Surgical Endoscopy (2020) 34:3176–3183
https://doi.org/10.1007/s00464-019-07089-y

2019 SAGES ORAL

Video-based assessment for laparoscopic fundoplication: initial development of a robust tool for operative performance assessment

E. Matthew Ritter[1] · Aimee K. Gardner[2] · Brian J. Dunkin[3] · Linda Schultz[4] · Aurora D. Pryor[5] · Liane Feldman[6]

Received: 7 May 2019 / Accepted: 21 August 2019 / Published online: 11 September 2019

Abstract

Introduction While better technical performance correlates with improved outcomes, there is a lack of procedure-specific tools to perform video-based assessment (VBA). SAGES is developing a series of VBA tools with enough validity evidence to allow reliable measurement of surgeon competence. A task force was established to develop a VBA tool for laparoscopic fundoplication using an evidence-based process that can be replicated for additional procedures. The first step in this process was to seek content validity evidence.

Methods Forty-two subject matter experts (SME) in laparoscopic fundoplication were interviewed to obtain consensus on procedural steps, identify potential variations in technique, and to generate an inventory of required skills and common errors. The results of these interviews were used to inform creation of a task inventory questionnaire (TIQ) that was delivered to a larger SME group ($n = 188$) to quantify the criticality and difficulty of the procedural steps, the impact of potential errors associated with each step, the technical skills required to complete the procedure, and the likelihood that future techniques or technologies may change the presence or importance of any of these factors. Results of the TIQ were used to generate a list of steps, skills, and errors with strong validity evidence.

Results Initial SMEs interviewed included fellowship program directors (45%), recent fellows (24%), international surgeons (19%), and highly experienced super SMEs with quality outcomes data (12%). Qualitative analysis of interview data identified 6 main procedural steps (visualization, hiatal dissection, fundus mobilization, esophageal mobilization, hiatal repair, and wrap creation) each with 2–5 sub steps. Additionally, the TIQ identified 5–10 potential errors for each step and 11 key technical skills required to perform the procedure. Based on the TIQ, the mean criticality and difficulty scores for the 11/21 sub steps included in the final scoring rubric is 4.66/5 (5 = absolutely essential for patient outcomes) and 3.53/5 (5 = difficulty level requires significant experience and use of alternative strategies to accomplish consistently), respectively. The mean criticality and frequency scores for the 9/11 technical skills included is 4.51/5 and 4.51/5 (5 = constantly used ≥ 80% of the time), respectively. The mean impact score of the 42/47 errors incorporated into the final rubric is 3.85/5 (5 = significant error that is unrecoverable, or even if recovered, likely to have a negative impact on patient outcome).

Conclusions A rigorous, multi-method process has documented the content validity evidence for the SAGES video-based assessment tool for laparoscopic fundoplication. Work is ongoing to pilot the assessment tool on recorded fundoplication procedures to establish reliability and further validity evidence.

Keywords Video-based assessment · Assessment · Fundoplication · Content validity · Test development

✉ E. Matthew Ritter
eritter@usuhs.edu

[1] Department of Surgery, Uniformed Services University of the Health Sciences and Walter Reed National Military Medical Center, 4301 Jones Bridge Road A 3020, Bethesda, MD 20878, USA

[2] Baylor College of Medicine, Houston, TX, USA

[3] Houston Methodist, Weill Cornell Medical College, Houston, TX, USA

[4] Society of American Gastrointestinal and Endoscopic Surgeons, Los Angeles, CA, USA

[5] Department of Surgery, Stony Brook University Renaissance School of Medicine, Stony Brook, NY, USA

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

Operative skills are an essential domain of surgical practice and conventional wisdom holds that surgeon proficiency impacts patient outcomes. Until recently, however, evidence for this association has been limited and hampered by challenges in measurement [1]. With improved techniques, emerging evidence now supports the association between technical performance (measured using intraoperative video) and postoperative complications in bariatric, pancreatic and colonic surgery [2–4]. This work is particularly important in that it goes beyond the traditional efforts to improve surgical quality—adherence to care processes and measurement of complications—and focuses on the "black box" of how the operative procedure was actually performed [5, 6].

