Received: 17 January 2024 | Revised: 28 February 2024 | Accepted: 19 March 2024

DOI: 10.1111/aogs.14840

AOGS Acta Obstetricia et Gynecologica Scandinavica

SYSTEMATIC REVIEW

Objective assessment tools in laparoscopic or robotic-assisted gynecological surgery: A systematic review

Freweini Martha Tesfai[1,2,3] | Jasleen Nagi[4] | Iona Morrison[5] | Matt Boal[1,2,3] | Adeola Olaitan[6] | Dhivya Chandrasekaran[2,7] | Danail Stoyanov[2,3] | Anne Lanceley[2] | Nader Francis[1,2,5]

[1]The Griffin Institute, Northwick Park & St Marks' Hospital, London, UK

[2]EGA Institute for Women's Health, University College London, London, UK

[3]Wellcome/EPSRC Center for Interventional and Surgical Sciences (WEISS), University College London, London, UK

[4]Imperial College London, London, UK

[5]Yeovil District Hospital, Somerset Foundation NHS Trust, Yeovil, UK

[6]The Wellington Hospital, HCA Hospitals, London, UK

[7]Department of Gynecological Oncology, University College of London Hospitals, London, UK

Correspondence
Nader Francis, The Griffin Institute, Northwick Park and St Mark's Hospital, Y Block, Watford Rd, Harrow HA1 3UJ, UK.
Email: n.francis@griffininstitute.org.uk

Abstract

Introduction: There is a growing emphasis on proficiency-based progression within surgical training. To enable this, clearly defined metrics for those newly acquired surgical skills are needed. These can be formulated in objective assessment tools. The aim of the present study was to systematically review the literature reporting on available tools for objective assessment of minimally invasive gynecological surgery (simulated) performance and evaluate their reliability and validity.

Material and methods: A systematic search (1989–2022) was conducted in MEDLINE, Embase, PubMed, Web of Science in accordance with PRISMA. The trial was registered with the Prospective Register of Systematic Reviews (PROSPERO) ID: CRD42022376552. Randomized controlled trials, prospective comparative studies, prospective single-group (with pre- and post-training assessment) or consensus studies that reported on the development, validation or usage of assessment tools of surgical performance in minimally invasive gynecological surgery, were included. Three independent assessors assessed study setting and validity evidence according to a contemporary framework of validity, which was adapted from Messick's validity framework. Methodological quality of included studies was assessed using the modified medical education research study quality instrument (MERSQI) checklist. Heterogeneity in data reporting on types of tools, data collection, study design, definition of expertise (novice vs. experts) and statistical values prevented a meaningful meta-analysis.

Results: A total of 19 746 titles and abstracts were screened of which 72 articles met the inclusion criteria. A total of 37 different assessment tools were identified of which 13 represented manual global assessment tools, 13 manual procedure-specific assessment tools and 11 automated performance metrices. Only two tools showed substantive evidence of validity. Reliability and validity per tool were provided. No

Abbreviations: APM, automated performance metric; MIGS, minimally invasive gynecological surgery; OSATS, objective structured assessment of technical skills; VR, virtual reality.

assessment tools showed direct correlation between tool scores and patient related outcomes.
Conclusions: Existing objective assessment tools lack evidence on predicting patient outcomes and suffer from limitations in transferability outside of the research environment, particularly for automated performance metrics. Future research should prioritize filling these gaps while integrating advanced technologies like kinematic data and AI for robust, objective surgical skill assessment within gynecological advanced surgical training programs.

KEYWORDS
minimally invasive gynecological surgery, objective assessment tools, surgical training

1 | INTRODUCTION

Minimally invasive surgery (MIS) in gynecology has a prominent role in the management of gynecological benign and oncological diagnoses. MIS reduces hospital stay and enhances postoperative recovery, making it one of the preferred routes of operation in many diagnoses.[1] In the last two decades, robotic-assisted laparoscopic surgery has emerged as a new entity within MIS.[2] However, with the introduction of new medical techniques and devices comes the risk of increased errors and unknown consequences.[3] In addition and distinct from open surgery, laparoscopic surgery requires specific surgical skills and endoscopic psychomotor skills to ensure patient safety.[4] Especially in laparoscopic surgery, depth perception is hindered and tactile feedback is reduced. Minimal movements are amplified and range of motion is decreased due to fixation of the trocars.[5] There is increasing evidence that simulation-based training and assessment such as lower fidelity physical/box video training and higher fidelity VR increase technical skills in the operating room, however, linkage to patient outcomes in minimally invasive gynecological surgery (MIGS) is lacking.[6] Furthermore, interpersonal skills, such teamwork and leadership, but also personal resourcefulness and advanced cognitive skills including error recognition and surgical planning play an important role in skills acquisition and intraoperative performance.[7,8] Moreover several studies have shown that surgical performance is associated with clinical outcomes and complication rates.[9,10]

Recently, there has been a focus on proficiency-based progression, dictating that the learner must meet specific performance benchmarks before progressing to the next stage in training.[11] To enable this, clearly defined metrics for those newly acquired surgical skills are needed. These can be formulated in objective assessment tools, defining and assessing the key steps of a specific procedure to support credentialing.

Global tools, such as the objective structured assessment of technical skills (OSATS), lack specificity which limits their applications in accreditation for a specific procedure, such as a hysterectomy.[12] To address this issue, an increasing number of recent cohort studies focusing on procedure-specific tools are being published. Furthermore, the emerging use of automated performance metrics (APMs) has not been reflected in previous systematic reviews assessing validity in MIGS.

The aim of this study was therefore to provide a comprehensive evaluation and updated review of the literature, reporting on all available assessment tools in MIGS. This evidence synthesis also appraised the reliability and validity of all reported tools including manual and automated in both simulated and live surgery.

Key message

There is a plethora of objective assessment tools in minimally invasive gynecological surgery. Further validation of already existing tools and integration of advanced technologies like kinematic data, should increase the usability in training curriculums.

2 | MATERIAL AND METHODS

2.1 | Data sources

The protocol for the study was developed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis guidelines (PRISMA).[13] The trial was registered with Prospective Register of Systematic Reviews (PROSPERO) (ID: CRD42022376552), a database of ongoing systematic reviews, to avoid duplication.[14] We searched for papers in the following databases: PubMed, MEDLINE, Embase and Web of Science from their inception until 17/11/2022. A broad search strategy was used (see Appendix S1). The search was performed capturing terms for minimally invasive including robotic-assisted laparoscopic gynecological procedures and assessment of performance. Finally, titles and abstracts were screened and full text eligible articles were reviewed.

2.2 | Eligibility criteria

Eligibility for inclusion was assessed by three independent assessors (FT, JN, IM). Articles reporting only on technical skills assessment in laparoscopic or robotic-assisted gynecological surgery were included in this review. This included randomized controlled trials, cohort and case–control studies. Furthermore, studies reporting on piloting these tools in any intraoperative, animal/wet laboratory and virtual reality (VR) simulated settings were also included. Exclusion criteria consisted of research reporting on open surgery, intrauterine and vaginal surgery, nontechnical skills tools, nonfull text available articles, abstracts or conference proceedings, pediatric studies, narrative reviews, commentary, editorials and non-English articles.

