Surgical Endoscopy (2024) 38:2219–2230
https://doi.org/10.1007/s00464-024-10675-4

2023 SAGES ORAL

Validity and reliability evidence support task-specific metrics for laparoscopic fundoplication

Alexis Desir[1,2] · Carolina Marques[2] · Emile Farah[1] · Shruti R. Hegde[1,2] · Carla Holcomb[1] · Daniel J. Scott[1] · Ganesh Sankaranarayanan[1,2]

Received: 12 September 2023 / Accepted: 30 December 2023 / Published online: 21 February 2024

Abstract

Background Laparoscopic hiatal hernia repair (LHHR) is a complex operation requiring advanced surgical training. Surgical simulation offers a potential solution for learning complex operations without the need for high surgical volume. Our goal is to develop a virtual reality (VR) simulator for LHHR; however, data supporting task-specific metrics for this procedure are lacking. The purpose of this study was to develop and assess validity and reliability evidence of task-specific metrics for the fundoplication phase of LHHR.

Methods In phase I, structured interviews with expert foregut surgeons were conducted to develop task-specific metrics (TSM). In phase II, participants with varying levels of surgical expertise performed a laparoscopic Nissen fundoplication procedure on a porcine stomach explant. Video recordings were independently assessed by two blinded graders using global and TSM. An intraclass correlation coefficient (ICC) was used to assess interrater reliability (IRR). Performance scores were compared using a Kruskal–Wallis test. Spearman's rank correlation was used to evaluate the association between global and TSM.

Results Phase I of the study consisted of 12 interviews with expert foregut surgeons. Phase II engaged 31 surgery residents, a fellow, and 6 attendings in the simulation. Phase II results showed high IRR for both global (ICC = 0.84, $p < 0.001$) and TSM (ICC = 0.75, $p < 0.001$). Significant between-group differences were detected for both global ($\chi^2 = 24.01$, $p < 0.001$) and TSM ($\chi^2 = 18.4$, $p < 0.001$). Post hoc analysis showed significant differences in performance between the three groups for both metrics ($p < 0.05$). There was a strong positive correlation between the global and TSM (rs = 0.86, $p < 0.001$).

Conclusion We developed task-specific metrics for LHHR and using a fundoplication model, we documented significant reliability and validity evidence. We anticipate that these LHHR task-specific metrics will be useful in our planned VR simulator.

Keywords Task-specific metrics · VR simulator · Nissen fundoplication · Surgical education

Achieving technical proficiency in laparoscopic surgery is critical as it remains the most frequently employed surgical technique by case volume [1]. Recent studies in bariatric [2] and colorectal [3] surgery have shown that greater technical skills are associated with better outcomes and fewer complications. Due to the difficulty in acquiring laparoscopic technical skills directly in the operating room, simulation-based training has emerged as a viable alternative [4–9]. Simulation training platforms provide a conducive learning environment to teach the technical and cognitive competencies necessary to master laparoscopic surgery in a safe, patient-free environment, without the cognitive load experienced in the operating room. An effective simulation-based training program is contingent upon having a robust curriculum with clearly defined and quantifiable performance metrics. Such metrics can be summative to establish a high stakes pass/fail determination or formative to provide trainees with targeted feedback for improvement. The most widely used summative tool for the evaluation of surgical performance

Podium presentation at SAGES 2023, Montreal, Canada.

✉ Ganesh Sankaranarayanan
Ganesh.Sankaranarayanan@utsouthwestern.edu

[1] Department of Surgery, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390, USA

[2] Artificial Intelligence and Medical Simulation Lab, Department of Surgery, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, TX 75390, USA

is the Objective Structured Assessment of Technical Skills (OSATS) [10], a validated global tool for assessing operative performance in 6 domains, typically through video-based review. Formative assessment usually requires the creation and validation of task-specific metrics tailored precisely to each procedure or task.

Laparoscopic hiatal hernia repair (LHHR) is a complex procedure requiring advanced surgical training [11]. Attaining proficiency in this procedure is crucial given the high recurrence rate for such hernias, which is up to 50%, especially for paraesophageal hernias [12]. LHHR remains a difficult procedure to master with reported learning curves ranging from 50 to 200 cases [13–15], emphasizing the need to optimize training to acquire the necessary skills. Traditional anatomic models like cadavers and live animal models have been useful for simulating many procedures but may fall short in replicating some important aspects of human HHR and can pose ethical, cost, logistical, and curricular challenges. Advances in technology have made Virtual reality (VR) simulators a potentially ideal solution, offering detailed anatomic representations that are characteristic of HHR and facilitating focused, deliberate practice [16]. Using standardized simulation scenarios, VR trainers also enable automated objective assessment and targeted feedback to improve performance without the need for an expert surgeon reviewer. Importantly, skills acquired in VR simulators have been shown to improve operating room performance [17, 18]. We are developing a VR simulator for LHHR training as part of an NIH-funded project. The purpose of this study was to develop and assess task-specific metrics for LHHR, specifically evaluating their reliability and validity for the fundoplication portion of the procedure.

Materials and methods

This study was approved by the UT Southwestern Institutional Review Board and was done in two phases. In phase 1, interviews were conducted with experts to create task-specific metrics for the assessment of performance in laparoscopic Nissen fundoplication. In phase II, a bench model study was performed to evaluate validity evidence supporting the newly created metrics.

Development of task-specific metrics for fundoplication

We performed a hierarchical task analysis (HTA) of the LHHR by conducting hour-long semi-structured interviews with local foregut surgeons and experts from the Society for American Gastrointestinal and Endoscopic Surgeons (SAGES) Foregut Task Force. HTA in surgery is a well-known method that breaks down any given surgical procedure into tasks, sub-tasks, and motion end effectors, and it has been successfully used to deconstruct various minimally invasive procedures [19–21]. To guide our expert interviews, we formulated an initial list of procedure steps, drawing from recorded operative videos, information from textbooks, and prior task analysis of the laparoscopic fundoplication procedure [22, 23]. Experts were then asked to describe how they perform the procedure, highlight key moments, identify variations in the procedure, and list common procedural errors in order of their severity. The recordings were then independently analyzed by two authors (SH and GS) to create task trees with variations and a list of errors. Any discrepancies were resolved by an expert author (CH) and through consultations with the interviewed experts.