To further work on measuring procedural competence in surgery, more assessment tools supported by a level of validity evidence that enables them to be used for competency assessment are needed. The Society of American Gastrointestinal and Endoscopic Surgeons (SAGES) envisions using such tools in the process of assessing a fellow's readiness to transition to independent practice or a practicing surgeon's successful completion of a curriculum for up skilling in a particular area of surgery using the SAGES Masters Program [7]. Part of this vision includes the identification of "index operations" that embody a technical skill set that is representative of a class of surgeries. Laparoscopic fundoplication (LF) is an example of such an "index operation" for foregut surgery.

A recent systematic review of the literature revealed limited validity evidence for existing tools to assess competency in LF [8]. As a result, SAGES has embarked on an effort to create a video-based assessment (VBA) tool to be used to determine if a surgeon entering or already in practice is performing LF with acceptable quality and safety. The tool is being created using best practices in assessment development and will be used for summative feedback to surgeons finishing minimally invasive surgery fellowships and those completing the foregut pathway in the SAGES Masters Program. The process used to create this assessment tool and generate the supportive evidence for using it as a highly reliable assessment of technical competence will serve as a template for creation of other assessment tools for additional anchor procedures.

The specific aim of this study was to demonstrate the content validity evidence of a novel SAGES VBA for LF, and highlight the process involved to achieve this level of evidence.

Methods

Test specifications

Prior to beginning the project, we first gathered a convenience sample of SAGES member subject matter experts (SMEs) to develop and confirm purposes and uses of the assessment, informing the development of the test specifications. The test specifications document describes the statement of purpose of the test, intended users and uses, intended interpretations, the construct domain to be measured, intended examinee population, performance level descriptors of the minimally qualified candidate, test format, test administration, security, and scoring procedures in the form of a test definition document. The group confirmed that the proposed procedures to develop the test conformed to industry standards jointly established by the National Council for Measurement in Education (NCME), the American Psychological Association (APA), and the American Educational Research Association (AERA) [9] in order to ensure comprehensiveness, rigor, and legal defensibility.

After confirmation of the purposes and uses of the assessment, we conducted a comprehensive job task analysis (JTA) to support the claim that the skills and abilities to be assessed are required for safe and independent performance of LF and are consistent with the purpose for which the test is being created [10, 11]. This process documents both the discrete and observable procedural activities required and the context in which those activities are performed. The JTA serves as the single most important stage to drive creation of appropriate tools and satisfy professional and legal requirements [9]. Thus, multiple data collection methodologies were used to meet best practice recommendations, including hierarchical task analysis, semi-structured interviews, and dissemination of a task inventory questionnaire [12, 13].

Subject matter expert interviews

As the responsibilities identified from the job task analysis serve as an anchor point in the validity argument, it is imperative that it is rigorously conducted and with a comprehensive group of SMEs, especially when seeking to inform competency decisions [14]. Thus, a variety of SMEs were identified for inclusion in the interview process based on practice types (e.g., academic, community, private), ranges of experience (e.g., recently completed Minimally Invasive Surgery (MIS) fellowship, experienced clinicians, and MIS fellowship educators/Program Directors), procedural experience, and demonstration of quality performance via case quality metrics (e.g., GERD HRQL and Dysphagia Scores).

One-hour semi-structured phone interviews with each SME were conducted with an Industrial-Organizational Psychologist using the critical incident technique (CIT) [15]. This method allowed for the collection of experiences and anecdotes about procedure-related incidents that describe particularly effective or ineffective job performance. The interview script was organized according to procedural steps, with prompts for SMEs to provide information on missing information, instances of effective performance,

instances of ineffective performance, errors experienced or observed, and skills required to perform each step of the procedure. SMEs also provided information regarding most critical step for optimal patient outcomes, common reasons to re-intervene, key competencies required for performing the procedure successfully, and most challenging aspects of the procedure. Surgeons who reported that they trained residents, fellows, or other colleagues on the procedure also indicated most common steps of the procedure in which they have to take over, why they have to take over, and common areas where trainees struggle. Finally, to ensure relevance and longevity of the tool, surgeons also indicated the extent to which they believed skills required to complete the procedure might change (i.e., become more or less important) in the future.

