in a pilot of a surgical WBA, the System for Improving and Measuring Procedural Learning in the OR (SIMPL OR). Residents received feedback from attendings via both scores (performance and autonomy ratings) and recorded dictations. Interview questions explored faculty and resident perceptions of feedback behaviors and perceived impacts on their teaching or learning practices. Three researchers analyzed transcripts using directed qualitative content analysis to generate themes and evaluated how the perceived impacts aligned with formative purposes for assessment and score use.

Results Both faculty and residents identified intended impacts of formative assessment, including (1) greater emphasis on feedback, (2) support for a postoperative feedback routine, and (3) facilitation of case-specific reflection. Residents also used score and verbal feedback for (1) calibrating case perceptions and (2) benchmarking performance to an external standard. The recorded dictations supported feedback by (1) providing context for ratings, (2) facilitating review of dictated feedback, and (3) prompting faculty for deliberate feedback. Unintended impacts included: (1) emotional discomfort during the assessment process, (2) increased feedback frequency but not diversity or quality, (3) inadequate support for feedback conversations, and (4) limited next steps for teaching or learning. Assessment usage declined over the pilot period.

Conclusions The validity evidence gathered in this study suggests an operative WBA can be used for formative purposes to improve perceptions of feedback, but unintended consequences and implementation challenges limited ultimate impacts on teaching and learning. User perspectives can add important elements to consequences validity evidence and should be further evaluated in different implementation settings to better understand how WBAs can achieve their formative goals.

Keywords Workplace-based assessment · Formative assessment · Validity · Consequences validity evidence

Introduction

As surgical training continues moving toward a competency-based model, educators have sought ways to better support trainee skill advancement through assessment [1–3]. To address this need, workplace-based assessments (WBAs) of trainee operative skills and performance have been implemented as a form of formative assessment to improve learning through feedback [4–6]. A key feature of WBAs is that they are observed assessments occurring during authentic practice to evaluate performance in context [7]. Although

Meeting Information This work was presented at the Association for Surgical Education Annual Meeting, San Diego, California, April 2023.

Extended author information available on the last page of the article

Published online: 05 April 2024

WBAs have gathered considerable momentum, prior systematic reviews of WBAs and other direct observation tools have revealed limited validity evidence supporting their use for formative purposes. Information on how these assessments support learning and performance improvement is lacking [7–9]. In fact, evidence suggests that WBAs are often caught between the dual purposes of formative and summative assessment, which may threaten their validity [9].

Validity is fundamental to any assessment. However, it is not a property of an assessment itself; rather, validity is an argument constructed from multiple sources of evidence to support (or refute) the interpretation and use of assessment results for a specific purpose [10, 11]. The *Standards for Educational and Psychological Testing* defines five sources of evidence based on Samuel Messick's work that underlie validity: content, response process, internal structure, relations to other variables, and consequences. While all sources of evidence contribute to a complete validity argument, evidence for consequences may be particularly important for formative assessments. Consequences evidence recognizes that all assessments have impacts, both intended and unintended. In the case of formative assessment, a key consequence is that the assessment must support learning by providing information that can be used by both learners and teachers for decisions and actions, such as determining what needs to be studied or changing instructional techniques based on assessment results [12]. As a result, the consequences validity argument for WBAs should focus on the claim that the results of these assessments can be used as feedback to stimulate further learning, change teaching, and/or improve performance.

Despite the importance of consequences validity evidence, it is rarely reported in medical and surgical education [13, 14]. When it is reported, consequences are often mistaken as setting pass/fail score benchmarks [6, 15], without evaluating the impacts of these cutoffs on trainees, programs, or other stakeholders. Ideally, consequences evidence should consist of multiple dimensions including impacts of not only the scores but also of the assessment activity as a whole [13]. Examining these potential consequences of WBAs can provide essential evidence for how these assessments can be used for formative feedback. Therefore, the purpose of our study was to collect and evaluate consequences validity evidence for a WBA implemented for operative performance assessment. Specifically, we sought to identify the perceived impacts of the WBA on faculty and resident operative teaching, learning, and feedback practices.

