Surgical Endoscopy
https://doi.org/10.1007/s00464-022-09466-6

2022 SAGES ORAL

Validity of video-based general and procedure-specific self-assessment tools for surgical trainees in laparoscopic cholecystectomy

Saba Balvardi[1,2] · Koorosh Semsar-Kazerooni[2] · Pepa Kaneva[2] · Carmen Mueller[1,2] · Melina Vassiliou[1,2] · Mohammed Al Mahroos[1] · Julio F. Fiore Jr.[1,2] · Kevin Schwartzman[3] · Liane S. Feldman[1,2,4]

Received: 11 March 2022 / Accepted: 10 July 2022

Abstract

Introduction Self-review of recorded surgical procedures offers new opportunities for trainees to extend technical learning outside the operating-room. Valid tools for self-assessment are required prior to evaluating the effectiveness of video-review in enhancing technical learning. Therefore, we aimed to contribute evidence regarding the validity of intraoperative performance assessment tools for video-based self-assessment by general surgery trainees when performing laparoscopic cholecystectomies.

Methods and procedures Using a web-based platform, general surgery trainees in a university-based residency program submitted recorded laparoscopic cholecystectomy procedures where they acted as the supervised primary surgeon. Attending surgeons measured operative performance at the time of surgery using general and procedure-specific assessment tools [Global Operative Assessment of Laparoscopic Skills (GOALS) and Operative Performance Rating System (OPRS), respectively] and entrustability level (O-SCORE). Trainees self-evaluated their performance from video-review using the same instruments. The validity of GOALS and OPRS for trainee self-assessment was investigated by testing the hypotheses that self-assessment scores correlate with (H1) expert assessment scores, (H2) O-SCORE, and (H3) procedure time and that (H4) self-assessment based on these instruments differentiates junior [postgraduate year (PGY) 1–3] and senior trainees (PGY 4–5), as well as (H5)simple [Visual Analogue Scale (VAS) ≤ 4] versus complex cases (VAS > 4). All hypotheses were based on previous literature, defined a priori, and were tested according to the COSMIN consensus on measurement properties.

Results A total of 35 videos were submitted (45% female and 45% senior trainees) and self-assessed. Our data supported 2 out of 5 hypotheses (H1 and H4) for GOALS and 3 out of 5 hypotheses (H1, H4 and H5) for OPRS, for trainee self-assessment.

Conclusions OPRS, a procedure-specific assessment tool, was better able to differentiate between groups expected to have different levels of intraoperative performance, compared to GOALS, a general assessment tool. Given the interest in video-based learning, there is a need to further develop valid procedure-specific tools to support video-based self-assessment by trainees in a range of procedures.

Keywords Video-based assessment · Self-assessment · Validity · Intraoperative assessment tool

Evidence suggests that surgical technique and skills directly influence safety and patient outcomes [1, 2]. A recent study has shown that almost one third of surgical graduates do not feel confident in their ability to perform certain procedures

✉ Liane S. Feldman
liane.feldman@mcgill.ca

[1] Department of Surgery, McGill University, Montreal, QC, Canada

[2] Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, Montreal, QC, Canada

[3] Respiratory Division, Department of Medicine, McGill University and McGill International Tuberculosis Centre, Research Institute of the McGill University Health Centre, Montreal, QC, Canada

[4] Department of Surgery, McGill University Health Centre, 1650 Cedar Ave, D6-156, Montreal, QC H3G 1A4, Canada

Published online: 03 August 2022

independently [3]. Modern challenges to surgical education include restriction of working hours as well as the COVID-19 pandemic that reduced trainees' exposure to elective surgery through mandated cessation of non-essential procedures [4, 5]. Hence, extension of technical learning outside the operating-room has become crucial [6]. Video-based assessment (VBA) of recorded operative procedures provides a new opportunity to measure surgeon performance while minimizing barriers related to direct in-theater evaluations. While video-assisted structured feedback by expert surgeons significantly improves laparoscopic skill acquisition in surgical trainees [7–9], this method is resource intensive and may have limited feasibility outside research settings. Accordingly, there is growing interest in the potential role of guided self-assessment of videorecorded surgical procedures to address this procedural training gap [10].

Self-assessment is an integral part of lifelong medical experiential learning. Evidence supports the utility of guided self-assessment to improve performance in non-medical fields such as sports and music [11]. However, systematic reviews report mixed results regarding the accuracy of trainee self-assessment [12–14]. These shortcomings can be mitigated using video recordings and implementing guided self-assessment strategies based on more robust intraoperative performance standards [12, 15]. Video-based tools with evidence supporting their use for self-assessment are required before the value of video-based self-assessment in enhancing surgical skill acquisition can be accurately investigated [16]. Thus the aim of this study was to contribute evidence regarding the validity of intraoperative assessment tools when used for formative video-based self-assessment by general surgery trainees performing laparoscopic cholecystectomies.