Task inventory questionnaire

Information from the SME interviews was organized into major domains to produce an inventory of tasks for LF. This task inventory then informed development of an online task inventory questionnaire (TIQ), which was organized by SME consensus of steps and sub steps of the procedure. SMEs indicated the importance (1 = not essential/optional: patient outcome not likely effected, 3 = important: may impact short term patient experience but not ultimate outcome, 5 = absolutely essential: Patient outcome likely to be affected) and difficulty (1 = easy: requires little experience and no use of alternative strategies to accomplish consistently, 3 = medium: requires some experience and use of alternate strategies to accomplish consistently, 5 = difficult: requires significant experience and use of alternative strategies to accomplish consistently) of performing each sub step. Errors associated with each step were also rated on their importance and likelihood to impact patient outcomes (1 = minimal: error causes delays of inefficiencies but if not corrected will have little or no effect on patient outcome, 3 = medium: recoverable error that if not corrected may have a negative impact on patient outcome, 5 = significant: critical error that is unrecoverable, or even if recovered, has a high likelihood for a negative impact on patient outcome). Finally, SMEs also rated the frequency (1 = rarely: < 20% of the time, 3 = about half the time: 40–60% of the time, 5 = constantly: > 80% of the time), current importance (1 = not essential/optional, 3 = important, 5 = absolutely essential, see definitions above) and perceptions of future importance (1 = importance decreases significantly, 3 = importance stays the same, 5 = importance increases significantly) of 11 skills associated with the procedure.

Surgeons credentialed to perform the procedure and who had performed at least twelve 360 degree fundoplications with our without paraesophageal hernia repairs over the past year were invited to complete the TIQ through email distribution to members of the Society of American Gastrointestinal and Endoscopic Surgeons (SAGES), the Japanese Society for Endoscopic Surgery (JSES), and the American Society for Metabolic and Bariatric Surgery (ASMBS). Demographics and experience data were collected from each SME who completed the TIQ. These data include race, gender, age, geographical location, years since training, fellowship completion (yes/no), number of years performing laparoscopic 360 degree fundoplications, and number of fundoplications performed in the past year according to four categories: (1) 360 degree fundoplications for reflux with < 4 cm hiatal hernia, (2) paraesophageal hernia repair with concomitant LF, (3) antireflux procedure during a Heller myotomy, and (4) antireflux procedure during another procedure for any other indication (e.g., esophageal injury).

Interview data were examined via thematic and frequency analysis. Basic descriptives, item-total correlations, Cronbach's alpha, and multivariate analysis of variance (MANOVA) were used to examine the data with SPSS 25.0. As no human subjects were involved, IRB approval was not required.

Results

Test specifications

Specifications were defined relative to the purpose, domain, and intended interpretations and uses of the assessment scores. The assessment is intended to inform competency judgments about the ability to safely, effectively, and independently perform a laparoscopic posterior fundoplication for surgeons in training, completing training, and/or in practice. The domain of measurement includes the skills and behaviors required for safe, effective, and independent performance of a laparoscopic posterior fundoplication procedure. The test is designed to differentiate between adequate and inadequate performance, based on a predetermined passing score.

Candidates who receive a passing score have demonstrated the required level of competence in LF (operationalized as a passing score specified by standard setting procedures). The assessment is not designed to differentiate among the passing candidates (i.e., a higher passing score does not necessarily indicate better performance). Additionally, a passing score signifies procedural competency only, as other cognitive skills and judgment are also required for safe and independent practice. Candidates who receive a failing score have not demonstrated the required level of competency in the domain. All statements above relate to the intended inferences to be made from the scores generated by the application of the assessment tool. The final scoring system for the tool has not yet been determined.

SME interviews

SMEs interviewed included Fellowship Program Directors (45%), recent Fellows (24%), international surgeons (19%), and highly experienced super SMEs with quality outcomes data (12%). Qualitative analysis of interview data confirmed 6 main procedural steps (visualization, hiatal dissection, fundus mobilization, esophageal mobilization, hiatal repair, and wrap creation) each with 2–5 sub steps (21 total). Additionally, interview data indicated there were 5–10 potential errors for each step (47 total) and 11 key technical skills required to perform the procedure.