Methods

We conducted a qualitative study that occurred in the context of a 6-month pilot of a surgical WBA, the System for Improving and Measuring Procedural Learning in the OR (SIMPL OR) in the Department of Otolaryngology—Head and Neck Surgery at a tertiary teaching institution. We approached the study from a constructionist point-of-view, recognizing reality as socially constructed and anticipating diverse perspectives on the experience of using the WBA [16]. Correspondingly, we used semi-structured interviews of residents and faculty as our primary method of data collection and designed interview questions to explore various perceived impacts of the WBA on different residents and faculty. Data analysis then sought to represent the range of participants' experiences.

Participants

All 21 resident physicians and 25 of 34 total clinical faculty in the department agreed to participate in the WBA pilot. Participants first completed rater training in January and February 2020 and began using the assessment in March 2020; however, the formal pilot period was delayed until July to December 2020 after disruptions from the COVID-19 pandemic. Beginning in October 2020, all pilot participants were invited via email to complete a semi-structured, one-on-one interview over Zoom (Zoom Video Communications, Inc. San Jose, CA), which is further described below. 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 (IRB #18–25337).

The WBA—SIMPL OR

SIMPL OR (hereafter referred to as SIMPL) is a smartphone-based assessment tool designed to improve the frequency and timeliness of operative feedback. It is intended to provide information regarding a resident's operative performance in two ways. First, SIMPL collects "real-time" formative ratings from both faculty and residents of resident autonomy and performance after surgical procedures. Faculty can also record dictated feedback for the resident to listen to after the surgery. Second, SIMPL creates a repository of longitudinal data so that various stakeholders, including residents, faculty, and programs, can track resident autonomy and performance in operative procedures over time for summative purposes. While both elements are beneficial, the developers believe that the most useful aspect of SIMPL is its

Table 1 Key interview questions for faculty and residents

1. How has SIMPL impacted the way in which you provide/receive feedback?
2. How has SIMPL impacted the way you think about a resident's/your own operative performance?
3. How has SIMPL impacted your teaching/learning?
4. How do you use the results or scores from the rating questions?

ability to function as a formative assessment to facilitate operative feedback, stimulate conversations, guide resident learning, and enhance faculty teaching [17–19], which is the focus of this study.

A SIMPL assessment consists of three questions: (1) a rating of resident operative autonomy using the 4 Zwisch levels from Show & Tell to Supervision Only [1, 18]; (2) a global rating of resident operative performance on a 5-point set of ordered categories from unprepared to exceptional; and (3) a rating of the difficulty of the procedure as easiest one-third, average, or hardest one-third [17]. Both residents and faculty answer all three questions for a specified case. For faculty, the questions are followed by a section where they can dictate recorded feedback to the resident. The ratings and faculty recordings are available for the resident to review immediately after submission, but only the program director and the individual residents have access to resident self-ratings. Either party can begin the assessment process, though during this pilot, residents were encouraged to initiate the request. The other party is then notified to complete a paired assessment within 72 h.

Framework for consequences validity evidence

Although the *Standards* describes three broad types of consequences, it does not specify how to collect or evaluate consequences evidence [10]. In this study, we adapted a framework described by Cook and Lineberry that organizes consequences evidence into several dimensions: (1) impact of the assessment results (scores), (2) impact of the assessment activity itself (independent of scores) and (3) the impact of classifications derived from scores, if present [13]. All three dimensions can include consequences that are intended/unintended or beneficial/harmful. To fit the characteristics of our assessment, we focused on consequences related to the assessment activity, the scores (i.e. the ratings to the three questions), and the dictated feedback, followed by unintended consequences. As our study team did not develop the assessment, we defined unintended consequences based on whether the consequence described by the participants aligned with goals of formative assessment.