2.3 | Main outcomes measures

Data were independently extracted by three independent assessors (FT, JN, IM) using the Covidence online platform to aid analysis. Disagreements were resolved by discussion and if consensus could not be reached, a final decision was taken by the primary reviewer (FT). Extracted data included: study aim, study design, multi- and single center, number of participants, levels of participants, assessor blinding and validity evidence according to a contemporary framework of validity, which was adapted from Messick's validity framework.[15]

The quality of data and risk of bias for each included study were evaluated independently by the three assessors using the modified medical education research study quality instrument (MERSQI), a checklist appraising the methodological quality of medical education research studies.[16]

The possibility of performing a meta-analysis was considered and deemed unfeasible due to the heterogeneity in data reporting on the types of tools, data collection, study design and definition of expertise (novice vs. experts).

2.4 | Data collection and analysis

All five aspects of the contemporary framework were used to assess the validity of the assessment tools. This included content validity: testing whether the items of the objective assessment tool were relevant and represented the procedure. This is usually achieved by performing a consensus study among experts. Response process: observing how well scores reflect the observed performance. This could be achieved by providing a manual for the objective assessment tool or making sure the raters were blinded from the assessed participant. In the case of APMs, response process was always achieved because rater-bias was not present. Internal structure: testing whether scores are reliably reproducible. This was commonly achieved by providing inter-rater reliability (degree of agreement among multiple raters who independently assess the same surgeon) or intrarater reliability (assesses the consistency of a single rater's judgments over time). APMs cannot demonstrate rater reliability but can demonstrate internal consistency: the degree to which different items of an objective assessment tool are able to measure the same skill. Furthermore, we assessed the relationship to other variables testing whether scores correlated to clinical outcomes (predictive validity), scores from other assessment tools (concurrent validity) or level of surgical experience (construct validity). Finally, we assessed the impact using the assessment tools (consequences). This could be represented in a pass–fail score or the development of a summative or formative assessment tool. We used a scoring system rating each tool from each study, provided initially by Beckman et al. and later adjusted by Ghaderi et al., Haug et al. and Grüter et al., but modified for this systematic review.[17–20] Each aspect of the validity framework would count for a score from 0 to 3. The maximum score was 15: A score of 1–5 was associated with limited validity, a score of 6–10 with moderate validity and 11–15 with substantial validity. The definitions, examples and scoring system for manual and APMs (simulation) are summarized in Table 1.

3 | RESULTS

3.1 | General characteristics of the studies

A total of 19 746 titles and abstracts were screened for their eligibility and four additional studies were identified through other sources (citation searching $n=4$, gray literature $n=0$). A total of 174 articles were included for full text review and 102 studies were deemed ineligible, primarily because these studies were not reporting solely on MIGS or because no assessment tools were reported in those studies. Finally, 72 studies were included for further analysis. A breakdown of inclusion and exclusion is shown in the PRISMA diagram (Figure 1).

Study characteristics are summarized in Table 2. Studies were predominately conducted in the USA (48.6%) or Europe (25.0%); however, authors were represented from five different continents (all except South America and Antarctica). Studies were published between 2002 and 2023. Included studies consisted of manual assessment tools ($n=26$) and APMs ($n=11$). The later ones were tools from which the scoring system was directly derived from kinematic data and systems events data, usually in a VR setting. A total of 36 out of 72 (50%) studies were designed to address the utilization of previously validated tools in an educational intervention setting, followed by 31 (43.7%) studies aimed to either develop a new tool or assess the validity of existing ones.

With the exception of one paper, scoring five points,[21] all other papers had a score ranging from 10 to 16.5 on the 18-MERSQI checklist. The main limitations were lack of randomized controlled trials (study design), lack of multicenter studies (sampling) and the absence of correlating study outcomes with clinical outcomes. The risk of bias per tool can be found in Tables 3–7 and risk of bias per study can be found in Table S1.

A total of 26 manual technical skills assessments were included. These consisted of 13 global rating tools, and 13 procedure-specific tools. Furthermore 11 APMs were identified. The results are summarized under three categories: global, procedure-specific and automated metrics tools.

TABLE 1 Framework of validity used in this study: Manual and APMs simulation.

Validity aspects	Definition	Score	Data extracted	Examples
Content	The relationship between the content of a test and the construct it is intended to measure	0	No data regarding the content validity	
		1	Expert judgment including small group discussion with limited data regarding the tool content	Expert judgment
		2	Task analysis/hierarchical task analysis References to a previously validated tool	Task analysis/hierarchical task analysis, based on previously validated tools
		3	Well defined developing process, both theoretical basis for the chosen items and systematic review by experts	Delphi method, pilot study
	The relationship between the content of the simulation and the construct it is intended to measure	0	No data regarding the content validity	Idem
		1	Expert judgment including small group discussion with limited data regarding the tool content	
		2	Listing the assessment items for the APMs simulation training content with some references to a panel of experts (limited description of the developing process)	
		3	Reference to previous validated content/items of the APM. Well defined developing process, both theoretical basis for the chosen items and systematic review by experts	
Response process	Relationship among data items within the assessment and how these relate to overarching construct	0	No data regarding the response process	
		1	Limited data reported. Use of an assessment tool without discussing the impact of the differences in response	User manuals
		2	Some data regarding different response of assessors. Some data about systems that reduce the variation between respondents	Structured assessor training before the assessment process. Blinding of raters
		3	Multiple sources of data examining response error through critical examination of response process and respondents Rater training	Validation of initial scores (pilot study), evaluation of response error after structured assessor training
	Assessment of how well the documented record (the metrics) reflects the observed performance	3	Performance metrics were recorded by the simulator (i.e., eliminating rater bias). Metrics for test score inclusion selected based on discriminative ability	Inherently, there is no rater bias in APMs
Internal structure	Degree to which these relationships are consistent with the construct underlying the proposed test score interpretations	0	No data regarding the internal structure	
		1	Limited data regarding internal structure (references to a single inter-rater reliability measure)	Simple measures of inter- or intrarater reliability
		2	A few measures of reliability reported, insufficiently item analysis	Inter-/intrarater reliability coefficient combined with single measure of inter-item inter-test reliability
		3	Multiple measures of reliability including inter rater reliability and item analysis (interitem reliability intertest reliability, item response theory)	Generalizability theory analysis, item response theory
	Assessment of the reliability of the simulation	0	No data regarding the internal structure	
		2	Limited data regarding internal structure references to a single internal consistency reliability measure	Internal consistency reliability: assessed through the objective metrics for each participant's attempts by calculating the Cronbach's alpha.
		3	Multiple measures of reliability including internal consistency	Multiple measure of internal consistency

TABLE 1 (Continued)

Validity aspects	Definition	Score	Data extracted	Examples
Relationship to other variables	The comparison of scores with other known outcomes, performance assessment scores or relevant variables	0	No data regarding the other variables	
		1	Correlation of assessment scores with experience or another tool (concurrent validity or criterion validity)	Too validated by experience or another tool
		2	Correlation of assessment scores with experience and another tool (concurrent validity and criterion validity)	Tool validated by experience and another tool
		3	Correlation between assessment scores and clinical outcomes (predictive validity)	Tool validated by clinical outcomes
	Idem		Idem	Idem
Consequences	The impact, beneficial or harmful and intended or unintended, of assessment	0	No data regarding the consequences	
		1	Limited data merely a discussion about future use	Describing feasibility and potential future use (data on assessment time, post assessment survey)
		2	The application of performance assessments to training programs	Describing education impact (formative/summative feedback, learning curve of trainees)
		3	The impact of assessment usage on trainees or patients	Criterion referenced score (pass/fail scores), cutoff scores for licensing purposes, predictive models
	Idem		Idem	Idem

Note: This table has been adapted and includes the modified framework of Messick's validity with evidence scoring list, adopted from Beckman et al.,[17] Ghaderi et al.,[18] Haug et al.,[19] Grüter et al.,[20] further adapted for this review.