Task inventory questionnaire

One hundred and eighty-eight surgeons (85% men, 69% caucasian) completed the TIQ. Surgeons were an average age of 47 (± 10) years, were an average of 15 (± 11) years past residency, and had been performing 360 degree fundoplications for approximately 13 (± 9) years. Approximately three quarters (77%) were fellowship-trained. The average number of fundoplication cases performed in the past year was 43 (± 36), which included both 360-degree fundoplication (22 ± 26) and paraesophageal hernia repairs (21 ± 19). Surgeons had considerably high levels of agreement among all items ***($\alpha = 0.95$), which captured sub step importance ($\alpha = 0.83$), sub step difficulty ($\alpha = 0.91$), errors within each step ($\alpha = 0.92$), skill frequency ($\alpha = 0.85$), skill importance ($\alpha = 0.85$), and future importance of skills ($\alpha = 0.96$).

Table 1 displays SME ratings of importance and difficulty for each of the sub steps of the procedure, along with the percentage of who responded that the patient outcome likely to be directly affected by the performance of each sub step (absolutely essential). As shown, there was a wide variability in the reported importance of each sub step, ranging from a mean 1.95 (reinforce closure with mesh) to a mean of 4.81 (ensure wrap positioned around esophagus). Difficulty was rated close to the middle of the scale for the majority of sub steps, with the most difficult sub steps including safely divide tissues surround the esophagus (3.91 ± 0.86), create adequate retroesophageal window (3.79 ± 0.86), and position the wrap around esophagus (3.74 ± 0.93). Errors associated with each step and their impact on patient outcomes are displayed in Table 2.

The current importance, frequency of use, and future importance of each skill associated with the procedure is displayed in Table 3. As shown, SMEs indicated that the most important skills to achieve a good outcome for the operation were suturing (4.76 ± 0.55), dissection (4.73 ± 0.51), and tissue handling (4.62 ± 0.63). However, the skills used most frequently include dissection (4.72 ± 0.61), tissue handling (4.71 ± 0.67), and two-handed technique (4.71 ± 0.62). Overall, SMEs indicated that the importance of these 11 skills would remain the same in the future, although all means reveal a slight leaning (i.e., > 3 on the 5-point Likert scale) towards more important in the future.

Inclusion in the assessment tool

The development team then critically analyzed the results of the TIQ to ensure adequate representation of the construct while maximizing content validity evidence. Cut scores for inclusion in the rubric were examined using both empirical (mean ± SD) and rational (criterion based scoring) approaches. Empirical approaches proved to be either overly inclusive or dismissive depending on the area in question. Ultimately, a rational approach was chosen for steps, errors, and skills. For steps and sub steps, only those with a mean importance score > 4.5 and a difficulty score > 3.0 were chosen. An importance score > 4.5 corresponds with steps of the operation that are highly essential and have the most potential effect on outcome. A difficulty score of > 3.0 ensures that easy tasks requiring only limited experience are not included. Of the original 6 steps and 21 sub steps, 5 steps and 11 sub steps met these criteria. The mean importance and difficulty of the sub steps included in the final rubric are 4.66 ± 0.1 and 3.53 ± 0.3 respectively. Similar decisions were made for skills with a mean importance score > 4.0 and mean frequency score > 3.0 and for errors with a mean impact score > 3.0. These decisions insured that skills measured were used the majority of the time, were thought to be necessary to achieve a good outcome, and that errors only included those factors important enough to potentially affect outcome. Of the original 11 skills, 9 made the final cut with a mean importance score of 4.51 ± 0.19 and mean frequency score of 4.51 ± 0.18. These 9 skills also showed potential to remain consistent or increase in importance in the future with a mean future importance score of 3.52 ± 0.12. Lastly, 46/47 identified errors met the impact score criteria for inclusion; however, 5 errors were listed in the visualization step which was excluded from the final rubric in the sub step selection process. The mean impact score of the remaining 42 errors is 3.85 ± 0.52, indicating only errors that were thought to possibly effect patient outcome are included. All steps, sub steps, skills, and errors incorporated into the final rubric are annotated in Tables 1, 2, and 3.

