Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Examining Tools for Assessing Technical Skills in Robotic Surgery: A Validity- Focused Systematic Review

Naimisha Badri

nsbadri@arizona.edu

The University of Arizona College of Medicine Tucson https://orcid.org/0009-0005-0044-5098

Amber Chen-Goodspeed

NewYork-Presbyterian Queens

Folarin Adeyemi

The University of Arizona College of Medicine Tucson

Iman Ghaderi

The University of Arizona College of Medicine Tucson

Research Article

Keywords:

Posted Date: June 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10037106/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Introduction Objective assessment of operative performance has become essential in competency-based surgical education. Although structured tools such as Objective Structured Assessment of Technical Skills (OSATS) are well established for open and laparoscopic training, equivalent methods for robotic surgery remain fragmented. This study reviews current tools used to assess operative performance in robotic surgery, identifying areas of strength and current gaps.

Methods A PubMed search of current robotic assessment tools was conducted. Each study identified was mapped to robotic equivalents of the 23 American College of Surgeons (ACS) and the Association of Program Directors in Surgery (APDS) skills modules. An established rating scale was applied to rate evidence strength in use of the unitary "sources of validity" framework—content, response process, internal structure, relationships to other variables, and consequences. Two independent reviewers scored each study, then came to a consensus score for each. Interrater reliability (IRR) was calculated using intraclass correlation coefficient (ICC) and weighted Cohen's κ.

Results A total of 13 of the 45 ACS/APDS skill modules had corresponding papers. The majority of the papers used previously studied assessment tools such as Global Evaluative Assessment of Robotic Skills (GEARS) and OSATS. Internal structure and response process had the highest level of evidence, while consequences had the lowest amount. Interrater reliability was substantial for content, internal structure, and relationship to other variables, and moderate for response process and consequences. Overall IRR was substantial ($\kappa = 0.754$).

Conclusion The findings indicate a lack of available assessment tools for robotic surgery with most skill modules having no corresponding papers. Future work should focus on developing assessment tools and establishing validity evidence for their use in competency-based training, especially as robotic surgery becomes increasingly integrated into surgical practice.

Introduction

Robust assessment of technical performance is central to competency-based surgical education. As training has moved beyond the traditional apprenticeship model, reliance on subjective faculty assessment alone is increasingly insufficient as measures of operative competence. Early efforts to standardize technical assessment, most notably through the Objective Structured Assessment of Technical Skills (OSATS), established that surgical performance could be assessed objectively using structured tasks, global rating scales, and reproducible scoring framework[1]. These developments marked a critical shift in surgical education by laying the groundwork for valid, transparent, and competency-focused technical assessment.

Following the introduction of OSATS, assessment of surgical technical skill expanded substantially, with increasing use of standardized and structured methods such as procedure-specific checklists, global rating scales, competency-based frameworks, and simulation-based evaluation tools[2]. As minimally invasive surgery became more prevalent, specialty-specific assessment instruments emerged to evaluate technical performance within these new operative environments. One notable example was the Global Operative Assessment of Laparoscopic Skills (GOALS), which adapted the principles of global rating scales to laparoscopic surgery and provided a structured framework for intraoperative assessment[27]. Additional instruments, including O-SCORE and other specialty-specific measures, were subsequently introduced to improve the consistency and objectivity of technical skills assessment across training contexts. Concurrently, simulation-based training and assessment programs such as Fundamentals of Laparoscopic Surgery (FLS) were introduced and increasingly incorporated into surgical education, further reinforcing the value of standardized assessment of technical performance[3].

Robotic surgery has expanded rapidly across surgical specialties and now constitutes an increasingly important component of surgical training. Unlike conventional laparoscopy, it requires distinct cognitive and psychomotor abilities, including console-based ergonomics, wristed instrument control, camera navigation, and operative performance in the absence of direct haptic feedback. Despite these unique demands, approaches to robotic training and competency assessment remain inconsistent, and technical proficiency is still often inferred from case volume rather than objectively demonstrated skills. In response, robotic-specific assessment tools such as the Global Evaluative Assessment of Robotic Skills (GEARS) were developed to provide structured

evaluation of technical performance. Although these instruments have shown promise in differentiating levels of performance and supporting competency-based training through objective feedback and performance tracking, the overall strength and breadth of validity evidence supporting their use remain uncertain[4].