Abbreviation: APMS, automated performance metrics.

3.1.1 | Global rating tools

The OSATS, global operative assessment of laparoscopic skills (GOALS), global evaluative assessment of robotic skills (GEARS), global rating scale (GRS—not further specified) and modified versions of these tools were used most frequently ($n=32$) in the included studies. With the exception of three tools (robotic-OSATs, modified GEARS, operative performance rating system [OPRS]), all other tools were validated intraoperatively. Other settings included wet and dry models. The (modified) OSATS was the only manual tool assessed in a VR setting.[22,23] The only error rating tool identified was the generic error rating tool (GERT).[24] Kilani proposed the global rating index of technical skills (GRITS) to intraoperatively assess the correlation of surgical skill performance scores between expert assessment and self-assessment in various laparoscopic gynecological procedures, concluding that self-assessments have a higher evaluation than expert assessments.[25] Finally, the OPRS was used in a dry laboratory setting, assessing robotic-assisted laparoscopic radical hysterectomy and pelvic lymphadenectomy performance.[26]

3.1.2 | Procedure-specific

A total of 10 procedure-specific tools were assessed intraoperatively. The robotic sacrocolpopexy simulation model,[18] the "assessment tool for total laparoscopic hysterectomy"[27] and the OSATS for laparoscopic suturing and intracorporeal knot tying[28] were assessed in a laboratory-based setting. The objective structured assessment of laparoscopic salpingectomy (OSA-LS) was based on both the original OSATS and a modified rating scale for laparoscopic cholecystectomy developed by Grantcharov et al.[29] The myTIPreport is a smartphone application where both the trainee performing the procedure and a faculty member assessed the technical skills on a checklist immediately after the procedure.[30] The laparoscopic salpingo-oophorectomy-OSATS (LSO-OSATS) was based on the OSA-LS but consisted of fewer items (6 in the LSO-OSATS vs. 10 in the OSA-LS). Remarkably, six minimal invasive hysterectomy procedure-specific tools were included: Objective scale for assessment of technical skills of TLH (H-OSATS), the objective structured assessment of TLH (OSA-TLH), laparoscopic hysterectomy-OSATS (LH-OSATS) and the assessment tool for TLH, competency assessment tool for laparoscopic supracervical hysterectomy (CAT-LSH) and the robotic hysterectomy assessment score (RHAS).[22,27,31–34] All manual assessment tools and studies are summarized in Tables 3–6.

3.1.3 | Automated performance metrics (APMs)

A total of 11 APMs were identified in this systematic review. These include APMs in robotic-assisted laparoscopic VR simulations; da Vinci Surgical Simulation, RobotiX Mentor Simulation, and laparoscopic VR simulations: LapSim, LapMentor, MIST, SurgicalSim,

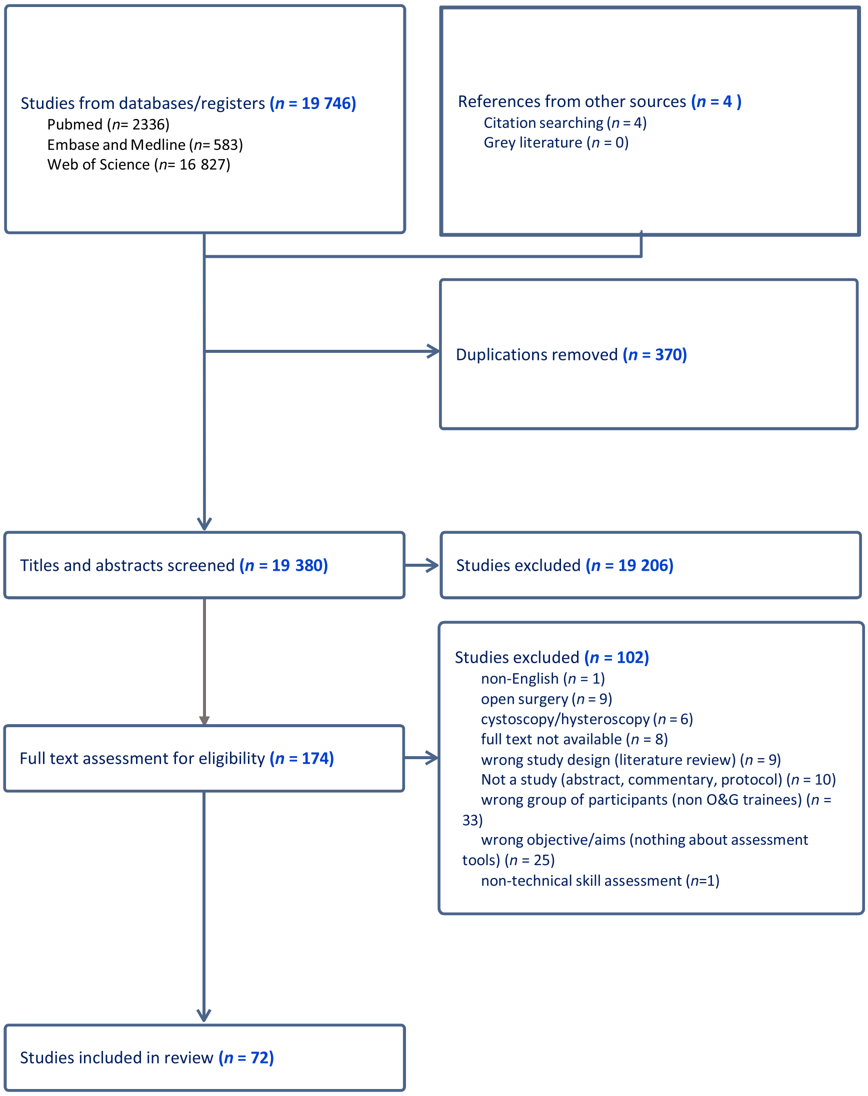

FIGURE 1 PRISMA flow diagram.

MISTELS, TRLCD05, FastTrack and LapVR simulator. No APMs were used intraoperatively.

3.1.4 | Validity of assessment tools

The following part of this review focuses on the sources of validity evidence of each included assessment tool, specified on the unitary framework for manual tools and APMs (Table 1). Given the heterogeneity of the interventions being investigated, each tool was categorized along the five dimensions of the contemporary framework (Tables 3–7; Table S2).