Discussion

Validity in assessment is defined as the degree to which evidence and theory support the interpretations of test *scores* for the proposed uses of the test. The word scores is emphasized to reinforce that a *test* or *assessment tool* is never valid or validated in and of itself, only the results of an assessment are valid when used for specific purposes. Validity of

Table 1 Importance and difficulty means for all sub steps of the procedure

Step	SubStep	Criticality Mean ± SD	Absolutely essential (%)	Difficulty Mean ± SD	Difficult (%)
Visualization of the operative field	Ensure safe & effective liver retraction	4.18 ± 0.91	49	2.20 ± 1.06	2
	Put pars flaccida in view	3.99 ± 1.04	42	1.93 ± 1.02	3
	Ensure diaphragmatic hiatus is in frame	4.50 ± 0.78	66	2.28 ± 1.13	4
Hiatal dissection[a]	Open gastrohepatic ligament	4.32 ± 0.96	60	1.86 ± 0.95	2
	Open/release the phrenoesophageal ligament[a]	4.60 ± 0.76	72	3.02 ± 0.87	3
	Create adequate retroesophageal window[a]	4.78 ± 0.54	82	3.79 ± 0.86	19
	Safely manipulate esophagus (with or without penrose)[a]	4.74 ± 0.58	80	3.48 ± 0.97	12
Fundus mobilization[a]	Mobilize fundus for wrap creation (with or without division of short gastrics)[a]	4.59 ± 0.74	71	3.25 ± 0.89	7
	Divide retrogastric attachments	4.29 ± 0.84	48	3.46 ± 0.88	10
	Complete visualization of base of left crus[a]	4.53 ± 0.68	61	3.73 ± 0.77	12
Esophageal mobilization[a]	Retract esophagus to optimize mediastinal dissection	4.30 ± 0.82	52	3.58 ± 0.90	14
	Safely divide tissues surrounding the esophagus[a]	4.53 ± 0.69	65	3.91 ± 0.86	26
Hiatal repair[a]	Expose posterior junction of right and left crus	4.49 ± 0.70	60	3.56 ± 0.89	12
	Close crura with sutures[a]	4.69 ± 0.59	76	3.54 ± 0.85	12
	Reinforce closure with pledgets	2.08 ± 1.24	5	3.15 ± 1.05	10
	Reinforce closure with mesh	1.95 ± 1.12	4	3.40 ± 0.98	13
Wrap creation[a]	Pass the fundus posteriorly[a]	4.60 ± 0.72	72	3.32 ± 0.97	11
	Position the wrap around esophagus (assess geometry/twist)[a]	4.76 ± 0.49	79	3.74 ± 0.93	21
	Ensure wrap positioned around esophagus (not stomach)[a]	4.81 ± 0.45	84	3.58 ± 0.97	18
	Assess wrap tension (with or without bougie)[a]	4.46 ± 0.81	62	3.62 ± 0.94	17
	Secure wrap with suture[a]	4.65 ± 0.68	75	3.50 ± 1.02	17

"Absolutely essential" indicates the percentage of subject matter experts (SMEs) who responded that the sub step was "absolutely essential: patient outcome likely to be directly affected by the performance of each step." "Difficult" indicates the percentage of SMEs who responded that the sub step was "Difficult: requires significant experience and use of alternative strategies to accomplish consistently."

[a]Indicates inclusion in the final rubric

educational and psychological assessments are guided by the Standards for Educational and Psychological Testing known commonly as *The Standards* [9]. The Standards are developed and published periodically by a joint committee of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. Since 1985, the Standards have supported a unitary concept of validity where a construct is supported by 5 recognized types of validity evidence. A construct is defined as a concept or characteristics that a test is designed to measure. Validity evidence to support the construct can be classified as content, relation to other variables, internal structure, response process, and consequences. Of these types of evidence, the one most critical for assessments resulting in certification or credentialing is content validity evidence [9]. It is essentially the relationship between the content of an assessment and the construct it is trying to measure. Failure to measure all important themes of the construct results in a threat to validity known as construct underrepresentation [16]. Similarly, measuring aspects not clearly related to the construct results in a threat to validity known as construct irrelevant variance [16]. For the purposes of our study, the construct is a safely and effectively performed laparoscopic posterior 360-degree fundoplication. An example of construct underrepresentation would be measuring only the completion time for the procedure, or the surgeon's skill in suturing. While these measurements might be indicators of skilled surgeons, and correlate highly with other aspects of performance, they lack the content validity evidence for use in a certification exam. Construct irrelevant variance would be seen if unimportant steps or meaningless errors were scored with an equal weighting to more significant aspects of the procedure, or if only a single, specific technique, such as use of a specific laparoscopic suturing device was required by the assessment tool. The experience, geographic distribution, and diversity of the SME groups involved in the creation of this tool along with the high levels of frequency of use, and estimated future importance of