Given the increasing reliance on performance assessments to evaluate technical competence and educational progression, the validity of the resulting score interpretations is critical to ensuring that decisions based on these assessments are appropriate, meaningful, and defensible.

Under the contemporary unitary framework of validity, validity does not reside within an instrument itself but in the degree to which evidence and theory support the interpretation and intended use of its scores. Accordingly, assessment tools must be evaluated through multiple sources of validity evidence, including content, response process, internal structure, relations to other variables, and consequences. As technical assessment assumes a greater role in high-stakes educational decisions, the evidentiary foundation supporting these instruments becomes increasingly consequential [5].

Despite the accelerating integration of robotic surgery into contemporary training, the landscape of robotic technical assessment and the validity evidence supporting available tools have not been comprehensively characterized. Therefore, the aim of this study was to systematically review current robotic technical assessment instruments, map them to the American College of Surgeons and Association of Program Directors in Surgery technical skills modules, and evaluate the strength of validity evidence supporting their use through an established rating scale for assessment of validity evidence. By doing so, we sought to clarify the current state of robotic technical assessment and identify critical strengths, limitations, and gaps in the literature.

Methods

Framework Development and Validity Rating

The ACS/APDS surgical skills curriculum consists of 45 total modules, of which the procedural modules are concentrated within Phase 1 and Phase 2. Specifically, these include 35 procedural modules across Phase 1 (20 basic skills modules) and Phase 2 (15 advanced skills modules). To create a parallel framework for robotic surgery, each of these ACS/APDS procedural modules was systematically mapped to a robotic equivalent procedure or skill domain, as shown in Table 1. Modules without a recognized robotic equivalent in the current literature were classified as unstudied domains, representing gaps in the existing evidence base. In total, 23 robotic equivalents of the ACS/APDS procedural skills modules were identified and used as the organizational template for this toolbox.

To evaluate the strength of evidence for these tools, an established rating scale for assessment of validity evidence, adapted from Ghaderi et. al[6] was applied. This scale rated evidence based on the unitary "sources of validity" framework, which includes five categories: Content, Response process, Internal structure, Relationships to other variables and Consequences. A rating scale was used to assign a numerical score to each of the five sources of validity evidence for every study. Scores ranged from 0 to 3, where a score of 0 indicated no discussion or data presented as evidence of validity; a score of 1 indicated limited data that weakly supported the validity of score interpretations; a score of 2 indicated an intermediate level of evidence with some supporting data but remaining gaps; and a score of 3 indicated multiple sets of data that strongly and comprehensively supported the validity of score interpretations. The criteria used to assign scores for each validity source are summarized in Table 2.

Table 1: ACS/APDS Skills Modules and Identified Robotic Equivalents

ACS/APDS Module	**Identified Robotic Equivalent**
Phase 1 Modules	
Asepsis and Instrument Handling	Asepsis and Instrument Handling
Knot Tying	Knot Tying
Suturing	Suturing
Urethral and Suprapubic Catheterization	Ø
Chest Tube Placement and Thoracentesis	Ø
Central Venous and Arterial Access	Ø
Surgical Biopsy Techniques	Ø
Arterial Anastomosis	Arterial Anastomosis
Laparotomy: Opening and Closure	Ø
Bone Fixation and Casting Principles	Ø
Upper Endoscopy	Ø
Lower Endoscopy	Ø
Hand-Sewn Anastomosis	Ø
Stapled Anastomosis	Stapled Robotic Anastomosis
Phase 2 Modules	
Laparoscopic Ventral Hernia Repair	Robotic Ventral Hernia Repair
Laparoscopic Colon Resection	Robotic Colon Resection
Laparoscopic and Open Bile Duct Exploration	Robotic Bile Duct Exploration
Abdominal Wall Stoma Creation	Robotic Wall Stoma Creation
Laparoscopic Appendectomy	Robotic Appendectomy
Laparoscopic Nissen Fundoplication	Robotic Nissen Fundoplication
Sentinel Lymph Node Biopsy and Axillary Dissection	Ø
Open Inguinal and Femoral Hernia Repair	Ø
Laparoscopic Inguinal Hernia Repair	Robotic Inguinal Hernia Repair
Laparoscopic and Open Splenectomy	Robotic Splenectomy
Laparoscopic and Open Cholecystectomy	Robotic Cholecystectomy
Thyroidectomy	Ø
Parathyroidectomy	Ø
Gastrectomy	Robotic Gastrectomy
Distal and Total Pancreatectomy	Robotic Total Pancreatectomy
Phase 3 Modules	
Laparoscopic Crisis Management	Robotic Crisis Management
Laparoscopic Troubleshooting	Robotic Troubleshooting
Latex Allergy-Induced Anaphylaxis	Ø
Patient Handoff	Ø
Postoperative Hypotension	Ø
Postoperative Myocardial Infarction and Cardiogenic Shock	Ø
Postoperative Pulmonary Embolism	Ø
Preoperative Briefing	Ø
Retained Sponge Identification on Postoperative Imaging	Ø
Trauma Team Training	Ø