Interviews

The first author (NZ) conducted all interviews. The interview first focused on the participants' experience with SIMPL, including usage patterns and understanding of the intentions for use to check if user perceptions were consistent with formative purposes. Questions then explored various forms of consequences (impact of assessment use, impact of scores, impact of dictated feedback), probing participant perceptions of feedback and their teaching or learning practices after the introduction of SIMPL. The interview protocols were piloted with two individuals (one attending and one resident) three months after the initial launch, and the questions were revised for clarity as needed. The pilot interviews were also included in the final analysis as they yielded data of similar quality to the regular interviews. Table 1 lists key interview questions.

Data analysis and reporting

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 interview transcripts, applying theoretically informed codes and also developing new codes as needed [20]. 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 consequences evidence using Cook and Lineberry's framework as a guide [13]. Impacts of the assessment activity were related to the process of using SIMPL, such as requesting and completing assessments. Impacts of the scores were related to the ratings from the three questions, including how faculty and residents did or did not use the ratings for teaching or learning. Since SIMPL also includes the option for dictated feedback, the consequences of this verbal component were also coded and analyzed.

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 assessment consequences. Finally, the first author compared and contrasted data in and among categories to understand relationships and determine broader themes. The researchers used the Standards for Reporting Qualitative Research to write study methods and results [21].

Reflexivity

The study team consisted of three individuals, each with training in qualitative analysis but from diverse professional backgrounds. The first author is an otolaryngologist who trained at the institution where the pilot was performed. Her stance affords her close insight into the culture and challenges of surgical training both as a whole and within the specific program. She is known to the participants, which can impact the way they share their experiences during the interviews. The other researchers included a geriatrics physician and a PhD-trained health professions education scholar. As non-surgeons, these individuals brought fewer assumptions to data interpretation yet still understood the nuances of medicine. The research team met frequently to discuss any differing interpretations and how their backgrounds shaped their understanding of the data.

Results

WBA use

A total of 267 SIMPL assessments were generated during the formal pilot period from July 1 and December 31, 2020. 212/267 (79.4%) assessments were paired, with both a faculty and resident submission for the same procedure, for a total of 106 assessment pairs. Of those who agreed to participate in the pilot, 14/21 (66.7%) residents and 12/25 (48%) faculty completed at least one SIMPL assessment. Usage declined over time, peaking in August 2020 with 38 completed assessment pairs and dropping to only 2 assessment pairs in November and December 2020. During the pilot, residents completed an average of 13.5 assessments, with a range of 1 to 47, and faculty completed an average of 9.9 assessments, with a range of 1 to 26.

Interview participants

Interviews from 8 residents and 9 faculty who used SIMPL during the pilot period were included in this study. Participant demographics are summarized in Table 2. Residents from all five post-graduate training years and faculty from a breadth of subspecialities participated. Faculty self-reported time in practice ranged from 5 to 31 years with a mean of 12.7 years.

Perceived purpose of SIMPL

Before exploring the consequences of SIMPL use, it was important to check if participants perceived SIMPL as serving formative, rather than summative, purposes and what they identified as desired outcomes of SIMPL use.

Table 2 Participant demographics

Residents ($n=8$)	n (%)	Faculty ($n=9$)	n (%)
Gender		Gender	
Men	1 (12.5)	Men	6 (66.7)
Women	7 (87.5)	Women	3 (33.3)
Year in training		Academic rank	
PGY1	1 (12.5)	Assistant professor	2 (22.2)
PGY2	2 (25.0)	Associate professor	4 (44.4)
PGY3	1 (12.5)	Full professor	3 (33.3)
PGY4	1 (12.5)	Subspecialty	
PGY5	3 (37.5)	Rhinology	3 (33.3)
		Pediatrics	1 (11.1)
		Laryngology	1 (11.1)
		Head and neck oncology	2 (22.2)
		Neurotology	1 (11.1)
		General/Sleep	1 (11.1)