Table 2 Mean impact scores for all identified errors

Step	Possible errors	Impact Mean ± SD
Visualization of the operative field	Inadequate liver retraction	2.75 ± 0.97
	Bleeding/injury to liver or surrounding structures	3.21 ± 0.81
	Suboptimal view from poor camera port placement	3.39 ± 0.94
	Suboptimal instrument port placement	3.26 ± 1.01
	Suboptimal instrument port placement	3.21 ± 1.17
Hiatal dissection*	Damage to esophagus	4.56 ± 0.66
	Damage to vagus nerves	3.84 ± 0.97
	Damage to stomach	3.25 ± 0.79
	Damage to major vascular structure	4.52 ± 0.77
	Damage to diaphragm/denuding	3.11 ± 0.95
	Damage to other surrounding structures	3.36 ± 0.81
	Dissecting in wrong areas	3.76 ± 0.87
	Unsafe use of surgical energy	4.38 ± 0.75
Fundus mobilization*	Damage to spleen	3.90 ± 0.80
	Improper mobilization (too much/little)	3.62 ± 0.81
	Damage to fundus or body of stomach	3.28 ± 0.86
	Unsafe use of surgical energy	4.11 ± 0.77
	Bleeding from short gastrics	3.29 ± 0.85
	Injury to major vascular structure	4.60 ± 0.59
	Inappropriate retraction (too much/little)	3.31 ± 0.80
	Incomplete division of phrenoesophageal ligament	3.54 ± 0.89
Esophageal mobilization*	Damage to pleura	2.65 ± 1.04
	Damage to vagus nerve	3.59 ± 1.02
	Damage to esophagus	4.59 ± 0.59
	Damage to stomach	3.41 ± 0.88
	Damage to major cardiovascular structure	4.79 ± 0.53
	Inadequate mobilization	3.83 ± 0.86
	Bleeding from mediastinal tissues	3.51 ± 0.91
	Unsafe use of surgical energy	4.29 ± 0.76
Hiatal repair*	Damage crura	3.22 ± 0.90
	Damage to major cardiovascular structure	4.80 ± 0.51
	Damage to esophagus from improper needle handling	4.12 ± 0.87
	Damage to stomach from improper needle handling	3.22 ± 0.95
	Closure too tight	3.96 ± 0.95
	Closure too loose	3.64 ± 0.94
	Improper bites (too small/big)	3.51 ± 0.80
	Poor knot tying technique	3.82 ± 0.93
Wrap creation*	Improper tension	3.87 ± 0.86
	Use wrong part of stomach	4.37 ± 0.72
	Wrap around stomach (not esophagus)	4.41 ± 0.75
	Injury to the esophagus	4.52 ± 0.65
	Injury to the stomach	3.51 ± 0.92
	Improper orientation (geometry twist)	4.33 ± 0.75
	Poor suturing/knot tying	3.90 ± 0.90
	Improper wrap length (too long/short	3.94 ± 0.92
	Wrap not sutured to esophagus	3.65 ± 1.02
	Vagus nerve injury	3.73 ± 1.04

*Indicates steps of procedure where all errors are included on the final rubric

Table 3 Current importance, frequency, and future importance of each of the skills required to perform the procedure

	Criticality Mean ± SD	Frequency Mean ± SD	Future Importance Mean ± SD
Suturing[a]	4.76 ± 0.55	4.53 ± 0.97	3.48 ± 1.00
Use of energy[a]	4.55 ± 0.71	4.53 ± 0.85	3.64 ± 0.91
Knot tying[a]	4.60 ± 0.74	4.47 ± 1.0	3.33 ± 1.13
Tissue handling[a]	4.62 ± 0.63	4.71 ± 0.67	3.65 ± 0.88
Dissection[a]	4.73 ± 0.51	4.72 ± 0.61	3.67 ± 0.87
Managing assistant	3.77 ± 0.97	3.95 ± 0.97	3.10 ± 1.12
Retraction[a]	4.21 ± 0.86	4.36 ± 0.85	3.46 ± 0.96
Two-handed technique[a]	4.43 ± 0.76	4.71 ± 0.62	3.56 ± 0.94
Endoscopy	3.05 ± 1.31	2.97 ± 1.43	3.28 ± 1.11
Needle handling[a]	4.35 ± 0.88	4.28 ± 1.06	3.38 ± 1.06
Working in confined space[a]	4.33 ± 0.82	4.26 ± 0.87	3.49 ± 0.98

[a]Indicates inclusion in the final rubric

the steps, skills, and errors included speaks directly to the representation of the construct, while the current importance and perceived impact establishes the relevance to the LF construct.