Validity evidence evaluation for the fundoplication task-specific metrics

In phase II, we assessed the validity evidence of the newly created metrics by conducting a study at the UT Southwestern Simulation Center using a porcine explant Nissen fundoplication model. Messick's unitary framework was used to evaluate the validity of our task-specific metrics [24]. Specifically, data were collected to evaluate validity evidence in the following domains: content alignment, response process, internal structure, and relationship to other variables.

Fundoplication simulator design

We created a Nissen fundoplication simulator using a porcine stomach explant, which was placed inside a modified version of a laparoscopic box trainer [4] (Fig. 1). A frozen porcine stomach and esophagus specimen (Animal Technologies Inc., Tyler, Texas) was thawed and positioned in the box trainer. The esophagus was passed through a small circular incision in the lap box, about 2 inches from the base and held taut using an Allis clamp. To prevent lateral movement, the stomach was secured with two alligator clips. To create a retroesophageal window for the fundoplication, a Penrose drain was inserted through a circular incision about 4 inches from the base to lift the stomach at the gastroesophageal junction and keep the model under tension. A 0° laparoscope connected to a standard equipment tower was used for visualization. A pair of standard laparoscopic needle drivers, curved graspers, and scissors were used to perform the procedure. In addition, 2–0 silk sutures pre-cut to 15 cm in length were placed on a foam box to be used for suturing.

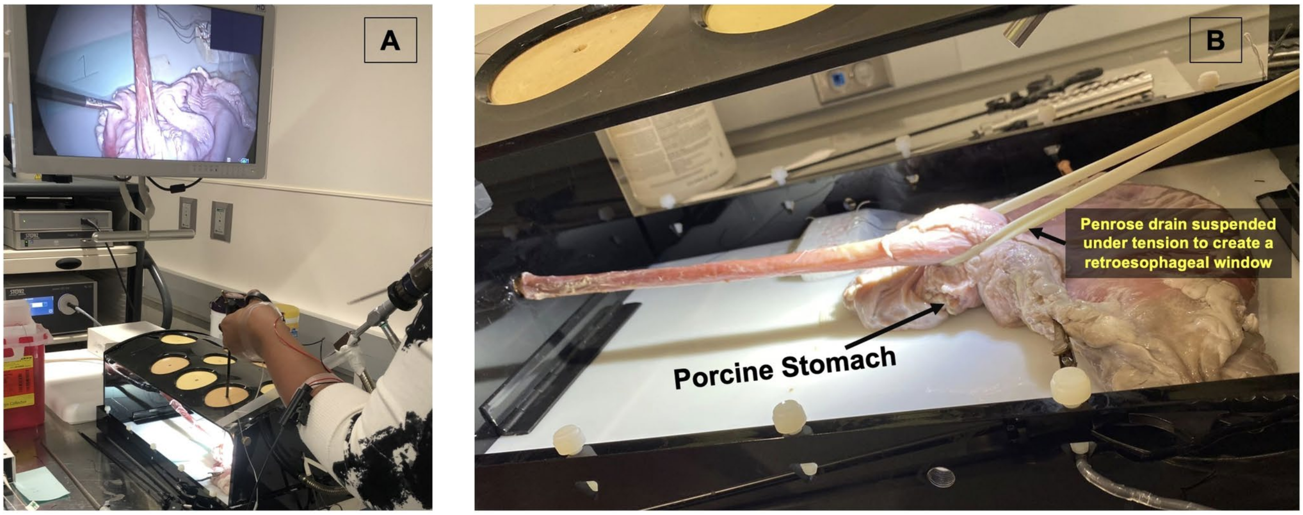

A
B
Penrose drain suspended under tension to create a retroesophageal window
Porcine Stomach

Fig. 1 Laparoscopic Nissen fundoplication simulator; **A** overall set-up; **B** specimen set-up

Study design and procedure

The study was performed at the UT Southwestern Simulation Center with a between-subjects design. Recruited participants were stratified into three groups by level of expertise: novice (post-graduate year [PGY] 1–2 residents in general surgery), intermediates (PGY 3–5 residents and a minimally invasive fellow), and experts (faculty).

Prior to starting the procedure, each participant completed a survey that captured demographic information, clinical experience, and simulator experience. After providing informed consent, participants were given general instructions explaining the study objective and the task, without any technical/operative guidance. Specifically, we did not provide any instructions on number and type of sutures, the distance between the sutures and the placement of the wrap. Participants were then asked to complete a Nissen fundoplication on the porcine stomach model. They were given 1 h to complete up to 2 unassisted attempts. Video recordings focused on the instruments actively utilized in the laparoscopic box trainer, the training model itself, and a card displaying the participants' random identification number to ensure anonymity during the video review process. Additionally, we also collected and analyzed the following real-time in situ metrics for each participant: (I) number of attempts completed (1 or 2), (II) number of sutures placed for fundoplication, (III) space between sutures measured in centimeters, and (IV) whether seromuscular bites were taken through the esophagus (dichotomized as 0 or 1).

At the conclusion of the study, participants were asked to complete a post-simulation survey to assess the quality of the simulator on a 5-point Likert scale. The survey covered 5 categories that included the visual appearance of the simulation, the quality of models and textures, the realism of the simulator interface, how closely the task mirrored the actual surgical procedure, and the simulator's overall effectiveness in teaching LHHR.

Two qualified raters, blinded to the participants' experience levels, independently evaluated the video recordings of performances using both global and task-specific metrics. Table 1 presents the global metrics derived from the OSATS rubric, whereas Table 2 displays the task-specific metrics grounded in the HTA [10, 25–28]. Among the OSATS domains, we excluded the scoring rubric for knowledge of instruments because all participants were provided with the same set of laparoscopic tools. Initially, the two raters assessed the performance of 5 participants, comparing their ratings to discuss the grading and to resolve discrepancies. They then evaluated another 5 videos to ensure concordance between their ratings and reviewed the intraclass correlation coefficient (ICC). Finally, each rater independently graded the remaining videos.