Table 2. Sources of Validity and Scoring Criteria

Type	Definition[6,7]	Scoring (0–3)
Content	Whether assessment items adequately represent the construct and domain of interest, including item wording, format, and expert evaluation of content appropriateness. *e.g., test blueprint, expert review, item writer qualifications*	**0** – No discussion or data **1** – Items listed without justification **2** – Some references and justifications, or reference to a previously validated instrument **3** – Explicit theoretical basis with systematic expert item review
Response Process	Whether rater actions, judgments, and scoring procedures align with the construct. Inconsistencies may introduce variance unrelated to what is being measured. *e.g., rater training, scoring quality control, pilot study*	**0** – No discussion or data **1** – Minimal discussion; no examination of thought processes **2** – Some data on rater cognition or systems reducing scoring error **3** – Critical examination of rater thought processes with several sources of data
Internal Structure	Whether assessment components statistically align with the underlying construct, evaluated through reliability and factor analysis. *e.g., ICC/TCC, interitem correlations, factor analysis, generalizability*	**0** – No discussion or data **1** – Single reliability measure only **2** – Partial factor analysis or a few reliability measures **3** – Multiple reliability measures; item-level analysis; generalizability analysis
Relationships to Other Variables	Whether scores correlate with external variables as theoretically expected. Commonly demonstrated by comparing scores across novice, intermediate, and expert performers. *e.g., novice vs. expert discrimination, test-criterion correlations*	**0** – No discussion or data **1** – Correlation with theoretically unimportant variables **2** – Correlation with variables of some theoretical importance **3** – Convergent or divergent evidence with theoretically predicted outcomes; anticipated a priori
Consequences	Whether the use of the assessment produces intended benefits without unintended harms, informing the appropriateness of score interpretation and use. *e.g., pass/fail cut scores, learner impact*	**0** – No discussion or data **1** – Post-assessment survey on perceived usefulness only **2** – Consequences described but not explicitly tied to score validity **3** – Consequences clearly impact validity; data-supported

Search Strategy and Study Selection

A systematic review was conducted via a PubMed search to identify current robotic surgical technical skills assessment tools. The search utilized keywords including "Objective," "Assessment," "Tool," "Skill," "Evaluation," "Proficiency," and "Robot," which were combined with procedure-specific terms (e.g., "Nissen") using Boolean operators to refine results. Additionally, relevant studies were extracted from the supplementary data of a base source focused on objective tools and artificial intelligence in robotic surgery[28].

Data Extraction and Quality Control

Two reviewers independently scored each identified study according to the validity framework. Following the independent review, the two raters met to reach a consensus score for each study. A summary of the studies identified for each robotic equivalent and scores for each validity domain are shown in Table 3.