Materials and methods

Participants and settings

This study is a single-centered prospective cohort study that took place at the adult hospital sites of the McGill University Health Center. This was a sub-study of a recently completed randomized controlled trial with data collected from August 2020 to August 2021 (effect of video-based guided self-reflection on intraoperative skills: a pilot randomized controlled trial; ClinicalTrials.gov Identifier NCT04643314). This study was approved by our Institutional Review Board (MUHC Ethics Approval ID 2020-6348). Inclusion criteria were: (1) postgraduate year (PGY) 2–5 trainees, (2) rotating through a General Surgery Clinical Teaching Unit, (3) performing elective or emergency (patients admitted through the emergency department) laparoscopic cholecystectomies (4) and performing more than 70% of the procedure. PGY 1 trainees and procedures where there was significant (more than 30%) supervising surgeon take over were excluded.

Measures and procedures

Demographic data from the trainees (i.e., age, gender, PGY, handedness, and number of previous laparoscopic and laparoscopic cholecystectomy procedures) and the characteristics of the operative procedures (diagnosis, urgency, and procedure time) were collected. Junior trainees were defined as PGY 2–3, and senior trainees were PGY 4–5. Total duration of the operation was defined as the time from skin incision to skin closure. Time to dissect the triangle of Calot was defined as the time from completion of adhesiolysis to clipping the first structure in the triangle of Calot. Duration of dissection of the gallbladder bed was defined as the time from division of the last structure in the hepatocystic triangle until detachment of the gallbladder from the gallbladder bed. In case of rescue techniques such as antegrade cholecystectomy or subtotal cholecystectomy, only the total procedure time was collected. The operative times were measured based on these definitions by one of the authors blinded to the operator and operative case characteristics.

Each resident was given a data storage device (USB key) to record elective and emergency laparoscopic cholecystectomy procedures in which they acted as the supervised primary surgeon for a significant portion of the operation (more than 70% of the whole procedure for senior trainees or more than 70% of triangle of Calot dissection and/or gallbladder wall dissection for junior trainees). These videos were uploaded to a secure web-accessible platform (Theaor.io) and any identifying features were removed by the platform. In addition to storage and facilitation of access to surgical videos, Theator segments the procedure into steps to enable more targeted review of different parts of the operation (i.e., preparation, triangle of Calot dissection, division of cystic structures, gallbladder separation, gallbladder packaging, extraction) and evaluates whether the critical view of safety was obtained. Trainees met with a member of the study team not involved in clinical supervision to receive coaching on the nature and the use of the intraoperative assessment tools and undergo rater training including demonstration of sample videos for low and high scores for each assessment items. The trainees were then asked to practice using the scales with calibrating videos in the same session. Subsequently, trainees in the intervention group were asked to review their own operating-room recordings and to assess themselves, entering their self-assessment scores directly into the Theator platform as shown in Supplementary Material 1.

Operative performance was assessed using two measures selected from tools identified by a systematic review of performance assessment tools for laparoscopic cholecystectomy [17]: Global Operative Assessment of Laparoscopic Skills

(GOALS) [18], a global rating tool for laparoscopic skills, and Operative Performance Rating System (OPRS) [19], a procedure-specific assessment tool. Both GOALS and OPRS have been supported by validity evidence for use in direct intraoperative and video-based evaluation of trainees by attending surgeons [17]. In the present study, 16 attending Acute Care Surgery and Minimally Invasive Surgery surgeons participated in our study. The attending surgeon underwent the same rater training as the participants. The attending surgeon received an email with a link to provide their GOALS and OPRS assessments immediately after the procedure with maximum allotted time of 72 h. They also completed a post-procedural questionnaire to report the degree of involvement of the trainee in the procedure (0–100%), case difficulty [using a Visual Analogue Scale (VAS: 1–10)] and overall trainee entrustability using the Ottawa Surgical Competency Operating-Room Evaluation (O-SCORE) [20].

GOALS is a general intraoperative performance assessment tool for laparoscopic procedures consisting of five items, each scored using a 5-point Likert scale where '1' represents the lowest level of performance and '5' is considered ideal performance. The total possible score ranges between 5 and 25 [18]. The items evaluate depth perception, bimanual dexterity, efficiency, tissue handling and autonomy (Supplementary Material 2) [18]. There is evidence for the validity of GOALS in direct intraoperative and video-based evaluation by attending surgeons [17]. OPRS is a procedure-specific 10-item intraoperative assessment tool [20]. A rating scale of 1–5 is used to evaluate each item with a rating of four or higher indicating technical proficiency and operative independence [17]. The final score is the mean score of the 10 items (Supplementary Material 3) [20]. OPRS has been recommended for use in the setting of direct observation, but it can also be used in assessment of recorded procedures [17]. The O-SCORE is a valid and reliable intraoperative assessment of operative competence using a 5-level scale. The expert clinician ranks the trainee's independence from 1 = "I had to do it" to 5 = "I did not need to be there" (Supplementary Material 4) [23]. This scale is only designed for direct observation and is not suited for video assessment [23]. This scale is included in trainee competency assessment by the Royal College of Physicians and Surgeons of Canada's competency-based medical education framework.