Data analysis

The ICC estimates and their 95% confident intervals for establishing interrater reliability (IRR) were calculated based on mean rating ($k=2$), absolute agreement, and 2-way mixed-effects model. An ICC value between 0.75 and 0.9 was deemed good, while a value above 0.9 was deemed excellent for IRR [29]. A total score was calculated by first averaging the individual metric scores from both raters and then summing them up for both global and task-specific evaluations. The Spearman rank correlation test was used to assess the association between the total global and task-specific scores. To determine performance differences between the three groups, the data were first evaluated for normality using the Shapiro–Wilk test. If the data were normally

Table 1 Rubric for assessing performance using global metrics

Metric	Domain of surgical performance	Rating (5-point Likert scale)				
1	Respect for tissue	Frequently used unnecessary force on tissue or caused damage		Careful handling of tissue but occasionally caused inadvertent damage though excessive force		Consistent handling of tissue, appropriate use of instruments, and force with minimal damage
	Score (scale 1-5)	**1**	**2**	**3**	**4**	**5**
2	Time and motion	Many unnecessary moves		Efficient time/motion but some unnecessary moves		Clear economy of movement and maximum efficiency
		1	**2**	**3**	**4**	**5**
3	Instrument handling	Repeatedly made tentative or awkward moves with instruments		Competent use of instruments but occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
		1	**2**	**3**	**4**	**5**
4	Flow of operation	Frequently stopped procedure and seemed unsure of next move		Demonstrated some forward planning with reasonable progression of procedures		Obviously planned course of procedure with effortless flow from one move to the next
		1	**2**	**3**	**4**	**5**
5	Knowledge of specific procedure	Deficient knowledge		Knew all important steps of procedure		Demonstrated familiarity with all aspects of procedure
		1	**2**	**3**	**4**	**5**
6	Overall performance	Very poor		Competent		Expert level
		1	**2**	**3**	**4**	**5**

distributed, a one-way analysis of variance (ANOVA) was conducted, followed by a pairwise *t* test with Bonferroni correction for post hoc analysis. If not normally distributed, the Kruskal–Wallis test was employed, followed by a pairwise Wilcoxon test with Benjamini–Hochberg correction. Post hoc effect size was reported when appropriate.

Sample size

A priori power analysis was conducted using the G*software [30] to test the difference in performance between the three groups, with $\alpha = 0.5$, a medium effect size $f = 0.5$ and power $\beta = 0.8$. The analysis showed that a total of 30 subjects equally distributed in three groups was needed to achieve the necessary power.

Results

Phase I results

Task analysis

A total of 12 expert foregut surgeons participated in interviews for task analysis, spanning 720 min in total. Table 3 displays the HTA of LHHR, outlining 6 major tasks, 27 subtasks, and 19 major errors. Using the HTA (Table 3) and the cataloged errors, we formulated metrics for video-based assessment of the LHHR (see Appendix 1).

Phase II results

Pre-survey results

Demographics A total of 38 participants were recruited to complete the fundoplication simulation (Table 4). Participants were grouped into novice ($n = 17$, 45%), intermediate ($n = 15$, 39%), and expert ($n = 6$, 16%). Additionally, 50% ($n = 19$) were male, 45% ($n = 17$) were under the age of 30, 87% ($n = 33$) self-reported being right-handed, and 58% ($n = 22$) were wearing corrective lenses.

Prior experience The overwhelming majority of novice and intermediate participants ($n = 28$, 88%) reported having observed 0–10 HHRs, while 3 (9%) reported observing 11–30 cases and 1 (3%) reported observing 30–50 cases. Among the attending surgeons, most had observed and/or participated in at least 100 cases and only 1 reported observing/participating in less than 100 cases. Overall, 42% ($n = 16$) of the participants self-reported a prior exposure to a robotic (Da Vinci) or laparoscopic (Fundamentals of Laparoscopic Surgery) simulation trainer, indicating that a subset of participants had previous hands-on engagement or familiarity with the technology being assessed. Additionally, 37%

Table 2 Task-specific metrics for assessing performance of the creation and securing the wrap portion of the laparoscopic fundoplication procedure

Metric	Score (0, 3, or 5 points)
Task-specific skills	
Wrap creation	a. Passes fundus posteriorly through retroesophageal window and performs shoeshine maneuver around the esophagus (5 points) b. Uses wrong part of stomach, or wrap is created with improper tension and orientation, or causes injury to esophagus, stomach, or vagus nerve (0 points)
Wrap position	a. Wrap on esophagus (5 points) b. Wrap on stomach (0 points)
Securing wrap	a. Three simple interrupted sutures incorporating stomach and esophagus (5 points) b. Poor knot-tying/suturing technique or injury to esophagus, stomach, or vagus nerves (0 points)
Wrap length	a. Appropriate 2–3-cm wrap length on esophagus demonstrated with ruler/grasper (5 points) b. Improper length (too short/too long; 0 points)
General laparoscopic skills	
Depth perception	a. Accurately directs instruments in the correct plane to target (5 points) b. Some overshooting or missing of target, but quick to correct (3 points) c. Constantly overshoots target, wide swings, slow to correct (0 points)
Bimanual dexterity	a. Expertly uses both hands in a complimentary manner to provide optimal exposure (5 points) b. Uses both hands, but does not optimize interaction between hands (3 points) c. Uses only one hand, ignores non-dominant hand, poor coordination between hands (0 points)
Efficiency	a. Confident, efficient, and safe conduct, maintains focus on task until it is better performed by way of an alternative approach (5 points) b. Slow, but planned movements are reasonably organized (3 points) c. Uncertain, inefficient efforts, many tentative movements, constantly changing focus or persisting without progress (0 points)
Tissue handling	a. Handles tissues well, applies appropriate traction, negligible injury to adjacent structures (5 points) b. Handles tissues reasonably well, minor trauma to adjacent tissue (occasional unnecessary bleeding or instrument slipping; 3 points) c. Rough movements, tears tissues, injures adjacent structures, poor instrument control, instruments frequently slip (0 points)
Knowledge of procedure	a. Demonstrates familiarity of all aspects of the procedure (5 points) b. Knew important steps of procedure only (3 points) c. Insufficient knowledge of procedure and instruments (0 points)
Flow of operation	a. Obviously planned operation with clear anticipation of next moves (5 points) b. Demonstrated some forward planning and reasonable progression of procedure (3 points) c. Frequently stopped operating and unsure of next move (0 points)
Instrument handling	a. Fluid movements (5 points) b. Competent use of instruments but occasionally awkward (3 points) c. Tentative and awkward moves with instruments with frequent collisions (0 points)
Bite size	a. Appropriate bite size (5 points) b. Improper bites (too small/big; 0 points)
Knot tying	a. Appropriate technique (no air knots or suture breakage; 5 points) b. Improper technique (air knots, suture breakage; 0 points)
Needle handling	a. Equidistant placement of sutures without tissue injury or suture breakage (5 points) b. Poor spacing of sutures with minor trauma to tissue and rare suture breakage (3 points) c. Frequent suture breakage and poor control, tearing tissue (0 points)