Table 3: Robotic Equivalents of ACS/APDS Modules with Associated Studies and Scores

#	Procedure	Tool	Content	Response Process	Internal Structure	Relationship to Other Variables	Consequences
1	Asepsis and Instrument Handling	Guni et al. 2018 (GEARS)[6]	3	2	3	3	1
2	Knot Tying	Guni et al. 2018 (GEARS)[6]	3	2	3	3	1
3	Suturing	Sanchez et al. 2016 (GEARS)[12]	2	2	3	2	1
		Vernez et al. 2017 (GEARS)[16]	2	2	3	2	2
		Moller et al. 2020 (OSATS)[17]	2	2	1	2	1
4	Tissue Handling, Dissection, Wound Closure	Hung et al. 2013 (GEARS, dVSS)[18]	2	2	2	3	1
		Ahmad et al. 2021 (OSATS)[13]	2	2	2	2	2
5	Advanced Tissue Handling, Flaps, Skin Grafts	Tam et al. 2017 (OSATS)[19]	2	3	3	3	1
6	Arterial Anastomosis	Ø					
7	Basic Robotic Skills	Puliatti et al. 2022 (Task-Specific)[9]	3	3	3	3	1
8	Advanced Robotic Skills	Puliatti et al. 2022 (Task-Specific)[9]	3	3	3	3	1
9	Handsewn Anastomosis	Ø					
10	Stapled Anastomosis	Schmidt et al. 2022 (A-OSATS)[10]	3	3	3	3	1
		Brocke et al. 2024 (A-OSATS)[20]	2	3	2	3	1
11	Ventral Hernia Repair	Ø					

Table 3: Robotic Equivalents of ACS/APDS Modules with Associated Studies and Scores (continued)

#	Procedure	Tool	Content	Response Process	Internal Structure	Relationship to Other Variables	Consequences
12	Colon Resection	Hung et al. 2012 (GOALS+, dVSS)[21]	2	3	3	3	1
		Panteleimonitis et al. 2018 (GAS)[15]	3	3	1	2	2
		Gómez Ruiz et al. 2022 (GEARS, Task-Specific)[11]	3	3	3	3	1
13	Bile Duct Exploration	Ø					
14	Abdominal Wall Stomas	Ø					
15	Appendectomy	Ø					
16	Nissen Fundoplication	Ø					
17	Inguinal Hernia	Nikolian et al. 2024 (IHR-OPSA)[22]	3	3	3	1	1
18	Splenectomy	Ø					
19	Cholecystectomy	Willuth et al. 2022 (OSATS)[23]	2	3	1	3	2
20	Gastrectomy	Ø					
21	Distal/Total Pancreatectomy	Hogg et al. 2016 (OSATS)[14]	2	3	3	3	3
22	Robotic Crisis	Ø					
23	Robotic Troubleshooting	Schreyer et al. 2022 (RAS-NOTECHS)[24]	3	3	3	0	1

Statistical Analysis

Data analysis was performed using R version 4.5.2[25] with the irr package[26]. R code and correlating data shown in supplementary materials. Interrater reliability (IRR) was calculated using two primary metrics:

- Weighted Cohen's κ: Utilized because of the linear scale of the ratings, with squared weights applied. This is a score that measures the degree to which two people agree when they are grading on a linear scale. We chose squared weights for this metric because our rating scale was linear, requiring a higher penalty for larger disagreements
- Intraclass Correlation Coefficient (ICC): Calculated using a two-way random, absolute agreement model.

Percent agreement was also recorded for each validity category.

Kappa values were interpreted using a standard scale, with 0.61–0.80 considered substantial and 0.81–1.00 almost perfect.

Results

Module Coverage and Tool Identification

Of the 45 ACS/APDS skill modules evaluated, only 13 (28.9%) had corresponding published papers. The majority of the identified papers utilized previously studied assessment tools, most notably the Global Evaluative Assessment of Robotic Skills (GEARS) and the Objective Structured Assessment of Technical Skills (OSATS).

Validity Evidence Levels

The strength of evidence varied significantly across the five sources of validity. Response process and internal structure demonstrated the strongest levels of validity evidence among the reviewed studies, whereas content and relationship to other variables showed moderate to strong supporting evidence. In contrast, consequences demonstrated the weakest level of evidence, suggesting that the impact of assessments on outcomes and decision-making remains comparatively underexplored within the literature. A summary of the average validity evidence scores across all five sources is presented in Table 4.