Validity assessment

The validity of GOALS and OPRS tool as formative video-based self-assessment tools by general surgery trainees in laparoscopic cholecystectomy was evaluated based on COSMIN best practice guidelines for examining psychometric properties [21]. Based on previous literature, we hypothesized a priori that if GOALS and OPRS are valid video-based self-assessment tools for general surgery trainees in laparoscopic cholecystectomy, trainee self-assessment scores will correlate with (1) expert assessment scores [22], (2) O-SCORE Entrustability Scale [20, 23], and (3) procedure time [24] and that (4) self-assessment based on these instruments can differentiate junior (PGY 1–3) from senior trainees (PGY 4–5) [25], as well as (5) simple (VAS ≤ 4) versus complex cases (VAS > 4) [20].

Statistical analysis

A sample size of 29 submissions was expected to be sufficient to detect moderate correlations, i.e., $r = 0.5$ (as defined by COSMIN best practice guidelines) [21] with an $\alpha = 0.05$ and $\beta = 0.80$. Continuous variables were reported using mean and standard deviation or median and interquartile range, as appropriate. Categorical variables were reported using frequencies and percentages.

Guidelines recommend that hypotheses testing be based on the expected direction and magnitude of differences or correlations rather than on sample size-dependent statistics, such as p values [21]. Hypotheses 1–3 were tested using Pearson or Spearman's rank Correlation where appropriate. We expected a moderate positive correlation (coefficient 0.3 to 0.5) between attending surgeon and trainee self-assessments, and a moderate negative correlation (coefficient -0.3 to -0.5) between trainee self-assessment and procedure time. Hypotheses 4–5 were tested using multiple linear regression while adjusting for gender (for Hypotheses 4 and 5) [26], case complexity (for Hypothesis 4) and PGY level (for Hypothesis 5). We hypothesized that the magnitude of difference between groups would be equal to or greater than the minimal important difference (MID) of 2 for GOALS [18] and 0.3 for OPRS [27, 28]. These MIDs were estimated based on distribution based method with differences above one-half of the standard deviation considered clinically meaningful [29]. To reduce the risk of bias arising from missing data, we used random-forest-based imputation of missing data using the missForest R package [30]. Statistical analyses were performed using RStudio (version 1.2.1577; RStudio, Inc., Boston, MA, USA).

Results

A total of 35 intraoperative recordings of laparoscopic cholecystectomy procedures were submitted by 11 trainees. Two trainees refused self-assessment citing time constraints. The trainee median age was 30 years, 45% were female and 45% were senior trainees (≥ PGY 4). Fifty-five percent of the submitted cases were done by trainees who had been the primary operating surgeon in more than 20 laparoscopic

cholecystectomies and 89% had been the primary operating surgeon in more than 50 laparoscopic procedures (Table 1).

Out of the 35 submitted intraoperative recordings of laparoscopic cholecystectomies 11 (37%) were of patients with acute cholecystitis and 8 (23%) were of patients with biliary colic. Twelve cases out of 35 (34%) were done on an emergency basis and 17 (48%) were deemed complex (VAS > 4) by the attending supervising surgeon (Table 2). In 9 (26%) of these cases, the supervising attending surgeon took over for less than 30% of the duration of the procedure. Median length of procedure was 85 min, 20.5 min for dissection of the triangle of Calot and 10.4 min for dissection of the gallbladder bed (Table 2).

Table 3 summarizes the intraoperative attending surgeon assessment and the trainee video-based self-assessment scores for GOALS and OPRS. Median length of time to completion of the intraoperative assessments by the attending surgeon was 0 days. However, trainee median length of time to video-based self-assessment was 10 days. The attending surgeon's GOALS and OPRS total scores were higher than the trainee's self-assessments [22 vs. 18 ($p = 0.001$) and 4.5 vs. 3.7 ($p < 0.001$); respectively].