($n = 14$) reported having gaming experience, with more than half of them ($n = 10$) playing at least 1–5 h a week. None of the participants included in the study reported any exposure to VR laparoscopic training.

Post-simulation survey results

After the Nissen fundoplication task, we conducted a post-simulation survey in which participants rate the realism and usefulness of their experience on a scale of 1–5, with 1 being not realistic/useful and 5 being very realistic/useful. The survey questions covered 5 categories that included the realism of the anatomy of the model, the realism of the ex vivo porcine model (texture), the realism of the simulator interface (instruments, display), the overall realism of the task compared to the actual surgical task, and the overall perceived usefulness of the simulator for learning laparoscopic hiatal hernia surgical skills. Table 5 shows the

Table 3 Hierarchical task analysis of the laparoscopic hiatal hernia repair showing major tasks, sub-tasks, and errors

Tasks	Sub-tasks	Errors
1. Patient positioning	1.1 Place patient in supine/split leg and then reverse Trendelenburg position	1. Patient placed in an incorrect position
2. Port placement	2.1 Establish pneumoperitoneum 2.2 Inspect site of next trocar incision, make incision, and place desired ports 2.3 Expose hiatus with left lobe liver retractor	1. Air embolism 2. Injury to nearby structures
3. Abdominal hernia sac dissection	3.1 Reduce herniated stomach/omentum/other structures if present using hand-over-hand gentle reduction of hernia and contents and the use of tension/countertension to perform dissection 3.2 Divide pars flaccida/gastrohepatic ligament until base of the right crus 3.3 Divide short gastric until base of left crus 3.4 Perform blunt dissection of phrenoesophageal membrane/medial border of crura 3.5 Perform circumferential dissection to create retroesophageal window	1. Perforation of stomach 2. Injury to replaced vessels and vagus nerves 3. Thermal injury to stomach or retroperitoneal structures 4. Injury to crural fibers, aorta, inferior vena cava, and esophagus
4. Mediastinal dissection	4.1 Identify gastroesophageal junction (GEJ) using either a bougie, endoscopy, or anatomic landmarks (fat pad) 4.2 Perform blunt proximal circumferential dissection along the areolar place as high as possible up to aortic arch and up to inferior pulmonary vein 4.3 Measure intraabdominal esophageal length using either a ruler or the grasper tip 4.4 Excise hernia sac to expose angle of His	1. Esophageal perforation 2. Injury to celiac axis, esophagus, gastric wall, vagus nerve, aorta, axygos vein, lymphatics, pleura, or pericardium 3. CO_2 pneumothorax 4. Thermal spread to esophagus if using energy device
5. Crural closure	5.1 Place posterior sutures using Ethibond/silk sutures posteriorly about 1 cm apart without angulating esophagus 5.2 Place anterior sutures if there is esophageal displacement/angulation 5.3 Check repair with bougie or endoscope 5.4 If there is tension, do right crus relaxing incision	1. Esophageal perforation 2. Injury to aorta, inferior vena cava, esophagus, and phrenic nerve
6. Fundoplication	6.1 Divide additional short gastric vessels and any posterior esophageal attachments 6.2 Pass fundus posteriorly through retroesophageal window 6.3 Perform shoeshine maneuver 6.4 Retract esophagus caudad with or without a Penrose 6.5 If Nissen (360°), place three simple interrupted sutures, one stomach–stomach and two stomach–esophagus–stomach, above GEJ 6.6 If Toupet (270°), place three simple interrupted sutures, stomach–esophagus, on each side above GEJ 6.7 Perform gastropexy if needed	1. Wrapped herniation 2. Wrapped on stomach instead of esophagus 3. Internal hernia (posterior stomach through potential space by the wrap) 4. Twisted, tight/loose wrap 5. Injury to anterior vagus nerve, liver, and esophagus
7. Closure	7.1 Perform endoscopy to assess wrap and repair 7.2 Remove ports under direct visualization 7.3 Close port sites	1. Not assessing wrap and repair

survey results for the degree of realism and usefulness of the fundoplication simulation model. The vast majority of participants from all three groups rated the simulator's realism aspects highly, recognizing its usefulness and capability to capture the essential features of the task, thus establishing the content alignment.

Reliability analysis

The IRR between the two blinded raters was good for both the global- (ICC = 0.84, 95% CI 0.79–0.87, $p < 0.001$) and task-specific metrics (ICC = 0.75, 95% CI 0.7–0.78, $p < 0.001$), thereby establishing internal structure validity.

Grading the videos with blinded raters mitigated potential errors due to rater bias, thus ensuring response process validity.