We believe this work represents the most robust development of an observational workplace based assessment of surgical skill to date. A 2011 systematic review of observational assessment tools for surgical skills showed that while there was some validity evidence for tools applied at the trainee level, there was no validity evidence to support use of the scores at the level of the practicing surgeon [17]. Interestingly, even this review did not analyze the validity evidence of the assessment tool results in the unitary validity framework [18] present in the Standards since 1985. This highlights how more work like this is needed to bring the science in surgical performance measurement into the twenty-first century [19]. At least one group did take a step in the positive direction with the development of the competency assessment tool (CAT) for laparoscopic colorectal surgery in the National Training Programme in England [20]. While the steps to ensure that the content aligned with the domain were similar when compared with our work, only 7 SMEs were initially interviewed to generate the task inventory, and only 15 additional surgeons from the UK alone were queried with a form of a task inventory questionnaire. Similarly, the Bariatric Objective Structured Assessment of Technical Skill (BOSATS) [21] tool upon which the work correlating surgeon performance to outcomes in bariatric surgery [2] was based, involved only 2 surgeons to generate the task inventory and then approximately 30 surgeons from 6 different countries completed the questionnaire. Our work represents a much more broadly inclusive approach with a 6–20 fold increase in the SME's used to represent the construct (42 for the task inventory and 188 on the TIQ) with purposeful inclusion of SMEs from North America, South America, northern and southern Europe, and Asia.

There are several limitations worth mentioning for an assessment developed with this methodology. First, the content representation of the construct is based on how the procedure has been performed and is currently being performed. Despite SME estimates that the measured skills will become slightly more important in the future, this may not be the case. Development of disruptive technology such as automated robotic suturing or dissection guided by Artificial Intelligence (AI) could render portions of this assessment tool obsolete. Thus the content must be revisited periodically to ensure it continues to align with the construct of laparoscopic fundoplication as it is being practiced clinically. A second limitation is lack of a current estimation of reliability or internal structure validity evidence. We are actively developing a rater training program and recruiting raters from outside of the core development group. Once these additional raters are trained, we plan to employ generalizability theory to estimate the number of performances and raters required to make a reliable competency decision. Lastly is the issue of scalability. Currently, application of the assessment tool to a video performance requires a trained human rater to view all steps of the procedure in its entirety. This involves hours of time from multiple individuals. Historically, assessments requiring expert rater review are cumbersome and expensive to deploy on a large scale. We are currently exploring scalability options including potential use of crowd sourcing and/or AI to lessen the burden on human expert raters.

Future steps in development of this tool in addition to the rater training program and reliability assessments mentioned above include an evidence-based standard setting process to establish a standardized scoring system along with a defensible pass/fail cut score, determining potential applicability of scores to similar procedures such as partial fundoplications, hiatal hernias > 4 cm, reoperative procedures, and potentially robotic assisted procedures. Given the high level of content validity evidence in the tool, focusing on the core principles of the procedure, the applicability of the tool to those procedures is likely.

Conclusion

A rigorous, multi-method process has documented the content validity evidence for the SAGES video-based assessment tool to measure performance within the construct of laparoscopic fundoplication. The high levels of current importance, future importance, frequency of use, and difficulty directly address the common threats to validity of construct underrepresentation and construct irrelevant variance.

Work is ongoing to establish reliability and further validity evidence.