3.1.5 | Content

A total of 8 out of 13 (61.5%) global assessment tools had content validity.[30,34–41] The studies reporting on the generic skills assessment tool, GERT and OPRS, did not demonstrate content validity. In contrast to the global rating tools, content validity for the procedure-specific tools was provided 12/13 (92.3%) studies. A total of 10 (76.9%) tools underwent a (hierarchical) task analysis,[21,22,27,28,31–33,42–50] usually followed by a consensus study with experts.

Content validity was demonstrated in 6 (54.5%) different APMs.[51–56] One APM study reported on consensus methodology to reach content validity.[56]

3.1.6 | Response process

Eight out of 13 global assessment tools (61.5%) provided evidence of rater training, either by a training session or by providing a manual for tool usage.[24–26,28,30,37,38,40,41,57,58] This was applicable to seven out of 12 (58.3%) procedure-specific tools.[31–33,42–50,59,60] Addison et al. used crowd-sourced assessment of technical skills (CSATS) for GEARS and raters were routinely trained and evaluated for their rating reliability.[41]

All APMs inherently demonstrate response process, as they are automated, hence removing rater bias, and theoretically having 100% reliability.

TABLE 2 Characteristics of 72 included studies reporting on objective assessment tools in minimally invasive gynecological surgery.

Characteristics	*N* (%)
Country	
USA	35 (48.6)
Denmark	8 (11.1)
Canada	7 (9.7)
UK	5 (6.9)
France	5 (6.9)
Australia	3 (4.2)
Other	9 (12.5)
Study aim	
Development of tools and/or validation study	31 (43.7)
Utilization of tools for educational intervention study	36 (50.0)
Development or validation + use in education intervention study	5 (7.0)
Type of assessment tools	
Total	37
Manual assessment tools	26 (70.3)
Automated performance metrics	11 (29.7)
Type of minimal invasive gynecological surgery	
Laparoscopic procedures	60 (83.3)
Robotic-assisted procedures	12 (16.7)
Benign/gynecology oncology	
Benign	66 (91.7)
Gynecology oncology	6 (8.3)
Study design	
Randomized controlled trial	11 (15.3)
Cohort	61 (84.7)

3.1.7 | Internal structure

Internal structure was assessed in different ways. The most common reported form was inter-rater reliability with 10 out of 13 (76.9%) global rating tools and 10 out of 13 (76.9%) procedure-specific tools. All manual global tools report good to excellent inter-rater reliability,[23–25,30,33,35,37–40,57,61–70] with the exception of one modified OSATS in a dry laboratory setting for a simple laparoscopic ovarian cystectomy by Chahine et al.[71] Inter-rater reliability among procedure-specific tools[31–33,42–44,46,47,59,60] showed good to excellent correlation except for specific domains for the H-OSATS and the dissection assessment for robotic technique.[43,60] Intrarater reliability was reported for 4/13 (30.8%) global tools demonstrating excellent intrarater reliability.[28,30,38,40] The H-OSATS was the only procedure-specific tool reporting excellent intrarater reliability.[31,60]

Only one APM study calculated internal structure, reporting poor internal consistency (Cronbach's alpha 0.58) on the RobotiX Mentor Simulator.[55]

TABLE 3 Validity evidence per tool, manual global assessment tools (1/2).

Tools	GOALS[61,69,72]	mGOALS[30,34,35,61–64,67]	OSATS[23,58,65,73–76]	mOSATS[22,36,37,65,66,71]	R-OSATS[40]	GRS[38,39,70,77]
Setting	Intraoperative Dry laboratory	Intraoperative Wet laboratory Dry laboratory	Intraoperative Wet laboratory Dry laboratory Virtual Reality	Intraoperative Dry laboratory Virtual Reality	Dry laboratory	Intraoperative Dry laboratory
Procedure	Robotic supracervical hysterectomy Laparoscopic colpotomy Total laparoscopic hysterectomy Laparoscopic myomectomy	Laparoscopic colpotomy Laparoscopic vaginal cuff suturing Laparoscopic sacrocolpopexy Laparoscopic supracervical hysterectomy Laparoscopic myomectomy	Laparoscopic oophorectomy, dissection and ligature of uterine artery Diagnostic laparoscopy Laparoscopic sacrocolpopexy Laparoscopic bilateral tubal ligation Laparoscopic salpingectomy Total laparoscopic hysterectomy	Laparoscopic suturing and intracorporeal knot tying Basic dry laboratory tasks Laparoscopic ovarian cystectomy Total laparoscopic hysterectomy Laparoscopic bilateral tubal ligation Laparoscopic salpingo-oophorectomy	Basic dry laboratory tasks	Laparoscopic vaginal cuff closure Bilateral tubal ligation Basic dry laboratory tasks Total laparoscopic hysterectomy Laparoscopic Salpingectomy/salpingo-oophorectomy
Number of participants	14–28	12–40	10–102	16–102	105	10–99

(Continues)

TABLE 3 (Continued)

Tools	GOALS[61,69,72]	mGOALS[30,34,35,61–64,67]	OSATS[23,58,65,73–76]	mOSATS[22,36,37,65,66,71]	R-OSATS[40]	GRS[38,39,70,77]
Level of expertise	Novices, intermediates and experts					
Content	Referred to previous literature for content validity	Referred to previous literature for content validity	None	Referred to previous literature for content validity	Delphi consensus to develop tool	Delphi consensus developed tool Referred to previous literature for content validity
Response process	Response process present in all studies except for one study	Response process present in all studies except for one study	Response process present in all studies except for two studies	Response process present in all studies except for one study	Response process present	Response process present in one study
Internal structure	Inter rater reliability: Strong–Excellent Intrarater reliability: Not assessed Item analysis: Not assessed	Inter rater reliability: Good–Excellent Intrarater reliability: Good Item analysis: Not assessed	Inter-rater reliability: Good Intrarater reliability: Not assessed Inter item analysis: Excellent	Inter-rater reliability: Weak–excellent Intrarater reliability: Not assessed Item analysis: Not assessed	Inter-rater reliability: Acceptable Intrarater reliability: Excellent Inter item analysis: not assessed	Inter-rater reliability: Good Intrarater reliability: Very strong Inter item analysis: Excellent
Relations to other variables Construct validity (training level or case experience) Concurrent validity (other performance scores) Predictive validity (relation to clinical outcomes)	Construct validity	Construct validity Concurrent validity	Construct validity Concurrent validity	Construct validity	Construct validity	Construct validity
Consequences		Pass mark score defined at 27/35 82/85 32/40 However, not applied for benchmarking/credentialing in training curriculum				
Level of evidence	2b	2b	1b, 2a, 2b, 3	1b, 2a, 2b	2b	1b, 2a, 2b
Quality assessment (MERSQI score, maximum score is 18)	13.5–14	12.5–15.5	11–16.5	12–16	15.5	12.5–14

Note: Level of evidence according to the 2011 Oxford CEBM Levels of Evidence.[103]

Abbreviations: GOALS, global operative sssessment of laparoscopic skills; GRS, global rating skills; m, modified; MERSQI, medical education research study quality[16]; OSATS, objective structured assessment of technical skills; R-OSATS, robotic objective structured assessment of technical skills.

TABLE 4 Validity evidence per tool, manual global assessment tools (2/2).