Analysis of metrics

The descriptive statistics of the metrics used for assessing performance are shown in Table 6. Due to the unequal

Table 4 Demographics of the participants

	Expertise level				Total
	Novice	Intermediate		Expert	
PGY level (n)	PGY 1 (11) PGY 2 (6)	PGY 3 (4) PGY 4 (4)	PGY 5 (6) Fellow (1)	Attending (6)	
Number of participants	17	15		6	38
Sex, female, *n* (%)	7 (41)	8 (53)		4 (67)	19 (50)
Age, mean (SD), years	29 (2)	31 (2)		43 (7)	32 (6)
Race, white, *n* (%)	9 (53)	12 (80)		5 (83)	26 (68)
Ethnicity, Hispanic, *n* (%)	5 (29)	0		0	5 (13)
Dexterity, right-handed, *n* (%)	15 (88)	12 (80)		6 (100)	33 (87)
Corrective lenses, yes, *n* (%)	9 (53)	8 (53)		5 (83)	22 (58)

Table 5 Survey completed after performing the Nissen fundoplication simulation on the porcine model

Score from 1 (not realistic) to 5 (very realistic)	1/5	2/5	3/5	4/5	5/5
Realism of the anatomy of the model, *n* (%)	0	0	5 (13)	19 (50)	14 (37)
Realism of the model (texture), *n* (%)	0	0	3 (8)	18 (47)	17 (45)
Realism of the simulator interface (for instrument, display), *n* (%)	0	1 (3)	7 (18)	16 (42)	14 (37)
Overall realism of the task compared to the actual surgery, *n* (%)	0	1 (3)	8 (21)	18 (47)	11 (29)
Overall usefulness of the simulator in learning LHHR skills, *n* (%)	0	1 (3)	3 (8)	8 (21)	26 (68)

Table 6 Median and interquartile range (IQR) of metrics used for the assessment of performance

Metric	Group	Median (IQR)	Kruskal–Wallis test (*p* value)	Post hoc effect size η^2 $0.01 \leq$ Small < 0.06 $0.06 \leq$ Moderate < 0.14 Large: $> = 0.14$
Total global score	Novice	12 (6.5)	< 0.001	0.645
	Intermediate	20 (4.75)		
	Expert	28 (2.87)		
Task-specific score	Novice	32 (26)	< 0.001	0.47
	Intermediate	52 (11.25)		
	Expert	64.75 (3)		
Number of attempts	Novice	1 (1)	0.001	0.301
	Intermediate	2 (0)		
	Expert	2 (0)		
Number of sutures placed	Novice	3 (1)	0.62	0.03
	Intermediate	3 (0)		
	Expert	3 (0)		
Sum of distance between sutures	Novice	1.4 (1.1)	0.04	0.115
	Intermediate	2 (0.5)		
	Expert	2.25 (1.25)		
Seromuscular bite	Novice	0 (1)	0.03	0.141
	Intermediate	1 (1)		
	Expert	1 (0)		

sample size of the groups and data violating normality using the Shapiro–Wilk test, non-parametric tests were used and are reported here.

Global metrics

Table 6 presents the median and interquartile range of the total global scores for all three groups. The Kruskal–Wallis test showed a significant difference in performance between the groups ($\chi^2 = 24.01$, $p < 0.001$). As depicted in Fig. 2, performance improved with increasing level of expertise. Post hoc analysis revealed significant differences among all three groups: novice vs. intermediate ($p = 0.001$), intermediate vs. expert ($p = 0.01$), and novice vs. expert ($p = 0.007$).

Task-specific metrics

The median and interquartile range of the total task-specific scores for all three groups are shown in Table 6. The Kruskal–Wallis test revealed a significant difference in performance among the groups ($\chi^2 = 18.4$, $p < 0.001$). As illustrated in Fig. 3 and mirroring the total global score, performance improved with increasing levels of experience. Post hoc analysis showed a significant difference in performance among all three groups: novice vs. intermediate ($p = 0.001$), intermediate vs. expert ($p = 0.03$), and novice vs. expert ($p = 0.001$). The Spearman rank correlation indicated a strong association between the total global score and the total task-specific scores (rs = 0.87, $p < 0.001$), as depicted in Fig. 4. In addition, Fig. 5 displays photos of subjects executing various components of the task-specific metrics.

In situ metrics

I. Number of attempts: all of the participants in the expert and intermediate groups were able to complete the maximum of 2 attempts in the allotted time except for 1 subject each in both groups; whereas, in the novice group, only 6 out of 17 subjects were able to proceed to the second attempt. The Kruskal–Wallis test showed a significant difference in the number of attempts between the groups ($\chi^2 = 12.5$, $p = 0.001$). Post hoc analysis showed a significant difference between the novice and intermediate groups ($p = 0.002$). No difference was found between the novice and expert group ($p = 0.07$) and the intermediate and expert group ($p = 0.54$).

II. Number of sutures: all the subjects in the expert group placed 3 sutures to complete the fundoplication. In the intermediate group, 13 subjects placed 3