Trainees' GOALS video self-assessment scores correlated with staff surgeon GOALS assessment scores (correlation coefficient 0.47) and mean self-assessment scores differed between senior versus junior trainees (adjusted mean difference 3.53 [3.06, 3.78]). However, they did not correlate with entrustability (O-SCORE) or total procedure time. GOALS scores were also not significantly different between complex versus simple procedures (Table 4). Trainees' OPRS self-assessment scores correlated with staff surgeon assessment (correlation coefficient 0.35), differed between senior versus junior trainees (adjusted mean difference 0.51 [0.17, 0.84])

Table 1 Characteristics of trainee operator

Variables	
Number of trainees	11
Age (years)	30.0 (29.0, 31.5)
Gender	
Female	5 (45%)
Male	6 (55%)
Training level	
Junior trainees (PGY 2–3)	6 (55%)
Senior trainees (≥ PGY 4)	5 (45%)
Handedness	
Right-handed	11 (100%)
Left-handed	0 (0%)
Previous laparoscopic cholecystectomy experience	
≤ 20	5 (45%)
> 20	6 (55%)
Previous laparoscopic experience	
≤ 50	2 (18%)
> 50	9 (89%)

Data are presented as median (IQR) or *n* (%)

IQR interquartile range, *PGY* postgraduate year

Table 2 Operative case characteristics

Variables	
Number of videos, *n*	35
Diagnosis, *n* (%)	
Acute cholecystitis	13 (37%)
Biliary colic	8 (23%)
Chronic cholecystitis	4 (11%)
Choledocholithiasis	3 (9%)
Gallbladder polyp	3 (9%)
Pancreatitis	4 (11%)
Operative priority, *n* (%)	
Emergency	12 (34%)
Elective	23 (66%)
Complex procedure (VAS > 4), *n* (%)	17 (48%)
Triangle of Calot dissection done by trainee, % (mean +/− SD)	89.3% ± 26.5%
Gallbladder bed dissection done by trainee, % (mean +/− SD)	96.9% ± 9.6%
Take-over by supervising surgeon, *n* (%)	
Yes	9 (26%)
No	26 (74%)
Procedure duration-min, median (IQR)	
Total procedure time	85.0 (66.0, 115.0)
Dissection of triangle of Calot duration	20.5 (16.1, 36.9)
Dissection of gallbladder bed	10.4 (7.8, 14.8)

IQR interquartile range, *VAS* Visual Analogue Score

Table 3 Intraoperative expert assessment and video-based self-assessment

Variables	Intraoperative expert assessment Median (IQR)	Trainee self-assessment Median (IQR)	*p* value
Time to assessment (days)	0 (0–1)	10 (4–28)	NA
O-SCORE	4 (3,4)	NA	NA
GOALS	22 (19, 23)	18 (17, 20)	0.001
Depth perception	5 (5, 5)	4 (4, 4)	< 0.001
Bimanual dexterity	4 (4, 5)	4 (3, 4)	0.01
Efficiency	4 (4, 5)	3 (3, 4)	< 0.001
Tissue handling	4 (4, 5)	4 (3, 4)	< 0.001
Autonomy	4 (3.2, 5)	4 (3, 4)	0.1
OPRS	4.5 (3.7, 4.9)	3.7 (3.3, 4)	< 0.001
Incision/port placement	5 (5, 5)	4 (4, 5)	0.001
Exposure	4 (4, 5)	4 (4, 4)	0.07
Cystic duct dissection	4 (4, 5)	4 (3, 4)	0.009
Cystic artery dissection	4 (4, 5)	4 (3, 4)	0.002
Gallbladder dissection	5 (4, 5)	4 (3, 4)	< 0.001
Instrument handling	4 (4, 5)	4 (3, 4)	0.003
Respect for tissue	5 (4, 5)	4 (3, 4)	< 0.001
Time and motion	4 (4, 5)	3 (3, 4)	< 0.001
Operation flow	4 (4, 5)	3 (3, 4)	< 0.001
Overall performance rating	5 (4, 5)	4 (3, 4)	< 0.001

IQR interquartile range, *O-SCORE* Ottawa Surgical Competency Operating-Room Evaluation, *GOALS* Global Operative Assessment of Laparoscopic Skills, *OPRS* Operative Performance Rating System, *NA* not applicable

and differed between complex versus simple procedures (adjusted mean difference 0.39 [0.03 to 0.74]). However, they did not correlate with entrustability scores or procedure time (Table 4). Hypothesis 3 was further investigated by testing the correlation of the self-assessment scores with the duration of the dissection of the triangle of Calot and the dissection of the gallbladder from the liver bed separately. Both GOALS and OPRS self-assessment scores correlated with the duration of dissection of the gallbladder bed but not the triangle of Calot dissection.

There was an 11% rate of missing attending surgeon intraoperative assessment (Supplementary Material 5). However, sensitivity analysis by testing these hypotheses after imputation of missing data yielded similar findings (Supplementary Material 6).