PGY post-graduate year

Table 3 Representative quotes describing perceived purpose of SIMPL

Purpose	Examples
Increase frequency and ease of feedback	"The purpose is at least to give more specific feedback to certain cases… actually have some feedback instead of just numbers." (F2) "…providing a framework to create an environment for feedback…or way to give feedback easily." (R7)
Improve process and culture of feedback	"The understanding I had was that we were going to be able to provide feedback for residents that was easy…because residents were requesting feedback, and this was a way for them to do it without having to like directly ask you for feedback, which might be uncomfortable." (F7) "…an easy, simple way to solicit feedback from an attending…so that trainees don't feel awkward constantly asking attendings for feedback." (R1)

As a whole, faculty (*F*) and residents (*R*) viewed SIMPL as a form of WBA that could improve the frequency, timeliness, and ease of operative feedback, suggesting they viewed the purpose as formative and desired consequences as improving the process and culture of feedback (Table 3).

Consequences

We found multiple impacts when examining consequences evidence. These impacts are organized in Table 4 and described below.

Table 4 Overview of themes for consequences evidence for the use of SIMPL as a formative assessment

Consequences Evidence Dimension and Definition	Themes
Impact of Assessment Activity *Consequences related to the process of using SIMPL, such as requesting and completing assessments*	Greater emphasis on feedback Support for a postoperative feedback routine Facilitation of case-specific reflection
Impact of Scores and Score Use *Consequences related to the ratings and their use for teaching and learning*	Calibrating resident case perceptions Benchmarking performance to an external standard
Impact of Dictated Feedback *Consequences related to the recorded dictated feedback*	Providing context for ratings Facilitating review of dictated feedback Prompting faculty for deliberate feedback
Unintended Impacts *Consequences that limited formative use of the assessment*	Emotional discomfort during the assessment process Increased feedback frequency but not diversity or quality Inconsistent support for feedback conversations Limited next steps for teaching or learning

Impact of assessment activity

Faculty and residents identified impacts of SIMPL use that were primarily related to perceptions of feedback practices, including: (1) greater emphasis on feedback, (2) support for a postoperative feedback routine, and (3) facilitation of case-specific reflection.

Greater emphasis on feedback The presence of the assessment helped bring feedback to the forefront for residents and faculty. This impact also extended beyond the app, where "…even if it [feedback] wasn't through the app… there was an overall benefit of just bringing feedback into the light." (R2)

Support for a postoperative feedback routine Both faculty and residents described that the assessment helped create a formal structure and expectation for feedback. For example, faculty noted that receiving an assessment request was a helpful indication that the resident desired feedback, which made it easier to provide feedback:

> "…it signals to me that the resident wants feedback…the very act of the resident asking, 'Can you complete a SIMPL thing?' And then me getting the SIMPL notification and then answering those three questions, provides an important routine and prompt to provide this feedback." (F1)

Participants also described that this type of postoperative feedback was distinct from other types of feedback, such as "in-the-moment" intraoperative feedback or rotation evaluations: "It's a different type of feedback, because it's not in the moment, it is a little bit distanced from the operating room…it's a bigger picture." (R7)

Facilitation of case-specific reflection In addition to impacts on feedback, both faculty and residents reported an increased awareness of reflective practices after the case that could support learning. One resident explained, "I think the value is that it forces you to kind of review how the case went again, in my head. And then thinking through the case again adds to the learning of the case." (R7) Similarly, faculty felt that "I actually do gain a benefit. It forces me to stop and reflect on the resident's actions and progress and progression." (F6)

Impact of scores and score use

Only residents discussed impacts of the scores from the three questions. These impacts were: (1) calibration of case perceptions and (2) benchmarking performance to an external standard.