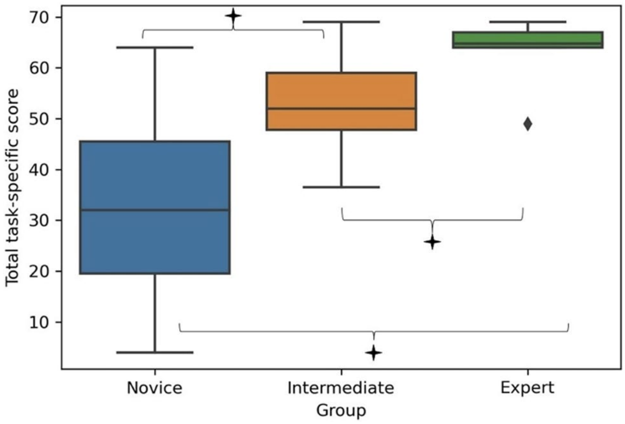


Fig. 3 Total task-specific score for the three groups

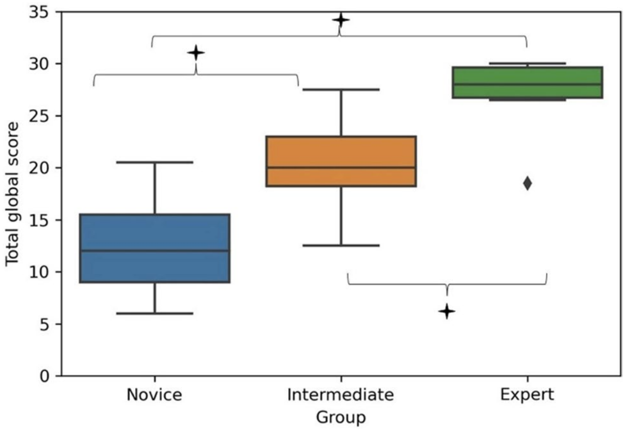


Fig. 2 Total global score for the three groups

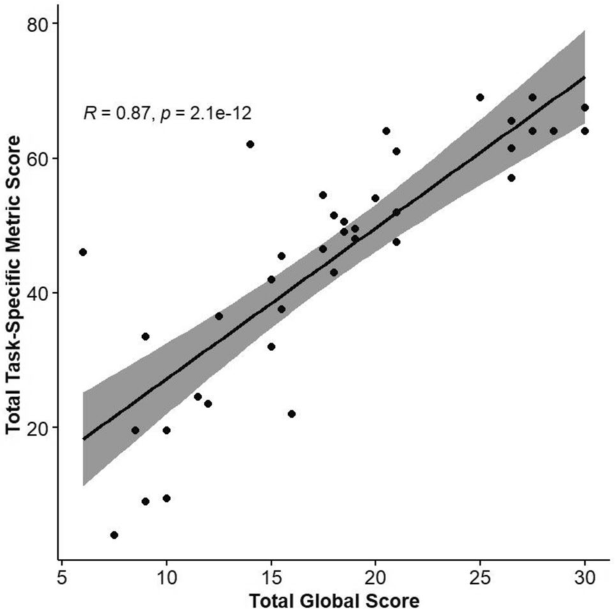
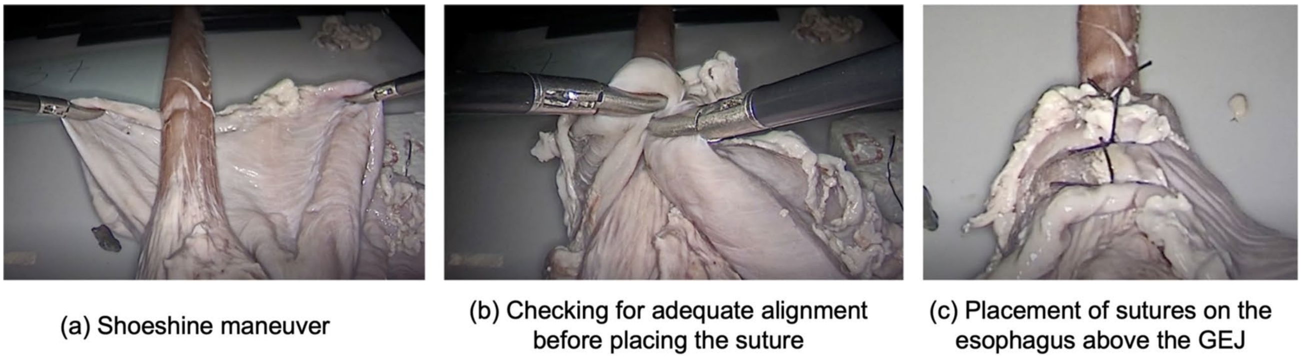


Fig. 4 Correlation between total global- and task-specific scores


(a) Shoeshine maneuver

(b) Checking for adequate alignment before placing the suture

(c) Placement of sutures on the esophagus above the GEJ

Fig. 5 Video-based assessment of the fundoplication task

sutures and 2 placed 4 sutures. In the novice group, 3 placed only 1 suture, 2 placed 2 sutures, 8 placed 3 sutures, and 4 placed 4 sutures. Kruskal–Wallis tests showed no significant difference in the number of sutures placed between the three groups ($\chi^2 = 0.94$, $p = 0.62$).

III. Sum of distance between sutures: experts sum of distance ranged from 2 to 3.5 cm, the intermediate group sum ranged from 1.4 to 4.4 cm, and the novice group sum ranged from 0 to 3.5 cm. The Kruskal–Wallis test showed significant differences between the groups in the sum of the distances for all the sutures ($\chi^2 = 6.04$, $p = 0.04$). Post hoc analysis could not find any significant differences between novice and intermediate groups ($p = 0.15$), novice and expert groups ($p = 0.08$), and between intermediate and expert groups ($p = 0.08$).

IV. Seromuscular bite: overall, 5 out of 6 subjects in the expert group, 10 out of 15 subjects in the intermediate group, and 5 out of 17 subjects in the novice group placed a seromuscular bite on the esophagus while performing fundoplication. The Kruskal–Wallis test showed a significant difference between the groups in seromuscular bite taken during fundoplication ($\chi^2 = 6.94$, $p = 0.03$). Post hoc analysis could not find any differences between novice and intermediate groups ($p = 0.06$), novice and expert groups ($p = 0.06$), and between intermediate and expert groups ($p = 0.49$).

The results from the analysis revealed the metrics' relationship to other variables, confirming construct validity.

Discussion

Our results demonstrate that task-specific metrics differentiate the performance in the wrap creation step of the laparoscopic fundoplication between novice, intermediate, and expert surgeons. A strong positive correlation was also observed between the validated global OSATS score and our task-specific scores. High IRR for both metrics established the feasibility of using our task-specific metrics for video-based assessment of performance. Additionally, it is noteworthy that 89% of participants rated the simulator's usefulness as either 4 or 5 on a scale of 5. This rating was further supported by informal comments from several non-expert participants throughout the study expressing their desire for this practice opportunity before performing the procedure in the operating room. Many trainees also mentioned how the experience enhanced their confidence when approaching such cases involving live patients.

Expertise in laparoscopic hiatal hernia surgery requires extensive training with high case volume. Learning curve studies have shown that for individual surgeons, a total of 20–40 cases, and for individual institutions, about 50 cases, are needed for stabilization of postoperative complication rates [13, 31]. In a 10-year institutional learning curve study, it was found that 200 fundoplication cases had to be performed before operative time, conversion rates, and complications plateaued [14]. Given the procedure's long learning curve, obtaining adequate training is further complicated by a substantial number of cases performed in high-volume centers, indicating centralization of this procedure to a few specialty centers [32]. This can affect the number of cases performed by residents, whose training pathways in complex foregut surgery are limited to their experience in the operating room. In our study, 88% of residents reported participating in 10 or fewer LHHRs. Simulation-based training can help bridge this gap by providing an opportunity for trainees to practice this task outside of the operating room.