Compliance with ethical standards

Disclosures Dr. Ritter reports royalties from the Henry M. Jackson Foundation for the Advancement of Military Medicine outside the submitted work. Dr. Gardner reports fees for contracted consultant work from the SAGES Foundation during the conduct of the study. Dr. Dunkin reports fees for contracted consultant work from the SAGES Foundation during the conduct of the study and is currently employed by Boston Scientific. His employment with Boston Scientific occurred outside of the submitted work. Ms. Shultz has nothing to disclose. Dr. Pryor reports speaking fees from Ethicon, Gore Medical, Medtronic, and Stryker, all outside the submitted work. Dr. Feldman reports research Grants from Medtronic and Merck, both outside the submitted work.

References

1. Fecso AB, Szasz P, Kerezov G, Grantcharov TP (2017) The effect of technical performance on patient outcomes in surgery: a systematic review. Ann Surg 265:492–501
2. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ, Michigan Bariatric Surgery C (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369:1434–1442
3. Hogg ME, Zenati M, Novak S, Chen Y, Jun Y, Steve J, Kowalsky SJ, Bartlett DL, Zureikat AH, Zeh HJ 3rd (2016) Grading of surgeon technical performance predicts postoperative pancreatic fistula for pancreaticoduodenectomy independent of patient-related variables. Ann Surg 264:482–491
4. Mackenzie H, Ni M, Miskovic D, Motson RW, Gudgeon M, Khan Z, Longman R, Coleman MG, Hanna GB (2015) Clinical validity of consultant technical skills assessment in the English National Training Programme for laparoscopic colorectal surgery. Br J Surg 102:991–997
5. Grenda TR, Pradarelli JC, Dimick JB (2016) Using surgical video to improve technique and skill. Ann Surg 264:32–33
6. Tam V, Zeh HJ 3rd, Hogg ME (2017) Incorporating metrics of surgical proficiency into credentialing and privileging pathways. JAMA Surg 152:494–495
7. Jones DB, Stefanidis D, Korndorffer JR Jr, Dimick JB, Jacob BP, Schultz L, Scott DJ (2017) SAGES University MASTERS Program: a structured curriculum for deliberate, lifelong learning. Surg Endosc 31:3061–3071
8. Bilgic E, Al Mahroos M, Landry T, Fried GM, Vassiliou MC, Feldman LS (2019) Assessment of surgical performance of laparoscopic benign hiatal surgery: a systematic review. Surg Endosc. https://doi.org/10.1007/s00464-019-06662-9
9. American Educational Research Association APA, National Council on Measurement in Education (2014) Standards for educational and psychological testing. AERA, Washington, DC
10. Ag DC (1986) The vaidity of credentialing examinations. Eval Health Prof 9:137–169
11. Kane MT (1985) The validity of licensure examinations. Am Psychol 7:911–918
12. Annett J (2003) Heirarchical task analysis. In: Hollnage E (ed) Handbook of cognitive task design. LEA, New Jersey, pp 17–36
13. Brannick MT, Pearlman K, Sanchez JI (2017) Work analysis. In: Farr J, Tippins NT (eds) Handbook of employee selection. Routledge, New York
14. Nelson D (1994) Job analysis for licnsure and certification exams: science or politics. Educ Measurement 13:29–35
15. Jc F (1954) The critical incident technique. Psychol Bull 51:317–358
16. Downing SM, Haladyna TM (2004) Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ 38:327–333
17. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB (2011) Observational tools for assessment of procedural skills: a systematic review. Am J Surg 202(469–480):e466
18. Downing SM (2003) Validity: on meaningful interpretation of assessment data. Med Educ 37:830–837
19. Korndorffer JR Jr, Kasten SJ, Downing SM (2010) A call for the utilization of consensus standards in the surgical education literature. Am J Surg 199:99–104
20. Miskovic D, Ni M, Wyles SM, Kennedy RH, Francis NK, Parvaiz A, Cunningham C, Rockall TA, Gudgeon AM, Coleman MG, Hanna GB, National Training Programme in Laparoscopic Colorectal Surgery in E (2013) Is competency assessment at the specialist level achievable? A study for the national training programme in laparoscopic colorectal surgery in England. Ann Surg 257:476–482
21. Zevin B, Bonrath EM, Aggarwal R, Dedy NJ, Ahmed N, Grantcharov TP, ATLAS groups (2013) Development, feasibility, validity, and reliability of a scale for objective assessment of operative performance in laparoscopic gastric bypass surgery. J Am Coll Surg 216:955–965 **(quiz 1029–1031, 1033)**