Tool	GEARS[41]	mGEARS[67]	GRIT[25]	GSAT[68]	GERT[24]	TCPE[78]	OPRS[26]
Setting	Intraoperative	Wet laborarory	Intraoperative	Intraoperative Dry laboratory	Intraoperative	Dry laboratory Virtual reality	Dry laboratory
Procedure	Robotic Assisted Hysterectomy	Robotic vaginal cuff closure	Laparoscopic salpingectomy/ salpingo-oophorectomy Laparoscopic resection of endometriosis, adhesiolysis and ovarian drilling	Laparoscopic salpingectomy	Total laparoscopic hysterectomy	Basic dry laboratory tasks	Robotic-assisted hysterectomy
Number of participants	30	30	8	20	14	192	16
Content	Referred to previous literature for content validity	Experts developed tool based on GEARS	Referred to previous literature for content validity	None	None	None	Referred to previous literature for content validity
Response process	Response process present	None	Response process present	None	Response process present	None	Response process present
Internal structure	None	Inter-rater reliability: Excellent Intrarater reliability: Not assessed Item analysis: Not assessed	Inter rater reliability: Excellent Intrarater reliability: Not assessed Item analysis: Good to excellent	Inter-rater reliability: substantial Intrarater reliability: not assessed Item analysis: not assessed	Inter-rater reliability: Excellent Intrarater reliability: Excellent Item analysis: Not assessed	None	None
Relations to other variables Construct (training level or case experience) Concurrent validity (other performance scores) Predictive validity (relation to clinical outcomes)		Construct validity		Construct validity	Concurrent Validity	Construct validity	Construct validity
Consequences		Pass mark defined at 27/35 However, not applied for benchmarking/ credentialing in training curriculum					
Level of evidence	3	2b	3	2a	2b	2b	2b
Quality assessment (MERSQI score)	11	14.5	12	15	14.5	13.5	13.5

Note: Level of evidence according to the 2011 Oxford CEBM Levels of Evidence.[103]

Abbreviations: GEARS, global evaluative assessment of robotic skills; GERT, generic error rating tool; GRIT, global rating index of technical skills; GSAT, generic skills assessment tool; m, modified; MERSQI, medical education research study quality[16]; OPRS, operating performance rating system; TCPE, time to correct performed exercise.

TABLE 5 Validity evidence per tool, manual procedure-specific assessment tools.

Tool	OSA-LS[46–49]	H-OSATS[31,60]	LSO-OSATS[59]	RHAS[21]	DART[43]	CAT-LSH[34,44]
Setting	Intraoperative	Intraoperative Virtual reality	Intraoperative	Intraoperative	Intraoperative	Intraoperative
Procedure	Laparoscopic salpingectomy	Total laparoscopic hysterectomy	Laparoscopic salpingo-oophorectomy	Robotic assisted laparoscopic hysterectomy	Robotic assisted dissection performance	Laparoscopic Supracervical hysterectomy
Numbers of participants	3–32	14–30	24	57	36	21
Content	Delphi consensus to develop tool Referred to previous literature	Experts developed task-analysis Referred to previous literature for content validity	Experts developed tool	Delphi consensus to develop tool	Delphi consensus to develop tool	Experts developed tool
Response process	All studies demonstrated response process	All studies demonstrated response process	None	None	None	None
Internal structure	Inter-rater reliability: Strong–Very Strong	Inter-rater reliability: Excellent Fair–excellent	Inter-rater reliability: Excellent	Inter-rater reliability: Poor–good	Inter-rater reliability: Poor–good	Inter-rater reliability: Excellent
	Intrarater reliability: not assessed	Intrarater reliability: Excellent	Intrarater reliability: none	Intrarater reliability: not assessed	Intrarater reliability: not assessed	Intrarater reliability: not assessed
	Item analysis: Moderate-perfect	Item analysis: not assessed	Item analysis: not assessed	Item analysis: not assessed	Item analysis: not assessed	Item analysis: not assessed
Relations to other variables Construct validity (training level or case experience) Concurrent validity (other performance scores) Predictive validity (relation to clinical outcomes)	Construct validity	Concurrent validity Construct validity		Construct validity	Construct validity	Construct validity
Consequences	None	Pass mark defined (90/150) However, not applied for benchmarking/ credentialing in training curriculum	None	None	None	None
Level of evidence	1b, 2a, 2b,3	2b	2a	2b	2b	2b
Quality assessment (MERSQI score)	10–16.5	13–13.5	15	14	14.5	12.5–13.5

Note: Level of evidence according to the 2011 Oxford CEBM Levels of Evidence.[103]

Abbreviations: CAT-LSH, competency assessment tool for laparoscopic supracervical hysterectomy; DART, dissection of assessment of robotic technique; H-OSATS, objective scale for assessment of technical skills of total laparoscopic hysterectomy (TLH); LSO-OSATS, OSATS of laparoscopic salpingo-oophorectomy; MERSQI, medical education research study quality[16]; OSA-LS, objective structured assessment of laparoscopic salpingectomy; RHAS, robotic hysterectomy assessment score.

TABLE 6 Validity evidence per tool, manual procedure-specific assessment tools.

Tool	OSA-TLH[32]	Surgical competency assessment tool for sentinel lymph node dissection[45]	Robotic sacrocolpopexy simulation model[21]	Assessment tool for TLH[27]	OSATS for Laparoscopic Suturing and Intracorporeal Knot Tying[28]	LH-OSATS[33]	myTIPreport[30]
Setting	Intraoperative	Intraoperative	Dry laboratory	None	Dry laboratory	Intraoperative	Intraoperative
Procedure	Total laparoscopic hysterectomy	Laparoscopic sentinel lymph node dissection	Robotic assisted laparoscopic sacrocolpopexy	Total laparoscopic hysterectomy	Laparoscopic suturing and intracorporeal knot tying	Total laparoscopic hysterectomy	Total laparoscopic Hysterectomy
Number of participants	16	35	6	51	14	20	28
Content	Delphi consensus to develop tool	Delphi consensus to develop tool	Experts developed tool	Delphi consensus to developed tool	Experts developed tool	Experts developed tool	Experts composed this checklist tool
Response Process	None	None	None	None	None	None	Response process present
Internal structure	Inter-rater reliability: Excellent Intrarater reliability: not assessed Item analysis: Excellent	Inter-rater reliability: not assessed Intrarater reliability: not assessed Item-analysis: good	None	None	Inter-rater reliability: Strong Intrarater reliability: Strong Item-analysis: not assessed	Intrarater reliability: fair Intrarater reliability: not assessed Item-analysis: not assessed	Inter-rater reliability: Strong Intrarater reliability: weak Item analysis: not assessed
Relations to other variables Construct validity (training level or case experience) Concurrent validity (other performance scores) Predictive validity (relation to clinical outcomes)	Construct validity	Construct validity	None	None	None	Construct validity	Construct validity
Consequences	Pass mark defined (29.3/55) However, not applied for benchmarking/credentialing in training curriculum	None	None	None	None	None	
Level of evidence	2b	2b	4	4	2b	1b	2b
Quality assessment (MERSQI score)	14.5	14.5	5	11.5	12	15	14

Note: Level of evidence according to the 2011 Oxford CEBM Levels of Evidence.[103]

Abbreviations: MERSQI, medical education research study quality[16]; OSA-TLH, OSATS of TLH, LH-OSATS is OSATS of TLH.