Calibration of case perceptions For residents, the scores allowed them to adjust their perceptions of the surgical case to those of the faculty, particularly "because sometimes there's a discrepancy between how you answered it and how they answer it." (R3) In addition, the different questions, such as autonomy and case difficulty, interacted to influence resident impressions:

> "The first [question], the amount of help that you got, helps calibrate your impression of the case…quantifying how much help you got may be different than what your gut instinct is…I've had cases where…the reason why it was easy was because there was a lot of active help and a lot of show-and-tell… Versus there are cases where…it was hard…because the attending was letting me struggle." (R4)

Benchmarking performance to an external standard Residents reported looking to the scores as feedback to where they were along their training continuum. In particular, the term 'practice-ready' served as an external benchmark for operative performance:

> "…hopefully, by the time I graduate, I have practice-ready performance on every case that I did with an attending. The other levels are just hopefully showing you that you're progressing a little bit. But what I'm looking for is I want to see that the attending thinks that I am ready to do this safely by myself." (R2)

Impact of dictated feedback

Faculty and residents perceived that the recorded dictations supported feedback in the following ways: (1) providing context for ratings, (2) facilitating review of dictated feedback, and (3) prompting faculty for deliberate feedback.

Providing context for ratings Residents described that the dictated feedback offered additional explanation for the scores, as an attending's reasoning for selecting a particular score was not always obvious:

> "I think the recordings are helpful, because... [if an attending] gives me an active help, and then the recording says, 'You did a great job today. You didn't seem like you needed much guidance...' Well, that's the opposite of what you just rated me. But I see what he's probably getting at as far as active help…" (R2)

Facilitating review of dictated feedback The saved dictations created a unique repository of recordings that allowed the residents to review helpful feedback:

> "[The faculty] often like to talk about why it was difficult. I think those are instances where I listen more than once to try to figure out or absorb what is it that could have been done differently, if they mention it…" (R4)

Prompting faculty for deliberate feedback For some faculty, the presence of the recordings helped them be more intentional about providing structured feedback: "I think of something to say instead of 'you did a good job.' So, I think because you're recording it… One thing to work on, one thing to improve on, one thing that was done very well. I think that's the easiest way to structure a lot of this." (F2)

Unintended impacts

We found unintended impacts in all the above domains that limited the formative use of the assessment. These impacts included: (1) emotional discomfort during the assessment process, (2) increased feedback frequency but not diversity or quality, (3) inconsistent support for feedback conversations, and (4) limited next steps in teaching and learning.

Emotional discomfort during the assessment process Both residents and faculty described instances of emotional discomfort in the assessment process. For residents, discomfort arose primarily when they perceived faculty were not interested in using the assessment. As one resident explained, "Dr. [attending] doesn't use it very often'… It's hard to tell them that… this brings sort of like an additional level to debriefing the case, and just for our learning purposes, this has a little bit more meaning." (R4) For faculty, the discomfort was primarily related to the recordings, which was a combination of personal preferences and uncertainty about whether the verbal feedback was effective:

> "I find [the recording] to be sort of the least useful way of providing feedback and also the way that makes me most uncomfortable and anxious… But I also…I don't feel like it's a great forum for them to listen to and generate questions and come back to me about it so that we can have a conversation." (F1)

Increased feedback frequency but not diversity or quality Although we found some evidence to support the perception that SIMPL was increasing the frequency of feedback in both groups, residents did not necessarily feel that a greater variety of attendings were providing feedback. As one resident described, "It objectively has [increased feedback] just because I have these, and I didn't before… And I think that maybe the attendings that participate are already the ones who are going to give us feedback anyway…" (R5) Similarly, faculty felt that while SIMPL helped increase the frequency of their feedback, it did not impact the quality of their feedback or their feedback practices.