Interrater Reliability (IRR)

The overall interrater reliability for the study was substantial, with an overall weighted kappa of 0.754 and an overall ICC of 0.757. When broken down by category, reliability was highest in the Content category ($\kappa = 0.746$, ICC = 0.758) and lowest in the Consequences ($\kappa = 0.510$, ICC = 0.525) and Response Process ($\kappa = 0.514$, ICC = 0.529) categories. Results show in Table 5.

Table 4: Average Scores Across Each Validity Source

Validity Source	**Average Score**
Content	2.47
Response Process	2.63
Internal Structure	2.53
Relationship to Other Variables	2.47
Consequences	1.32

Table 5: Interrater Reliability by Validity Category

Category	Weighted Kappa	Percent Agreement	ICC
Content	0.746	88.2%	0.758
Response Process	0.514	76.5%	0.529
Internal Structure	0.728	70.6%	0.740
Relationship to Other Variables	0.643	64.7%	0.657
Consequences	0.510	64.7%	0.525
Overall	0.754	72.9%	0.757

Discussion

In this review, only a limited number of studies addressed objective assessment tools for robotic surgery, and even fewer provided comprehensive validity evidence across the five sources of validity as shown in Table 2. Most studies adapted existing assessment instruments rather than developing novel, procedure-specific tools. Furthermore, assessments frequently focused on isolated technical domains or simulator-based tasks and provided limited evaluation of multiple sources of validity.

Content

Content validity assesses whether an instrument adequately represents the construct of interest and relevant performance domain through appropriate content, item design, and expert review.

The majority of papers relied on previously studied assessment tools such as the Global Evaluative Assessment of Robotic Skills (GEARS) or the Objective Structured Assessment of Technical Skills (OSATS) rather than developing a new instrument. A few studies described processes to establish content validity during tool development. For example, Guni et al.[8] developed a checklist for robotic suturing by identifying key procedural steps from expert videos and common technical errors from novice performances before validating the checklist against GEARS scores. Similarly, Puliatti et al.[9] incorporated expert consensus through a modified Delphi process in which experienced robotic surgeons evaluated and refined performance metrics until full agreement was reached. These approaches represent stronger sources of content validity evidence because they involve systematic expert review and deliberate alignment between assessment items and the procedural domain being measured.

Response Process

Response process examines whether raters' actions, judgments, and scoring procedures align with the construct being measured and are applied consistently. In the studies included in this review, established instruments such as GEARS or OSATS were commonly used to assess surgical performance. These tools rely on Likert-style global rating scales, which depend on subjective judgment and may introduce variability based on how individual raters interpret scale anchors or domains. Some studies attempted to strengthen response process evidence by incorporating measures to reduce bias in scoring. For example, Schmidt et al.[10] used blinded video review in which two raters independently scored recorded procedures to evaluate intra- and inter-rater reliability and the ability of the assessment to discriminate between novice, intermediate, and expert surgeons. Gomez et al.[11] incorporated formal rater training, with consultant colorectal surgeons trained to apply proficiency-based progression binary metrics alongside GEARS scoring while evaluating procedural steps, errors, and critical errors; inter-rater reliability between assessors was also examined. Reliability metrics such as inter-rater agreement and Cronbach's were frequently reported to quantify consistency between raters; however, these analyses focused on statistical agreement rather than examining the cognitive processes underlying scoring.

Internal Structure

Internal structure evaluates whether assessment components align with the underlying construct, typically through analyses of reliability. Sanchez et al.[12] examined internal consistency and interobserver variability when applying the GEARS scale to robotic surgical performance, using reliability analyses to evaluate

agreement between raters and the coherence of the assessment domains. These approaches were frequently used to demonstrate that the scoring systems produced consistent results across evaluators.

Relationship to Other Variables

Relationship to other variables assesses whether scores demonstrate expected associations with external measures that are theoretically related to the construct. Evidence for relationships to other variables was commonly demonstrated by examining how assessment scores varied across levels of surgical experience or training. Several studies compared performance scores between novice, intermediate, and expert surgeons to determine whether the assessment tools were able to distinguish between expected levels of expertise.