TABLE 7 Validity evidence per tool, automated performance metrics.

Tool	DvSS[51,67,72,79]	LapSim[23,34,48,52,80–84,102]	Lapmentor Express[85,86,104]	VBLAST-PT[53]	MIST[54,101]
Setting	Virtual reality	Virtual reality	Virtual reality	Virtual reality	Virtual reality
Procedure	Basic robotic modules	Basic laparoscopic VR modules Laparoscopic Salpingectomy	Basic laparoscopic VR modules Laparoscopic Salpingectomy/ Salpingo-oophorectomy	Basic Laparoscopic VR module	Basic Laparoscopic VR modules
Numbers of participants	11–20	22–63	24–31	27	26–44
Content	Expert assessed content of metrics	Expert assessed content of metrics	None	Expert assessed content of metrics	Expert assessed content of metrics
Response process	Automated performance metrics (no rater bias)	Automated performance metrics (no rater bias)	Automated performance metrics (no rater bias)	Automated performance metrics (no rater bias)	Automated performance metrics (no rater bias)
Internal structure	None	None	None	None	None
Relations to other variables Construct validity (training level or case experience) Concurrent validity (other performance scores) Predictive validity (relation to clinical outcomes)	All studies showed construct validity	All studies showed construct validity	All studies showed construct validity	Construct validity	None
Consequences	Thresholds for robotic modules experts scores were provided However, not applied for benchmarking/ credentialing in training curriculum				
Level of Evidence	2b	2a, 2b, 3	2b	2b	2b
Quality assessment (MERSQI score)	12–12.5	11–13.5	11–13.5	11.5	11.5

Note: Level of evidence according to the 2011 Oxford CEBM Levels of Evidence.[103]

Abbreviations: DvSS, Da Vinci surgical system; LapSim, laparoscopic simulator; LapVR simulator, laparoscopic virtual reality simulator; MERSQI, medical education research study quality[16]; MIST, minimally invasive surgical trainer, McGill inanimate system for training and evaluation of laparoscopic skills; VBLAST-PT, virtual basic laparoscopic skill trainer.

3.1.8 | Relationship to other variables

A total of 10 out of 13 (79.6%) global tools reported relationships to other variables by either comparing novices to experts (construct validity) or showing significant correlation between scoring outcomes and other performance assessment tools, considered the gold standard (concurrent validity).[22–24,27,28,30,35,36,38–40,57,58,61–77] Nine out of 13 (69.2%) procedure-specific tool studies[31–33,42–44,46–50,60] showed construct or concurrent validity. Nine out of 11 (81.8%) APM showed construct or criterion

SurgicalSim[87]	Dv-Trainer mimic[88]	TRLCD05[88]	RobotiX Mentor simulator[55]	Fastrack[89]	LapVR simulator[56,90]
Virtual reality	Virtual reality	Virtual reality	Virtual reality	Wet laboratory	Virtual reality
Basic Laparoscopic VR modules	Robotic VR modules	Laparoscopic VR modules	Robotic-assisted vaginal cuff closure	Laparoscopic pelvic lymphadenectomy	Laparoscopic salpingectomy and salpingotomy on a right sided isthmic tubal pregnancy
22	16	16	22	20	34
None	None	None	Expert assessed content of metrics	None	Content was validated by experts
Automated performance metrics (no rater bias)	Completion time was measured	Completion time was measured	Automated performance metrics (no rater bias)	Movement analysis was performed by tracking the position of the Fastrack transducers (no rater bias)	Automated performance metrics (no rater bias)
None	None	None	Inter-consistency reliability: poor	None	None
Construct validity	Construct validity	Construct validity	Construct validity	Construct validity	Both studies showed construct validity Concurrent validity
	Residents in both groups (laparoscopic and robotic group) were more comfortable performing surgery in their method of training	Residents in both groups (laparoscopic and robotic group) were more comfortable performing surgery in their method of training	Pass mark defined at 75/110[48] However, not applied for benchmarking/ credentialing in training curriculum		
2b	2a	2a	2b	2b	2b,[49] 3
12.5	14	14	13.5	14.5	13.5

validity.[51–53,55,56,67,72,78–90] None of the included studies reported on the association between intraoperative performance of practicing surgeons to clinical/postoperative outcomes of patients (predictive validity).

3.1.9 | Consequences

Four out of 26 (15.4%) manual (global and procedure-specific tools) provided benchmark scores.[30,32,35,57,60,64,67] One out of 11 (9.1%)

APMs provided a benchmark score on the RobotiX Mentor providing a pass/fail score of 75/110.[55]

3.1.10 | Validity evidence

Table 8 summarizes the evidence of validity of all tools based on the scoring tool from Table 1. Only one manual tool showed substantial evidence of validity (score 11–15): the total laparoscopic hysterectomy procedure specific tools: OSA-TLH. The RobotiX Mentor Simulator was the only APM showing substantive evidence. A total of 17 tools showed moderate evidence (score 6–10) of which 10 were global tools, six were procedure specific and one APMs. Finally, 18 tools showed limited evidence of validity, of which nine were APMs, three manual global tools and six procedure specific tools.

4 | DISCUSSION

This systematic review of 72 articles identified 37 surgical performance assessment tools that have been studied in a laparoscopic and robotic-assisted gynecological surgery setting. This review provided a comprehensive evaluation of the validity and reliability of assessment tools, using a contemporary validity framework (Table 1). These included 26 manual tools and 11 APMs. Interestingly, none of the studies were able to show predictive validity (correlating the tool score with clinical outcomes).

Tough achieving predictive validity often necessitates a more demanding endeavor, and there is still a significant opportunity to develop study settings correlating tool scores with clinical outcomes.[91,92] The General Medical Council (GMC) in the UK has even stated that in the absence of the gold standard, exploring the strength of the relationship between similar established assessment tools, from different surgical specialities, might offer itself as an alternative.[93] Furthermore, more granular analysis of surgical skills, such as the objective clinical human reliability analysis (OCHRA) could enhance the likelihood of achieving predictive validity, associating technical kills with clinical outcomes, regardless of level of expertise.[94]

When looking at current training programs, such as the Royal College of Obstetrics and Gynecology (RCOG) in the UK, it interesting to see that the most frequent used objective assessment tool is the OSATS.[95] Global assessment tools are easily available for different procedures. However, this systematic review showed that the only manual tool showing substantive evidence was a procedure specific tool. It should also be noted that the exchange of constructive feedback within the trainer-trainee dialogue often plays a greater role in shaping learning outcomes.

Culligan et al. proposed a robotic surgery simulation training curriculum and established predictive validity by demonstrating a correlation between program completion and improved live surgery outcomes.[72] These included reduced estimated blood loss, shorter operating times, and enhanced intraoperative GOALS scores.

TABLE 8 Objective assessment tools arranged by strength of validity based on the validity evidence scoring list from Table 1 (substantial, moderate and limited evidence).