Inadequate support for feedback conversations There were few perceptions that the assessment process, scores, or recorded dictations helped support feedback conversations. Only one participant recalled a brief follow-up conversation about the scores; most did not describe any dialogue. Some residents described instances where the presence of the assessment facilitated feedback outside of the app, such as in the setting of faculty forgetting to provide a dictation or expressing a preference to give feedback outside of SIMPL. In addition, several faculty felt that the dictations allowed them to feel as though they were speaking to the residents in

real-time; however, that feeling did not necessarily translate to increased face-to-face conversations about the feedback:

> "I don't think anyone the last few months have been like, 'What did you mean when you said that on that dictation?' There's been no clarification. So, if that did start happening that would mean that that was really encouraging a conversation' That doesn't mean it doesn't…it just means I can't tell what it's done." (F8)

Some residents felt the feedback was limited by lack of dialogue: "[It] is much more helpful to have it [feedback] be in-person in a conversation as opposed to having an audio recording where they said, 'You did a good job. This was a difficult case. You used the laser really well.' (R6)

Limited next steps for teaching or learning We did not find any evidence to suggest that faculty used the scores to guide their teaching decisions for future procedures. However, the autonomy ratings did lead one faculty member to briefly alter their intraoperative teaching behaviors to try to give residents the 'best' possible score:

> "When I first started to use it, I wanted to try to give residents a higher 'passive help' as opposed to 'active help' grading. So, I was consciously trying to give less feedback during the surgery, and then… I decided that I didn't want to restrict what I thought would be helpful feedback, based on a grading system." (F6)

Furthermore, despite the previously described impacts for case calibration and benchmarking, residents did not perceive the scores to be particularly useful as feedback to inform learning without additional context: "I notice they answer these [rating questions], and then they don't really give reasons why. So, I think they can say 'intermediate performance,' but…why was it intermediate?" (R3)

Discussion

The purpose of this study was to examine the consequences validity evidence for a WBA implemented as a formative assessment of operative performance, focusing on the primary purpose for formative assessment to support feedback and learning [12]. The results revealed that during a short-term pilot of SIMPL, both faculty and residents identified multiple impacts of the assessment process and scores on feedback perceptions and practices, including the development of a feedback routine, filling a need for postoperative feedback, increasing the perception of feedback frequency, and promoting case reflection. Residents also felt that the scores helped calibrate their perceptions of the case and that the dictated feedback could provide helpful context for the scores as well as be reviewed to enhance learning.

While these impacts appeared consistent with the developers' claim that using SIMPL could increase the frequency of operative performance assessments and provide natural language feedback [17], we also uncovered additional impacts that limited the extent to which the WBA could be used as a formative assessment to ultimately improve teaching or learning. First, there was the perception from the residents that the group of faculty providing feedback were those who already made feedback a priority, so the addition of the WBA did not necessarily improve the amount and diversity of the feedback the residents received. Second, parts of the assessment process led to emotional discomfort for both residents and faculty that may impact its use. We were also unable to provide much evidence to support the developers' claim that SIMPL's use may stimulate face-to-face feedback conversations [17], though it is possible these impacts may have occurred if use had been greater. Finally, we found that the scores had limited impacts on guiding teaching or learning. The scores could potentially have the opposite effect if faculty decrease their intraoperative teaching behaviors to try to give residents higher autonomy ratings. Although this was not a sustained practice in our study, we felt this result was important to highlight, as it underscores the potential for unique findings when examining consequences validity evidence.

The results of this study must be taken in the context of how the assessment was implemented as well as in the setting of declining use over time. During the pilot period, there was an unplanned decline in SIMPL use. Our goal was to evaluate consequences based on natural use, so we did not add extra external incentives to increase use. Previous work by Eaton et al. highlighted structural barriers to use, including forgetting to create or respond to assessment requests, being too busy, as well as perceived issues related to added value of the assessment [22]. There may also be a negative cycle in which low perceived value perpetuates users' forgetfulness. This value issue may be partly because any type of feedback given through SIMPL is inherently unidirectional from teacher to learner, which is inconsistent with how current literature characterizes optimal feedback in the form of bidirectional conversations [23, 24].