As the exposure of surgical trainees to LHHR varies based on whether or not they are at a high-volume center, a simulator for training in this procedure is essential. Such a simulator should not only be capable of training the important cognitive and technical aspects of this procedure but should also be capable of both high-stakes summative and low-stakes formative assessment of skills. Several tools exist for video-based assessment of performance in LHHR with limited validity evidence [33]. A majority of training programs use a global tool for assessment of laparoscopic performance, such as the OSATS and the Global Operative Assessment of Laparoscopic Surgery (GOALS) [34–37] or a combination of global scales and procedure-specific assessment tools in the form of checklists [38, 39]. In a study by Peyre et al. [40], investigators focused on a detailed 65-step procedural checklist previously developed based on task analysis for the evaluation of technical performance in laparoscopic Nissen fundoplication [41]. Sixty-four of the 65 steps showed high degree of reliability (>0.8) when expert operative performance of Nissen fundoplication was graded by five surgeons using the checklist. More recently, as part of its master's program, SAGES developed a video-based assessment tool for laparoscopic fundoplication and demonstrated its content validity [22]. In our work, we independently developed metrics for assessment using the well-established HTA method. Overall, major tasks and subtasks aligned with prior HTA findings for this procedure [20, 22, 41]. Using HTA, we identified 19 major errors and developed task-specific metrics to evaluate performance for LHHR. Such task-specific metrics developed using HTA and expert consensus have been validated for the assessment of performance in endotracheal intubation and colorectal anastomosis procedures [27, 28, 42]. Though only the task-specific metrics for the creation and securing the wrap portion of the procedure were tested in our work, we were able to clearly establish validity evidence in the following domains defined by Messick's unitary framework, namely, content alignment, response process, internal structure, and relationship to other variables.

One unique aspect of this study was the incorporation of in situ metrics in addition to our task-specific metrics for assessment. Both the number of attempts and placement of seromuscular bite were found to be useful metrics, which could be easily incorporated in the VR simulator for assessment. Though the goal of the work was to develop assessment metrics to incorporate in our VR simulator, the developed metrics with their validity evidence can also be used for video-based assessment of performance in laparoscopic fundoplication procedures. We showed the relationship of our metrics to other variables by comparing our task-specific metrics to OSATS but due to constraints in time in performing video-based assessment, it is not yet known how our task-specific metrics correlate with other instruments developed for this procedure, which will be part of our future work.

The transferability of skills from simulation to live OR must be a priority when creating a simulator. Transferability would both encourage usage and result in an actual improvement in live operative technical skills and patient outcomes. Although we did not test the initial dissection and reduction of the hernia sac with its contents and the assessment of intraabdominal esophageal length due to constraints in creating a physical model, we plan to test those aspects later in a VR model. We have created a model of the crura with an enlarged esophageal hiatus and are performing studies to establish validity of the metrics for the crural repair portion of the procedure, which will be reported separately. Our fundoplication simulation closely mimics a portion of the actual LHHR operation with a few differences, namely the simulation's lack of a diaphragm; hence, it does not replicate the exact constraints experienced in the real surgery. The realism of our simulation is evident from the feedback of the participants, 79% of whom graded it 4 or 5 on a 5-point scale of realism.

Limitations of this study include a relatively small sample size and varying participant numbers across the groups. While we were able to maintain sufficient representation from each level of surgical expertise, the intermediate and expert groups had comparatively fewer participants. This could be attributed to the escalating operative and clinical responsibilities associated with each PGY level, leaving less availability for participating in research studies. The smaller and unequal sample size also resulted in small or moderate effect size with no clear post hoc comparison results for our in situ metrics. Furthermore, despite blinding of the identities of participants in the videos, there might still be some reluctance and apprehension regarding skill evaluation among participants. Finally, due to resource constraints, we could not use the flexible endoscopy in our study to assess quality and securement of the wrap, such as tightness and potential full-thickness bites.

Using an ex vivo fundoplication model, this study established the validity and reliability of task-specific metrics developed for assessment of performance in the creation and securing the wrap portion of the LHHR. The developed simulator and the video-based assessment metrics can be used for training and assessment in this procedure. Our next step is to incorporate the validated task-specific metrics in our VR simulator for automated assessment.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00464-024-10675-4.

Acknowledgements We acknowledge support from the UT Southwestern Simulation Center for the conduct of this study. The authors thank Dave Primm of the UT Southwestern Department of Surgery for help in editing this article.

Funding This work was funded by a grant from NIH/NIBIB R01: EB025247.

Declarations

Disclosures Alexis Desir, Carolina Marques, Emile Farah, Shruti Hegde, Carla Holcomb, Daniel J. Scott, and Ganesh Sankaranarayanan have nothing to disclose.