Level of evidence according to score	Tool name	Total
Substantial evidence (score 11–15)	OSA-TLH	11
	RobotiX Mentor simulator	11
Moderate evidence (score 6–10)	mGOALS	10
	H-OSATS	10
	OSATS	8
	OSA-LS	8
	R-OSATS	7
	GRS	7
	RHAS	7
	DART	7
	Surgical competency assessment tool for sentinel lymph node dissection	7
	DvSS	7
	LH-OSATS	6
	LapVRsimulator	6
	mGEARS	6
	GRIT	6
	GERT	6
	mOSATS	6
	GOALS	6
Limited evidence (score 0–5)	LapSim	5
	Lapmentor Express	5
	Fastrack	5
	MIST	5
	TRLCD05	5
	OPRS	5
	CAT-LSH	5
	VBLAST-PT	4
	SurigcalSim	4
	Dv-Trainer mimic	4
	LSO-OSATS	4
	Assessment tool for TLH	3
	OSATS for laparoscopic suturing and intracorporeal knot tying	3
	GEARS	3
	myTIPreport	3
	Robotic sacrocolpopexy simulation model	2
	GSAT	2
	TCPE	1

However, the study's generalizability was limited by its restriction to board-certified obstetrics and gynecology surgeons. Despite other studies reporting pass/fail scores for a modified GOALS, modified

AOGS Acta Obstetricia et Gynecologica Scandinavica

GEARS, H-OSATS, OSA-TLH (both TLH procedure specific tool), the RobotiX Mentor simulator and DvSS, none of them showed any evidence of successful implementation of curriculums for credentialing. Future research should not only focus on investigating other aspects of validity, but also on benchmarking already available objective assessment tools to make them useful additions to surgical national curriculums. This will ultimately enhance the standardization and effectiveness of resident and fellow training in MIGS.

This systematic review had some limitations. First, it was limited by only including studies in English. Another limitation was that the majority of the studies were small, conducted once or in a nonrandomized single center setting, risking potential biases and compromising reproducibility of results. Often, different thresholds and definitions were used, producing heterogeneity and the subsequent inability to perform a meaningful meta-analysis, highlighting that tools should be evaluated more thoroughly in large, well-run studies. Furthermore, assessment tools for intrauterine and vaginal surgery were not included.

However, a significant increase in numbers of assessment tools ($n=37$) in MIGS were identified, making it, to our knowledge, the most comprehensive and detailed systematic review on the subject of minimal invasive gynecology surgery. It is important to inform the gynecological surgical community of all available tools that can be applied not only in the research settings but to support learning and teaching.

Ferriss et al., published a systematic review of intraoperative assessment tools in MIGS, focusing mainly on manual performance metrics.[96] They concluded that procedure-specific tools are more thoroughly evaluated, however described their limited use due to poor quality studies and borderline reliability. Another scoping review by Hennings et al. explained that most surgical assessment scales were validated in simulation settings, compromising transferability to the operating room.[12] However, comprehensive evaluations of the tools' validity were not reported, mainly lacking the consequence component.

This review also provided a comprehensive review of APMs available in MIGS. One of the advantages of APMs compared to manual tools is the automated collection of the performances, preventing rater bias. Furthermore, it is less time consuming and manual tools require a degree of subjective scoring. Furthermore, research has been suggested that skills in laparoscopic surgery can be increased by proficiency-based procedural VR simulator training. However, this review showed that the majority of APMs (81.8%) has limited validity evidence. This low level of validity hinders transferability outside of the research environment. Moreover, these metrics alone cannot be considered substitutes of experts' input towards surgical competencies. Until true objective assessment tools are in place to provide expert opinion on trainees' performances within the clinical context, APMs are useful adjunct to support objective assessments of surgical skills. This systematic review did not identify APMs using kinematic data from live surgery.

However, recent studies in different specialities have been able to correlate kinematic data derived from recording devices with technical performance to create a scoring index in dry lab surgeries and live operations. Lyman et al., were able to correlate the operating robotic index model to level of experience in a dry laboratory robotic-assisted laparoscopic hepaticojejunostomy reconstruction, showing construct validity.[97] Another example of appliance is the emerging interdisciplinary field of surgical data science (SDS) aiming to improve quality of interventional healthcare by capturing, organizing, analyzing, and modeling data. Mascagni et al., were able to assess the critical view of safety criteria in laparoscopic cholecystectomy through annotating anatomical segments and training a deep neural network to predict critical view of safety occurrence.[98] Utilization of these tools, including kinematic data from advanced computing devices and surgical systems and artificial neural networks will become essential to better understand factors in surgical performance and ultimately standardize safe operation. One key challenge for developing these approaches further is the current absence of large-scale datasets that fully represent the domains of variation; for example, experience level, subtask, instrument type, in order to allow robust training of AI models with only limited clinical datasets currently available.[99,100] Future validation of APMs will support utilization while they are likely to expand in the future with artificial intelligence and machine learning.

5 | CONCLUSION

This comprehensive review offers an up-to-date overview of existing assessment tools for MIGS. With 37 tools identified, including both established manual techniques and APMs, it provides a valuable resource for researchers, educators, and practitioners alike. While global assessment tools remain readily available, procedure-specific tools hold great educational potential.

Importantly, the review highlights the gap in evidence regarding predictive validity—linking assessment scores to patient outcomes. Additionally, it underscores the limitations of current APMs, mainly due to insufficient content validity assessments. Nonetheless, APMs show promise in their objective data collection and potential for reducing rater bias. Future research should focus on addressing these limitations while continuing to explore the integration of advanced technologies like kinematic data analysis and artificial intelligence.

AUTHOR CONTRIBUTIONS

Freweini Martha Tesfai: Concept, design, data collection, analysis and interpretation, manuscript preparation. Jasleen Nagi and Iona Morrison: Data collection, data analysis and interpretation, manuscript preparation. Matt Boal: Concept and design, data analysis and interpretation, manuscript preparation. Adeola Olaitan, Dhivya Chandrasekaran, Danail Stoyanov, Anne Lanceley and Nader Francis: Data analysis and interpretation, manuscript preparation.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest.

ORCID

Freweini Martha Tesfai https://orcid.org/0009-0007-4303-666X
Anne Lanceley https://orcid.org/0000-0001-9143-9710