While the developers hoped that SIMPL could stimulate conversations, that intent was not realized in our study. Although feedback was traditionally viewed as unidirectional, over the past decade, research has emphasized the complex interpersonal nature of feedback. Rather than a one-way transmission of information, present recommendations are to view feedback interactions as an educational alliance between teachers and learners with a strong learner involvement in the process [23, 24]. Therefore, it is unsurprising that implementing SIMPL without faculty development and reinforcement of a new approach to feedback did not alter perceived feedback behaviors. In addition, because

the scores did not provide meaningful information on how to improve, the practical implications of the ratings themselves were limited. This issue is consistent with our group's prior work examining response processes for rating decisions, which revealed that receiving a set of scores without understanding the rater's cognitive processes or the rating context is likely insufficient to support learning [25].

Based on these findings, we suggest several ways programs can improve the use of WBAs such as SIMPL for formative purposes and maximize gains for teaching and learning. First, prior work has shown that WBAs are often perceived as both formative and summative [26]. Therefore, the goals for formative assessment use should not only be made explicit to faculty and trainees at the beginning, but also be revisited throughout implementation to ensure consistency between intended and actual use. Second, faculty and residents can be encouraged to review results from prior assessments to formulate teaching and learning goals and to have follow-up conversations to align with best practices for feedback as a dialogue [23, 24]. The assessment scores and dictations should be viewed as an invitation for conversation, rather than an endpoint. Furthermore, other solutions may be needed to increase faculty and resident participation. As evidenced by the steep decline in usage in our study, simply implementing a new assessment program is not sufficient for engagement.

Limitations to this study include the scope and duration of the implementation and nature of the sample. The assessment was only implemented in a single department within a single institution for a short-term pilot. It is likely that the consequences will evolve over the longer term, and some impacts that may initially seem beneficial may lead to more negative effects over time. In addition, the assessment was only used for formative purposes, so it is unclear how the impacts would change if the program also used results for summative decisions, which is likely to occur in many other programs. Finally, the interview descriptions of changes to feedback practices are limited to only what the participants can recall. Further study with longitudinal pre/post data or substantial fieldwork observations of feedback practices could more completely investigate assessment consequences.

Conclusions

The validity evidence gathered in this study suggests an operative WBA can be used for formative purposes to improve perceptions of feedback. However, unintended consequences and implementation challenges such as low use limited ultimate gains in teaching and learning; these issues can be addressed to achieve desired impacts. User perspectives can add important elements to consequences validity evidence and should be considered when designing and implementing assessments to better understand the impacts on users and whether these impacts align with intended purposes. Further research is needed to evaluate WBAs in different implementation settings to continue to refine their design and implementation for optimal impact as a formative assessment.

Author contributions All authors contributed to the study conception and design. Data collection was performed by NWZ, MD, MAEd. Material preparation and data analysis were performed by NWZ, MD, MAEd; LMH, MD, MAEd; and BCOB, PhD. The first draft of the manuscript was written by NWZ, MD, MAEd, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Reporting guideline This study was reported in accordance with the Standards for Reporting Qualitative Research (SRQR)

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

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethics approval This study was performed in line with the principles of the Declaration of Helsinki. The University of California, San Francisco Institutional Review Board approved this study as exempt (IRB #18–25337).

References

1. DaRosa DA, Zwischenberger JB, Meyerson SL, George BC, Teitelbaum EN, Soper NJ, et al. A theory-based model for teaching and assessing residents in the operating room. J Surg Educ. 2013;70:24–30.
2. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84:273–8.