Hung et al.[21] evaluated the relationship between simulator training and subsequent performance metrics, demonstrating that participants who received simulator training showed measurable advantages in key performance measures compared with those who did not receive training. Similarly, Ahmad et al.[13] assessed this source of validity through pre- and post-training comparisons within a robotic surgery curriculum. Fellows completed baseline assessments followed by simulator-based training and repeat assessment, allowing investigators to examine how changes in training exposure were reflected in subsequent performance scores.

Consequences

Consequences examines the intended and unintended effects of assessment use to determine whether score interpretation and application are appropriate. Consequences validity evidence was rarely examined directly. Discussion of consequences was generally limited to potential implications rather than empirical evaluation of outcomes resulting from the use of the assessment tools. Hogg et al.[14] examined the relationship between technical performance scores and postoperative outcomes, demonstrating that video-based assessments of surgical skill were associated with the risk of postoperative pancreatic fistula following pancreaticoduodenectomy. Panteleimonitis et al.[15] discussed the potential patient safety implications of robotic surgery training pathways, emphasizing the role of structured training and simulation outside the operating room during early skill acquisition. These discussions illustrate how assessment tools may influence patient outcomes and surgical training practices.

Conclusion

Assessment tools for surgical technical skills are widely used in surgical education and practice to evaluate operative performance, provide structured feedback, guide trainee progression, and support competency-based training models. As robotic surgery has become increasingly integrated into modern surgical practice, these tools are also being applied to the assessment of robotic operative skills. This review provides an overview of currently available instruments used to assess robotic surgical performance and examines the validity evidence supporting their use. Overall, the literature evaluating these instruments remains limited, and relatively few studies have examined robotic surgical skill assessment using the contemporary unitary framework of validity. As a result, many tools currently used in robotic surgical training have only partial evidence supporting the interpretation and appropriate application of their scores.

As competency-based training models continue to evolve in surgical education, additional validation work will be necessary to clarify how results from these assessment tools should be interpreted and used in training and evaluation. Future research should focus on developing and validating global and procedure-specific assessment tools for robotic surgery, including adaptation of existing laparoscopic instruments to robotic equivalents. For example, a laparoscopic cholecystectomy assessment tool could potentially be adapted to evaluate robotic cholecystectomy performance. Procedure-specific tools may better assess the critical steps, decision-making, and technical execution of an operation itself, rather than only general robotic technical ability. Additional validation work is also needed to determine how assessment scores relate to meaningful educational and clinical outcomes, including operative safety, trainee progression, and patient outcomes. Strengthening the validity evidence for robotic surgical assessment tools will be essential to ensure these instruments can be confidently used for training, feedback, and competency evaluation.

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

References:

1. O'Regan, D. J., Caddy, C., Rajandram, R., & Barach, P. (2026). A Scoping Review of the Objective Structured Assessment of Technical Surgical Skills (OSATS). *Journal of Surgical Education, 83*(6), Article 103946. https://doi.org/10.1016/j.jsurg.2026.103946

2. Levin, M., McKechnie, T., Khalid, S., Grantcharov, T. P., & Goldenberg, M. (2019). Automated Methods of Technical Skill Assessment in Surgery: A Systematic Review. *Journal of Surgical Education, 76*(6), 1629–1639. https://doi.org/10.1016/j.jsurg.2019.06.011

3. Peters, J. H., Fried, G. M., Swanstrom, L. L., Soper, N. J., Sillin, L. F., Schirmer, B., Hoffman, K., & the SAGES FLS Committee. (2004). Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. *Surgery, 135*(1), 21–27. https://doi.org/10.1016/S0039-6060(03)00156-9

4. Goh, A. C., Goldfarb, D. W., Sander, J. C., Miles, B. J., & Dunkin, B. J. (2012). Global Evaluative Assessment of Robotic Skills: Validation of a Clinical Assessment Tool to Measure Robotic Surgical Skills. *The Journal of Urology, 187*(1), 247–252. https://doi.org/10.1016/j.juro.2011.09.032

5. Goldenberg, M., & Lee, J. Y. (2018). Surgical Education, Simulation, and Simulators—Updating the Concept of Validity. *Current Urology Reports, 19*, 52. https://doi.org/10.1007/s11934-018-0799-7

6. Ghaderi, I., Manji, F., Park, Y. S., Juul, D., Ott, M., Harris, I., & Farrell, T. M. (2015). Technical skills assessment toolbox: a review using the unitary framework of validity. *Annals of Surgery, 261*(2), 251–262.

7. Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal of General Internal Medicine, 20*(12), 1159–1164.

8. Guni, A., Raison, N., Challacombe, B., Khan, S., Dasgupta, P., & Ahmed, K. (2018). Development of a technical checklist for the assessment of suturing in robotic surgery. *Surgical Endoscopy, 32*(11), 4402–4407. https://doi.org/10.1007/s00464-018-6407-6

9. Puliatti, S., Amato, M., Mazzone, E., Rosiello, G., De Groote, R., Piazza, P., Sarchi, L., Farinha, R., Mottrie, A., & Gallagher, A. G. (2022). Development and validation of the metric-based assessment of a robotic vessel dissection, vessel loop positioning, clip applying and bipolar coagulation task on an avian model. *Journal of Robotic Surgery, 16*(3), 677–685. https://doi.org/10.1007/s11701-021-01293-6

10. Schmidt, M. W., Haney, C. M., Kowalewski, K. F., Bintintan, V. V., Abu Hilal, M., Arezzo, A., Bahra, M., Besselink, M. G., Biebl, M., Boni, L., Diana, M., Egberts, J. H., Fischer, L., Francis, N., Hashimoto, D. A., Perez, D., Schijven, M., Schmelzle, M., Soltes, M., Swanstrom, L., Welsch, T., Müller-Stich, B. P., & Nickel, F. (2022). Development and validity evidence of an objective structured assessment of technical skills score for minimally invasive linear-stapled, hand-sewn intestinal anastomoses: the A-OSATS score. *Surgical Endoscopy, 36*(6), 4529–4541. https://doi.org/10.1007/s00464-021-08806-2

11. Gómez Ruiz, M., Tou, S., Gallagher, A. G., Cagigas Fernández, C., Cristobal Poch, L., & Matzel, K. E. (2022). Intraoperative robotic-assisted low anterior rectal resection performance assessment using procedure-specific binary metrics and a global rating scale. *BJS Open, 6*(3), zrac041. https://doi.org/10.1093/bjsopen/zrac041

12. Sánchez, R., Rodríguez, O., Rosciano, J., Vegas, L., Bond, V., Rojas, A., & Sanchez-Ismayel, A. (2016). Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS). *Journal of Robotic Surgery, 10*(3), 227–231. https://doi.org/10.1007/s11701-016-0572-1

13. Ahmad, S. B., Rice, M., Chang, C., Hamad, A., Kingham, T. P., He, J., Pimiento, J. M., Zureikat, A. H., Zeh, H. J., III, & Hogg, M. E. (2021). Will It Play in Peoria? A Pilot Study of a Robotic Skills Curriculum for Surgical Oncology Fellows. *Annals of Surgical Oncology, 28*(11), 6273–6282. https://doi.org/10.1245/s10434-021-09913-z

14. Hogg, M. E., Zenati, M., Novak, S., Chen, Y., Jun, Y., Steve, J., Kowalsky, S. J., Bartlett, D. L., Zureikat, A. H., & Zeh, H. J., III. (2016). Grading of Surgeon Technical Performance Predicts Postoperative

Pancreatic Fistula for Pancreaticoduodenectomy Independent of Patient-related Variables. *Annals of Surgery, 264*(3), 482–491. https://doi.org/10.1097/SLA.0000000000001862

15. Panteleimonitis, S., Popeskou, S., Aradaib, M., Harper, M., Ahmed, J., Ahmad, M., Qureshi, T., Figueiredo, N., & Parvaiz, A. (2018). Implementation of robotic rectal surgery training programme: importance of standardisation and structured training. *Langenbeck's Archives of Surgery, 403*(6), 749–760. https://doi.org/10.1007/s00423-018-1690-1

16. Vernez, S. L., Huynh, V., Osann, K., Okhunov, Z., Landman, J., & Clayman, R. V. (2017). C-SATS: Assessing Surgical Skills Among Urology Residency Applicants. *Journal of Endourology, 31*(Suppl 1), S95–S100. https://doi.org/10.1089/end.2016.0569

17. Møller, S. G., Dohrn, N., Brisling, S. K., Larsen, J. C. R., & Klein, M. (2020). Laparoscopic Versus Robotic-assisted Suturing Performance Among Novice Surgeons: A Blinded, Cross-Over Study. *Surgical Laparoscopy, Endoscopy & Percutaneous Techniques, 30*(2), 117–122.