Discussion

GOALS and OPRS are two commonly used general and procedure-specific intraoperative assessment tools in laparoscopic cholecystectomy. There is evidence for their validity as formative assessment tools for surgical trainees evaluated by attending surgeons [17]. In this study we contribute evidence regarding the validity of their use for video-based self-assessment by general surgical trainees. Of the 5 a priori hypotheses tested for validity, 2 were supported for GOALS while 3 were supported for OPRS, suggesting stronger support for the use of self-assessment tools with procedure-specific items in this context.

Trainees' GOALS self-assessment scores correlated with expert GOALS assessment scores [22] and self-assessment scores were significantly higher in senior surgical trainees (PGY 4–5) compared to junior trainees (PGY 2–3) [25]. In

Table 4 Validity hypothesis testing

Hypothesis	GOALS		OPRS	
	Coefficient (95% CI)[a]	Hypothesis confirmed	Coefficient (95% CI)[a]	Hypothesis confirmed
(1) Correlation of self-assessment with expert score	0.47	Yes	0.35	Yes
(2) Correlation of self-assessment with expert entrustability score	0.17	No	0.18	No
(3) Correlation of self-assessment with total procedure time	−0.11	No	−0.13	No
(a) Correlation with duration of TC dissection	−0.05	No	−0.06	No
(b) Correlation with duration of GB bed dissection	−0.41	Yes	−0.32	Yes
(4) Mean difference in self-assessment score for senior vs. junior trainees	3.53 (3.06, 3.78)	Yes	0.51 (0.17, 0.84)	Yes
(5) Mean difference in self-assessment score for complex vs. simple cases	−1.56 (−3.40, 0.28)	No	−0.39 (−0.74, −0.03)	Yes

GOALS Global Operative Assessment of Laparoscopic Skills, *OPRS* Operative Performance Rating System, *O-SCORE* Ottawa Surgical Competency Operating-Room Evaluation, *TC* triangle of Calot, *BG* gallbladder bed

[a]95% CI is reported for regression coefficients

contrast to previous literature demonstrating that intraoperative technical skill scores obtained by direct observation by expert surgeons correlate with entrustability score [23], procedure duration [24], and operative case complexity [20], these were not observed in our study for GOALS self-assessment scores. Trainees' OPRS self-assessment scores correlated with expert OPRS assessment scores [22] and self-assessment scores were significantly higher in senior surgical trainees (PGY 4–5) compared to junior trainees (PGY 2–3) [25] and in simple (VAS ≤ 4) compared to more complex cases (VAS > 4) [20]. However, the previously demonstrated correlation between intraoperative technical skill assessed by attending surgeons and entrustability score [23] and procedure duration [24] were not detected using OPRS self-assessment scores.

Neither OPRS nor GOALS self-assessment scores correlated with the O-SCORE evaluating entrustability, despite studies reporting correlation between expert assessment scores and O-SCORE [23]. This could be due to inherent differences between the constructs that these tools are designed to measure. O-SCORE is a tool that is designed to assess surgical competence (i.e., technical skills, cognitive skills and non-technical skills including communication and leadership) and hence readiness for independent performance of a procedure [20]. In contrast, the assessment items in OPRS and GOALS are largely directed towards technical skills performance with one or two elements assessing cognitive skills (namely elements evaluating flow of the operation or trainees' autonomy) [17]. This is supported by previous studies showing that self-assessment of cognitive tasks to be fundamentally different and less accurate than that of more objective technical tasks in trainees [12]. Furthermore, O-SCORE assessment incorporates an established external reference criterion (independent performance of a procedure as an attending surgeon) but OPRS and GOALS items are susceptible to relative scoring by trainees based on training level (e.g., "I did well as a junior resident in an emergency case"), especially in more junior trainees who lack the full range of surgical skills. This can in turn result in end-aversion bias (i.e., avoidance of low scores during self-assessment due to an incorrect external reference) [31]. Therefore, the discrepancy between our findings with previous literature that reported a correlation between O-SCORE and OPRS may be due to the lower risk of end-aversion bias and superior cognitive task assessment in expert attending assessors compared to trainee self-assessment in our study.

Similarly, neither OPRS nor GOALS self-assessment scores correlated strongly with total procedure time, in contrast to what was previously reported in the literature [24]. In our analysis, procedure length was defined a priori as the time from skin incision to skin closure. We performed a sensitivity analysis looking at the association of self-assessment score with duration of dissection of the triangle of Calot and duration of dissection of the gallbladder bed separately. We observed a significant inverse correlation between self-assessment scores and time for dissection of the gallbladder bed. We hypothesize that the lack of correlation with total operative duration can be due to the variations in operative characteristics such as difficulty obtaining intra-abdominal access, presence of intra-abdominal adhesions, gallbladder extraction, or variability in involvement of junior trainees in closure that are independent from technical skills but can affect the procedure duration. Furthermore, previous studies have also suggested a significant disagreement between surgeons regarding when the 'critical view of safety' is achieved or when dissection of the triangle of Calot can be deemed adequate [32, 33]. Therefore, the lack of correlation of the self-assessment scores and the duration of dissection of the triangle of Calot may reflect the variability in defining the endpoint of this dissection between supervising attending surgeons.