3. Trehan A, Barnett-Vanes A, Carty MJ, McCulloch P, Maruthappu M. The impact of feedback of intraoperative technical performance in surgery: a systematic review. BMJ Open. 2015;5:e006759.
4. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide no. 31. Med Teach. 2007;29:855–71.
5. Pelgrim EAM, Kramer AWM, Mokkink HGA, van der Vleuten CPM. The process of feedback in workplace-based assessment: organisation, delivery, continuity. Med Educ. 2012;46:604–12.
6. 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;246:342–78.
7. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. BMJ. 2010;341:c5064.
8. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA. 2009;302:1316–26.
9. Martin L, Blissett S, Johnston B, Tsang M, Gauthier S, Ahmed Z, et al. How workplace-based assessments guide learning in postgraduate education: a scoping review. Med Educ. 2023;57:394–405.
10. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, D.C: American Educational Research Association; 2014.
11. Govaerts M. Workplace-based assessment and assessment for learning: threats to validity. J Grad Med Educ. 2015;7:265–7.
12. Hopster-den Otter D, Wools S, Eggen TJHM, Veldkamp BP. A general framework for the validation of embedded formative assessment. J Educ Meas. 2019;56:715–32.
13. Cook DA, Lineberry M. Consequences validity evidence: evaluating the impact of educational assessments. Acad Med. 2016;91:785–95.
14. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(166):e7-16.
15. Rasmussen NK, Carlsen JF, Olsen BH, Stærk D, Lambine T-L, Henriksen B, et al. Ensuring competence in ultrasound-guided procedures—a validity study of a newly developed assessment tool. Eur Radiol. 2022;32:4954–66.
16. Rees CE, Crampton PES, Monrouxe LV. Re-visioning academic medicine through a constructionist lens. Acad Med. 2020;95:846.
17. Bohnen JD, George BC, Williams RG, Schuller MC, DaRosa DA, Torbeck L, 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:e118–30.
18. George BC, Teitelbaum EN, Meyerson SL, Schuller MC, DaRosa DA, Petrusa ER, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. J Surg Educ. 2014;71:e90-96.
19. George BC, Bohnen JD, Schuller MC, Fryer JP. Using smartphones for trainee performance assessment: a SIMPL case study. Surgery. 2020;167:903–6.
20. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res. 2005;15:1277–88.
21. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Acad Med. 2014;89:1245–51.
22. Eaton M, Scully R, Schuller M, Yang A, Smink D, Williams RG, et al. Value and barriers to use of the SIMPL tool for resident feedback. J Surg Educ. 2019;76:620–7.
23. Ramani S, Könings KD, Ginsburg S, van der Vleuten CPM. Twelve tips to promote a feedback culture with a growth mind-set: swinging the feedback pendulum from recipes to relationships. Med Teach. 2019;41:625–31.
24. Ramani S, Könings KD, Ginsburg S, van der Vleuten CPM. Feedback redefined: principles and practice. J Gen Intern Med. 2019;34:744–9.
25. Zhao NW, Haddock LM, O'Brien BC. Are you thinking what i'm thinking? Exploring response process validity evidence for a workplace-based assessment for operative feedback. J Surg Educ. 2022;79:475–84.
26. Mughal Z, Patel S, Gupta KK, Metcalfe C, Beech T, Jennings C. Evaluating the perceptions of workplace-based assessments in surgical training: a systematic review. Ann R Coll Surg Engl. 2023;105:507–12.

Authors and Affiliations

Nina W. Zhao[1,2] · Lindsey M. Haddock[3,4] · Bridget C. O'Brien[3]

✉ Nina W. Zhao
nina.zhao.md@gmail.com

[1] Department of Otolaryngology–Head and Neck Surgery, University of California, 2233 Post Street, 3rd Floor, San Francisco, CA 94115, USA

[2] Present Address: Department of Otolaryngology–Head and Neck Surgery, University Hospitals Cleveland Medical Center, 11100 Euclid Avenue, Lakeside Suite 4500, Cleveland, OH 44106, USA

[3] Department of Medicine, University of California, 521 Parnassus Avenue, 2nd Floor, San Francisco, CA 94143, USA

[4] Present Address: Section of Geriatric Medicine, Division of Primary Care and Population Health, Department of Medicine, Stanford University School of Medicine, 211 Quarry Road, Suite 402, Palo Alto, CA 94304, USA