18. Hung, A. J., Jayaratna, I. S., Teruya, K., Desai, M. M., Gill, I. S., & Goh, A. C. (2013). Comparative assessment of three standardized robotic surgery training methods. *BJU International, 112*(6), 864–871. https://doi.org/10.1111/bju.12045

19. Tam, V., Zenati, M., Novak, S., Chen, Y., Zureikat, A. H., Zeh, H. J., III, & Hogg, M. E. (2017). Robotic Pancreatoduodenectomy Biotissue Curriculum has Validity and Improves Technical Performance for Surgical Oncology Fellows. *Journal of Surgical Education, 74*(6), 1057–1065. https://doi.org/10.1016/j.jsurg.2017.05.016

20. Brocke, T. K., Fox, C., Clanahan, J. M., Klos, C. L., Chapman, W. C., Jr., Wise, P. E., Awad, M. M., & Ohman, K. A. (2024). Extrapolative Validity Evidence of the Anastomosis Objective Structured Assessment of Technical Skill (A-OSATS) for Robotic Ileocolic Anastomosis. *Journal of Surgical Education, 81*(11), 1577–1584. https://doi.org/10.1016/j.jsurg.2024.07.021

21. Hung, A. J., Patil, M. B., Zehnder, P., Cai, J., Ng, C. K., Aron, M., Gill, I. S., Desai, M. M., & Goh, A. C. (2012). Concurrent and predictive validation of a novel robotic surgery simulator: a prospective, randomized study. *Journal of Urology, 187*(2), 630–637. https://doi.org/10.1016/j.juro.2011.09.154

22. Nikolian, V. C., Camacho, D., Earle, D., Lehmann, R., Nau, P., Ramshaw, B., & Stulberg, J. (2024). Development and preliminary validation of a new task-based objective procedure-specific assessment of inguinal hernia repair procedural safety. *Surgical Endoscopy, 38*(3), 1583–1591. https://doi.org/10.1007/s00464-024-10677-2

23. Willuth, E., Hardon, S. F., Lang, F., Haney, C. M., Felinska, E. A., Kowalewski, K. F., Müller-Stich, B. P., Horeman, T., & Nickel, F. (2022). Robotic-assisted cholecystectomy is superior to laparoscopic cholecystectomy in the initial training for surgical novices in an ex vivo porcine model: a randomized crossover study. *Surgical Endoscopy, 36*(2), 1064–1079. https://doi.org/10.1007/s00464-021-08373-6

24. Schreyer, J., Koch, A., Herlemann, A., Becker, A., Schlenker, B., Catchpole, K., & Weigl, M. (2022). RAS-NOTECHS: validity and reliability of a tool for measuring non-technical skills in robotic-assisted surgery settings. *Surgical Endoscopy, 36*(3), 1916–1926. https://doi.org/10.1007/s00464-021-08474-2

25. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2025. Available at: https://www.R-project.org/

26. Gamer M, Lemon J, Singh IFP. *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1; 2019. doi:10.32614/CRAN.package.irr.

27. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, Fried GM. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg. 2005 Jul;190(1):107-13. doi: 10.1016/j.amjsurg.2005.04.004. PMID: 15972181.

28. Boal MWE, Anastasiou D, Tesfai F, Ghamrawi W, Mazomenos E, Curtis N, Collins JW, Sridhar A, Kelly J, Stoyanov D, Francis NK. Evaluation of objective tools and artificial intelligence in robotic

surgery technical skills assessment: a systematic review. Br J Surg. 2024 Jan 3;111(1):znad331. doi: 10.1093/bjs/znad331. PMID: 37951600; PMCID: PMC10771126.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials1.docx
- Supplementarymaterials2dataforR.xlsx