A systematic review of self-assessment of technical tasks in surgery by Zevin et al. reported mixed results regarding the accuracy of trainee self-assessment [12]. These findings have been partly attributed to methodological limitations of previous studies and to factors such as recall bias (i.e., poor recall of intraoperative events by trainees after the fact) [12]. Cognitive factors such as 'memory bias' have also been reported to affect accuracy of self-assessment. Memory bias is a defense mechanism that encourages poor recall of personal failures to decrease unhappiness and despair [34]. The use of intraoperative recording review and valid and reliable assessment tools with unambiguous behavioral anchors have been associated with improved accuracy of self-assessment [12, 35]. Furthermore, video-based self-reflection has been found to readily address factors such as recall bias and memory bias, and valid assessment tools with clear performance anchors have the potential to address the lack of accuracy and inconsistency in interpretation of items [12, 15]. Our findings corroborated these previously outlined observations as we observed that OPRS (as an assessment tool that includes procedure-specific performance anchors) had stronger evidence of validity as a self-assessment tool compared to GOALS (a general assessment tool).

However, although GOALS and OPRS self-assessment scores significantly correlated with expert scores, they were consistently lower than expert scores. This discrepancy could be due to participant characteristics such as self-confidence, level of training or trainee gender [12]. Trainees who are women and trainees with low self-confidence have been reported to more frequently underestimate their performance [12]. Furthermore, rater training is an important avenue for minimizing information bias as it improves accuracy and reliability of assessment using standardized tools [36]. In our study we used a personal session to familiarize the trainees with the assessment tools and performance anchors, and

provide them with video examples. This method is formally known as 'Performance Dimension Training' [36]. Even though this method has been shown to improve rater accuracy, raters remain susceptible to the 'drift effect' where assessment accuracy can decline with time after initial training [37]. In future studies, providing longitudinal self-assessment feedback by comparison to expert assessments (i.e., Frame-of-reference Training) may lead to more significant and sustained positive impact on self-assessment accuracy [38]. Consequently, lack of adequate rater training or the drift effect could have introduced non-differential information bias in this study [39].

The strength of our study lies in the robust methodology used for validity assessment. We followed COSMIN best practice guidelines and hypotheses were posed a priori to prevent reporting bias [21, 40]. We observed that the median time to completion of intraoperative attending assessment was 0 days, with 75% of evaluations being completed by 1 day after the procedure decreasing the chance of recall bias of direct intraoperative assessments by attending surgeons. The median time to completion of trainee self-assessments was 10 days with a larger interquartile range. The use of intraoperative recording for self-assessment reduces concern about recall bias. However, since our data came from trainees who were interested in self-assessment, an element of selection bias cannot be excluded. Another limitation of our study is that the 35 videos analyzed were submitted by 11 trainees, introducing a clustering effect between submissions by the same trainee (i.e., values for videos obtained from the same trainee have a different relationship to one another than values for videos obtained from different trainees). Lack of accounting for clustering of data through statistical methods can introduce type 1 error [41]. However, given the size of the clusters (with two to five videos submitted by a given trainee), cluster analysis is not recommended and hence it was not performed [42]. Hierarchical linear modeling (HLM) is another statistical solution to decreasing type 1 error when analysing clustered data. However, previous research has shown that this strategy in sparsely clustered data (cluster size < 5) is not recommended due to significant decrease in power [43].

In summary, our study contributes evidence supporting the validity of GOALS and OPRS for formative trainee video-based self-assessment. There was stronger support for the use of OPRS, with three of five validity hypotheses supported, suggesting a potential advantage for assessments that include procedure-specific items compared to global assessments alone. These tools and their procedure-specific performance anchors can act as a guide for more accurate introspection and therefore may enhance their educational value for procedural learning. Given the reduced operative exposure of surgical trainees [3], use of these strategies to expand skills training outside the operating-room is crucial [10]. Future research should focus on developing procedure-specific VBA tools with robust measurement properties. This is an important step that will be required to investigate whether video self-review can improve procedural learning by surgical trainees.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00464-022-09466-6.

Funding This study was supported by Fonds de la recherche en Sante du Quebec (FRSQ); Educational Grant from Theator; and Society of University Surgeons STORZ Award for Residents.