References

1. St John A, Caturegli I, Kubicki NS, Kavic SM (2020) The rise of minimally invasive surgery: 16 year analysis of the progressive replacement of open surgery with laparoscopy. JSLS 24(4):e202000076
2. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJO (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
3. Stulberg JJ, Stulberg JJ, Huang R, Kreutzer L, Ban K, Champagne BJ, Steele SR, Johnson JK, Holl JL, Greenberg CC, Bilimoria KY (2020) Association between surgeon technical skills and patient outcomes. JAMA Surg 155(10):960–968
4. Scott DJ, Bergen PC, Rege RV, Laycock R, Tesfay ST, Valentine RJ, Euhus DM, Jeyarajah DR, Thompson WM, Jones DB (2000) Laparoscopic training on bench models: better and more cost effective than operating room experience? J Am Coll Surg 191:272–283
5. Ritter EM, Scott DJ (2007) Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery. Surg Innov 14(2):107–112
6. Stefanidis D, Korndorffer JR, Markley S, Sierra R, Scott DJ (2006) Proficiency maintenance: impact of ongoing simulator training on laparoscopic skill retention. J Am Coll Surg 202(4):599–603
7. Korndorffer JR, Dunne JB, Sierra R, Stefanidis D, Touchard CL, Scott DJ (2005) Simulator training for laparoscopic suturing using performance goals translates to the operating room. J Am Coll Surg 201(1):23–29
8. Fried GM, Feldman LS, Vassiliou MC, Fraser SA, Stanbridge D, Ghitulescu G, Andrew CG (2004) Proving the value of simulation in laparoscopic surgery. Ann Surg 240(3):518;discussion 525–528
9. Scott DJ, Dunnington GL (2008) The new ACS/APDS skills curriculum: moving the learning curve out of the operating room. J Gastrointest Surg 12:213–221
10. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84(2):273–278
11. Roman S, Kahrilas PJ (2014) The diagnosis and management of hiatus hernia. BMJ. https://doi.org/10.1136/bmj.g6154
12. Targarona EM, Grisales S, Uyanik O, Balague C, Pernas JC, Trias M (2013) Long-term outcome and quality of life after laparoscopic treatment of large paraesophageal hernia. World J Surg 37:1878–1882
13. Watson DI, Baigrie RJ, Jamieson GG (1996) A learning curve for laparoscopic fundoplication. Definable, avoidable, or a waste of time? Ann Surg 224(2):198
14. Zacharoulis D, O'Boyle CJ, Sedman PC, Brough WA, Royston CMS (2006) Laparoscopic fundoplication: a 10-year learning curve. Surg Endosc Interv Tech 20:1662–1670
15. Gill J, Booth MI, Stratford J, Dehn TCB (2007) The extended learning curve for laparoscopic fundoplication: a cohort analysis of 400 consecutive cases. J Gastrointest Surg 11:487–492
16. Satava RM (1993) Virtual reality surgical simulator. Surg Endosc 7:203–205
17. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, Satava RM (2002) Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 236:458-463;discussion 463–454
18. Hashimoto DA, Sirimanna P, Gomez ED, Beyer-Berjot L, Ericsson KA, Williams NN, Darzi A, Aggarwal R (2015) Deliberate practice enhances quality of laparoscopic surgical performance in a randomized controlled trial: from arrested development to expert performance. Surg Endosc Other Interv Tech 29:3154–3162
19. Cristancho SM (2008) Quantitative modelling and assessment of surgical motor actions in minimally invasive surgery. Doctor of Philosophy - PhD thesis, University of British Columbia, Vancouver
20. MacKenzie L, Ibbotson JA, Cao CGL, Lomax AJ (2001) Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Minim Invasive Ther Allied Technol 10:121–127
21. Sarker SK, Chang A, Albrani T, Vincent C (2008) Constructing hierarchical task analysis in surgery. Surg Endosc Other Interv Tech 22:107–111
22. 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 34:3176–3183
23. MacKenzie L, Caroline GLC, Ibbotson JA, Alan JL (2001) Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Minim Invasive Ther Allied Technol 10(3):121–127
24. Messick S (1994) Validity of phsychological assessment: vallidation of inferences from persons' responses and performances as scientific inquirey into score meaning. Res Rep 45:1–28
25. de Montbrun S, Roberts PL, Satterthwaite L, MacRae H (2016) Implementing and evaluating a national certification technical skills examination. Ann Surg 264(1):1–6
26. De Montbrun SL, Roberts PL, Lowry AC, Ault GT, Burnstein MJ, Cataldo PA, Dozois EJ, Dunn GD, Fleshman J, Isenberg GA, Mahmoud NN, Reznick RK, Satterthwaite L, Schoetz D, Trudel JL, Weiss EG, Wexner SD, MacRae H (2013) A novel approach to assessing technical competence of colorectal surgery residents: the development and evaluation of the colorectal objective structured assessment of technical skill (COSATS). Ann Surg 258(6):1001–1006
27. Sankaranarayanan G, Parker LM, Khan A, Dials J, Demirel D, Halic T, Crawford A, Kruger U, De S, Fleshman JW (2022) Objective metrics for hand-sewn bowel anastomoses can differentiate novice from expert surgeons. Surg Endosc 37(2):1282–1292
28. Sankaranarayanan G, Parker LM, Jacinto K, Demirel D, Halic T, De S, Fleshman JW (2022) Development and validation of task-specific metrics for the assessment of linear stapler-based small bowel anastomosis. J Am Coll Surg 235:881–893
29. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med 15(2):155–163
30. Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*power 3: a flexible statistical power analysis program for the social,

behavioral, and biomedical sciences. Behav Res Methods 39:175–191

31. Neo EL, Zingg U, Devitt PG, Jamieson GG, Watson DI (2011) Learning curve for laparoscopic repair of very large hiatal hernia. Surg Endosc 25:1775–1782
32. Schlottmann F, Strassle PD, Allaix ME, Patti MG (2017) Paraesophageal hernia repair in the USA: trends of utilization stratified by surgical volume and consequent impact on perioperative outcomes. J Gastrointest Surg 21:1199–1205
33. 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 33:3798–3805
34. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, Fried GM (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190(1):107–113
35. Bilgic E, Watanabe Y, McKendy K, Munshi A, Ito YM, Fried GM, Feldman LS, Vassiliou MC (2016) Reliable assessment of operative performance. Am J Surg 211:426–430
36. Ahlberg G, Kruuna O, Leijonmarck CE, Ovaska J, Rosseland A, Sandbu R, Strömberg C, Arvidsson D (2005) Is the learning curve for laparoscopic fundoplication determined by the teacher or the pupil? Am J Surg 189(2):184–189
37. Hogle NJ, Liu Y, Ogden RT, Fowler DL (2014) Evaluation of surgical fellows' laparoscopic performance using global operative assessment of laparoscopic skills (GOALS). Surg Endosc 28:1284–1290
38. Ghaderi I, Auvergne L, Park YS, Farrell TM (2015) Quantitative and qualitative analysis of performance during advanced laparoscopic fellowship: a curriculum based on structured assessment and feedback. Am J Surg 209:71–78
39. Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J, Reznick R, Macrae HM (2004) Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. Surg Endosc 18:1800–1804
40. Peyre SE, Peyre CG, Hagen JA, Sullivan ME (2010) Reliability of a procedural checklist as a high-stakes measurement of advanced technical skill. Am J Surg 199:110–114
41. Peyre SE, Peyre CG, Hagen JA, Sullivan ME, Lipham JC, DeMeester SR, Peters JH, DeMeester TR (2009) Laparoscopic Nissen fundoplication assessment: task analysis as a model for the development of a procedural checklist. Surg Endosc 23:1227–1232
42. Ryason A, Petrusa ER, Kruger U, Xia Z, Wong VT, Jones DB, De S, Jones SB (2020) Development of an endotracheal intubation formative assessment tool. J Educ Perioper Med 22(1):E635

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.