Declarations

Disclosures Saba Balvardi, Koorosh Semsar-Kazerooni, Pepa Kaneva, Carmen Mueller, Melina Vassiliou, Mohammed Al Mahroos, Julio F. Fiore Jr., Kevin Schwartzman and Liane S. Feldman have no conflict of interest or financial ties to disclose.

References

1. Birkmeyer JD, Finks JF, O'Reilly A et al (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442. https://doi.org/10.1056/NEJMsa1300625
2. Mackenzie H, Ni M, Miskovic D et al (2015) Clinical validity of consultant technical skills assessment in the English National Training Programme for Laparoscopic Colorectal Surgery. Br J Surg 102(8):991–997. https://doi.org/10.1002/bjs.9828
3. Friedell ML, VanderMeer TJ, Cheatham ML et al (2014) Perceptions of graduating general surgery chief residents: are they confident in their training? J Am Coll Surg 218(4):695–703. https://doi.org/10.1016/j.jamcollsurg.2013.12.022
4. Purdy AC, de Virgilio C, Kaji AH et al (2021) Factors associated with general surgery residents' operative experience during the COVID-19 pandemic. JAMA Surg 156(8):767–774. https://doi.org/10.1001/jamasurg.2021.1978
5. Canadian Medical Association (2021) Clearing the backlog: the cost to return wait times to pre-pandemic levels Deloitte LLP and affiliated entities. Canadian Medical Association. https://www.cma.ca/sites/default/files/pdf/Media-Releases/Deloitte-Clearing-the-Backlog.pdf. Accessed 8 June 2021
6. Hamad GG, Brown MT, Clavijo-Alvarez JA (2007) Postoperative video debriefing reduces technical errors in laparoscopic surgery. Am J Surg 194(1):110–114. https://doi.org/10.1016/j.amjsurg.2006.10.027
7. Bonrath EM, Dedy NJ, Gordon LE, Grantcharov TP (2015) Comprehensive surgical coaching enhances surgical skill in the operating room: a randomized controlled trial. Ann Surg 262(2):205–212. https://doi.org/10.1097/SLA.0000000000001214
8. Grantcharov TP, Schulze S, Kristiansen VB (2007) The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. Surg Endosc 21(12):2240–2243. https://doi.org/10.1007/s00464-007-9356-z
9. Trehan A, Barnett-Vanes A, Carty MJ, McCulloch P, Maruthappu M (2015) The impact of feedback of intraoperative technical performance in surgery: a systematic review. BMJ Open 5(6):e006759. https://doi.org/10.1136/bmjopen-2014-006759
10. Green JL, Suresh V, Bittar P, Ledbetter L, Mithani SK, Allori A (2019) The utilization of video technology in surgical education:

a systematic review. J Surg Res 235:171–180. https://doi.org/10.1016/j.jss.2018.09.015
11. Liebermann DG, Katz L, Hughes MD, Bartlett RM, McClements J, Franks IM (2002) Advances in the application of information technology to sport performance. J Sports Sci 20(10):755–769. https://doi.org/10.1080/026404102320675611
12. Zevin B (2012) Self versus external assessment for technical tasks in surgery: a narrative review. J Grad Med Educ 4(4):417–424. https://doi.org/10.4300/JGME-D-11-00277.1
13. Stern J, Sharma S, Mendoza P et al (2011) Surgeon perception is not a good predictor of peri-operative outcomes in robot-assisted radical prostatectomy. J Robot Surg 5(4):283–288. https://doi.org/10.1007/s11701-011-0293-4
14. Ganni S, Chmarra MK, Goossens RHM, Jakimowicz JJ (2017) Self-assessment in laparoscopic surgical skills training: is it reliable? Surg Endosc 31(6):2451–2456. https://doi.org/10.1007/s00464-016-5246-6
15. Bull NB, Silverman CD, Bonrath EM (2019) Targeted surgical coaching can improve operative self-assessment ability: a single-blinded nonrandomized trial. Surgery. https://doi.org/10.1016/j.surg.2019.08.002
16. Ritter EM, Gardner AK, Dunkin BJ, Schultz L, Pryor AD, Feldman L (2020) Video-based assessment for laparoscopic fundoplication: initial development of a robust tool for operative performance assessment. Surg Endosc 34(7):3176–3183. https://doi.org/10.1007/s00464-019-07089-y
17. Watanabe Y, Bilgic E, Lebedeva E et al (2016) A systematic review of performance assessment tools for laparoscopic cholecystectomy. Surg Endosc 30(3):832–844. https://doi.org/10.1007/s00464-015-4285-8
18. Vassiliou MC, Feldman LS, Andrew CG et al (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190(1):107–113. https://doi.org/10.1016/j.amjsurg.2005.04.004
19. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL (2005) Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. Surgery 138(4):640–647; discussion 647–649. https://doi.org/10.1016/j.surg.2005.07.017
20. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ (2012) The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. Acad Med 87(10):1401–1407. https://doi.org/10.1097/ACM.0b013e3182677805
21. Terwee CB, Mokkink LB, Knol DL, Ostelo RW, Bouter LM, de Vet HC (2012) Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. Qual Life Res 21(4):651–657. https://doi.org/10.1007/s11136-011-9960-1
22. Bull NB, Silverman CD, Bonrath EM (2020) Targeted surgical coaching can improve operative self-assessment ability: a single-blinded nonrandomized trial. Surgery 167(2):308–313. https://doi.org/10.1016/j.surg.2019.08.002
23. Thanawala RM, Jesneck JL, Seymour NE (2019) Education management platform enables delivery and comparison of multiple evaluation types. J Surg Educ 76(6):e209–e216. https://doi.org/10.1016/j.jsurg.2019.08.017
24. Aggarwal R, Grantcharov T, Moorthy K et al (2007) An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Ann Surg 245(6):992–999. https://doi.org/10.1097/01.sla.0000262780.17950.e5
25. Williams RG, Sanfey H, Chen XP, Dunnington GL (2012) A controlled study to determine measurement conditions necessary for a reliable and valid operative performance assessment: a controlled prospective observational study. Ann Surg 256(1):177–187. https://doi.org/10.1097/SLA.0b013e31825b6de4
26. Ali A, Subhi Y, Ringsted C, Konge L (2015) Gender differences in the acquisition of surgical skills: a systematic review. Surg Endosc 29(11):3065–3073. https://doi.org/10.1007/s00464-015-4092-2
27. Kim MJ, Williams RG, Boehler ML, Ketchum JK, Dunnington GL (2009) Refining the evaluation of operating room performance. J Surg Educ 66(6):352–356. https://doi.org/10.1016/j.jsurg.2009.09.005
28. Calland JF, Turrentine FE, Guerlain S et al (2011) The surgical safety checklist: lessons learned during implementation. Am Surg 77(9):1131–1137
29. Norman GR, Sloan JA, Wyrwich KW (2003) Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care 41(5):582–592. https://doi.org/10.1097/01.mlr.0000062554.74615.4c
30. Hong S, Lynn HS (2020) Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol 20(1):1–12
31. Steiner DL, Norman GR (2008) Health measurement scales. A practical guide to their development and use. Oxford University Press, Oxford
32. Stefanidis D, Chintalapudi N, Anderson-Montoya B, Oommen B, Tobben D, Pimentel M (2017) How often do surgeons obtain the critical view of safety during laparoscopic cholecystectomy? Surg Endosc 31(1):142–146. https://doi.org/10.1007/s00464-016-4943-5
33. Sgaramella LI, Gurrado A, Pasculli A et al (2021) The critical view of safety during laparoscopic cholecystectomy: Strasberg Yes or No? An Italian Multicentre study. Surg Endosc 35(7):3698–3708. https://doi.org/10.1007/s00464-020-07852-6
34. Eva KW, Regehr G (2005) Self-assessment in the health professions: a reformulation and research agenda. Acad Med 80(10 Suppl):S46-54. https://doi.org/10.1097/00001888-200510001-00015
35. Ganni S, Botden S, Schaap DP, Verhoeven BH, Goossens RHM, Jakimowicz JJ (2018) "Reflection-Before-Practice" improves self-assessment and end-performance in laparoscopic surgical skills training. J Surg Educ 75(2):527–533. https://doi.org/10.1016/j.jsurg.2017.07.030
36. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D (2012) Rater training to support high-stakes simulation-based assessments. J Contin Educ Health Prof 32(4):279–286. https://doi.org/10.1002/chp.21156
37. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS (2009) Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. J Gen Intern Med 24(1):74–79. https://doi.org/10.1007/s11606-008-0842-3
38. Woehr DJ, Huffcutt AI (1994) Rater training for performance appraisal: a quantitative review. J Occup Organ Psychol 67(3):189–205
39. Tripepi G, Jager KJ, Dekker FW, Zoccali C (2010) Selection bias and information bias in clinical research. Nephron Clin Pract 115(2):c94–c99
40. Mokkink LB, Prinsen C, Patrick DL et al (2018) COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). User Manual 78(1)
41. Hahn S, Puffer S, Torgerson DJ, Watson J (2005) Methodological bias in cluster randomised trials. BMC Med Res Methodol 5:10. https://doi.org/10.1186/1471-2288-5-10
42. Dalmaijer ES, Nord CL, Astle DE (2020) Statistical power for cluster analysis. arXiv preprint. arXiv:200300381
43. Clarke P (2008) When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. J Epidemiol Community Health 62(8):752–758