REFERENCES

1. Chao L, Lin E, Kho K. Enhanced recovery after surgery in minimally invasive gynecologic surgery. *Obstet Gynecol Clin N Am*. 2022;49:381-395.
2. Haidegger T, Speidel S, Stoyanov D, Satava RM. Robot-assisted minimally invasive surgery—surgical robotics in the data age. *Proc IEEE*. 2022;110(7):835-846.
3. Institute ECR. *Top 10 Health Technology Hazards Executive Brief 2020*. 2020. Accessed July 15, 2016. https://www.ecri.org/landing-2020-top-ten-health-technology-hazards
4. Whitson MJ, Williams RL, Shah BJ. Ensuring quality in endoscopic training: tools for the educator and trainee. *Tech Innov in Gastrointest Endosc*. 2022;24:354-363.
5. Vergis A, Steigerwald S. Skill acquisition, assessment, and simulation in minimal access surgery: an evolution of technical training in surgery. *Cureus*. 2018;10:e2969.
6. Orejuela FJ, Aschkenazi SO, Howard DL, et al. Gynecologic surgical skill acquisition through simulation with outcomes at the time of surgery: a systematic review and meta-analysis. *Am J Obstet Gynecol*. 2022;227:29.e1-29.e24.
7. Madani A, Vassiliou MC, Watanabe Y, et al. What are the principles that guide behaviors in the operating room?: creating a framework to define and measure performance. *Ann Surg*. 2017;265:255-267.
8. Suliburk JW, Buck QM, Pirko CJ, et al. Analysis of human performance deficiencies associated with surgical adverse events. *JAMA Netw Open*. 2019;2:e198067.
9. Cooperberg MR, Odiho AY, Carroll PR. Outcomes for radical prostatectomy: is it the singer, the song, or both? *J Clin Oncol*. 2012;30:476-478.
10. Vickers ASC, Bianco F, Mulhall J, et al. Cancer control and functional outcomes after radical prostatectomy as markers of surgical quality: analysis of heterogeneity between surgeons at a single cancer center. *Eur Urol*. 2011;59:317-322.
11. Gallagher AG, de Groote R, Paciotti M, Mottrie A. Proficiency-based progression training: a scientific approach to learning surgical skills. *Eur Urol*. 2022;81:394-395.
12. Hennings LI, Sorensen JL, Hybscmann J, Strandbygaard J. Tools for measuring technical skills during gynaecologic surgery: a scoping review. *BMC Med Educ*. 2021;21:402.
13. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ*. 2009;339:b2700.
14. *PROSPERO International Prospective Register of Systematic Reviews*. Accessed November 17, 2022. https://www.crd.york.ac.uk/prospero/
15. Watanabe Y, Bilgic E, Lebedeva E, et al. A systematic review of performance assessment tools for laparoscopic cholecystectomy. *Surg Endosc*. 2016;30:832-844.
16. Al Asmri M, Haque MS, Parle J. A Modified Medical Education Research Study Quality Instrument (MMERSQI) developed by Delphi consensus. *BMC Med Educ*. 2023;23:63.
17. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20:1159-1164.
18. Ghaderi I, Manji F, Park YS, et al. Technical skills assessment toolbox: a review using the unitary framework of validity. *Ann Surg*. 2015;261:251-262.
19. Haug TR, Orntoft MBW, Miskovic D, Iversen LH, Johnsen SP, Madsen AH. How can surgical skills in laparoscopic colon surgery be objectively assessed?—a scoping review. *Surg Endosc*. 2022;36:1761-1774.
20. Grüter AAJ, van Lieshout AS, van Oostendorp SE, et al. Video-based tools for surgical quality assessment of technical skills in laparoscopic procedures: a systematic review. *Surg Endosc*. 2023;37:4279-4297.
21. Myers EM, Anderson-Montoya BL, Fasano HT, Vilasagar S, Tarr ME. Robotic sacrocolpopexy simulation model and associated hierarchical task analysis. *Obstet Gynecol*. 2019;133:905-909.
22. Crochet P, Aggarwal R, Knight S, Berdah S, Boubli L, Agostini A. Development of an evidence-based training program for laparoscopic hysterectomy on a virtual reality simulator. *Surg Endosc*. 2017;31:2474-2482.
23. Enciso S, Diaz-Guemes I, Perez-Medina T, et al. Validation of a structured intensive laparoscopic course for basic and advanced gynecologic skills training. *Int J Gynaecol Obstet*. 2016;133:241-244.
24. Husslein H, Shirreff L, Shore EM, Lefebvre GG, Grantcharov TP. The generic error rating tool: a novel approach to assessment of performance and surgical education in gynecologic laparoscopy. *J Surg Educ*. 2015;72:1259-1265.
25. Kilani R. Comparing self-assessment of laparoscopic technical skills with expert opinion for gynecological surgeons in an operative setting. *Gynecol Surg*. 2018;15:16.
26. Umemura K, Kawai Y, Machida H, et al. An innovative tissue model for robot-assisted radical hysterectomy and pelvic lymphadenectomy. *Eur J Gynaecol Oncol*. 2021;42:482-487.
27. Tremblay C, Grantcharov T, Urquia ML, Satkunaratnam A. Assessment tool for total laparoscopic hysterectomy: a Delphi consensus survey among international experts. *J Obstet Gynaecol Can*. 2014;36:1014-1023.
28. Chang OH, King LP, Modest AM, Hur HC. Developing an objective structured assessment of technical skills for laparoscopic suturing and intracorporeal knot tying. *J Surg Educ*. 2016;73:258-263.
29. Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surg*. 2004;91:146-150.
30. DeStephano CC, Chen AH, Heckman MG, et al. Use of the limbs and things hysterectomy model to describe the process for establishing validity. *J Minim Invasive Gynecol*. 2018;25:1051-1059.
31. Knight S, Aggarwal R, Agostini A, Loundou A, Berdah S, Crochet P. Development of an objective assessment tool for total laparoscopic hysterectomy: a Delphi method among experts and evaluation on a virtual reality simulator. *PLoS One*. 2018;13:e0190580.
32. Savran MM, Hoffmann E, Konge L, Ottosen C, Larsen CR. Objective assessment of total laparoscopic hysterectomy: development and validation of a feasible rating scale for formative and summative feedback. *Eur J Obstet Gynecol Reprod Biol*. 2019;237:74-78.
33. Jokinen E, Mikkola TS, Härkki P. Simulator training and residents' frst laparoscopic hysterectomy: a randomized controlled trial. *Surg Endosc*. 2020;34:4874-4882.
34. Goderstad JM, Fosse E, Sandvik L, LienG M. Development and validation of a curriculum for laparoscopic supracervical hysterectomy. *Facts Views Vis Obgyn*. 2020;12:83-90.
35. Newcomb LK, Bradley MS, Truong T, et al. Correlation of virtual reality simulation and dry lab robotic technical skills. *J Minim Invasive Gynecol*. 2018;25:689-696.
36. Mannella P, Malacarne E, Giannini A, et al. Simulation as tool for evaluating and improving technical skills in laparoscopic gynecological surgery. *BMC Surg*. 2019;19:146.
37. Jokinen E, Mikkola TS, Harkki P. Effect of structural training on surgical outcomes of residents' first operative laparoscopy: a randomized controlled trial. *Surg Endosc*. 2019;33:3688-3695.

38. Fuchs Weizman N, Manoucheri E, Vitonis AF, Hicks GJ, Einarsson JI, Cohen SL. Design and validation of a novel assessment tool for laparoscopic suturing of the vaginal cuff during hysterectomy. *J Surg Educ*. 2015;72:212-219.
39. Chen CC, Green IC, Colbert-Getz JM, et al. Warm-up on a simulator improves residents' performance in laparoscopic surgery: a randomized trial. *Int Urogynecol J*. 2013;24:1615-1622.
40. Siddiqui NY, Galloway ML, Geller EJ, et al. Validity and reliability of the robotic objective structured assessment of technical skills. *Obstet Gynecol*. 2014;123:1193-1199.
41. Addison P, Bitner D, Chung P, et al. Blinded intraoperative skill evaluations avoid gender-based bias. *Surg Endosc*. 2022;36:8458-8462.
42. Frederick PJ, Szender JB, Hussein AA, et al. Surgical competency for robot-assisted hysterectomy: development and validation of a robotic hysterectomy assessment score (RHAS). *J Minim Invasive Gynecol*. 2017;24:55-61.
43. Vanstrum EB, Ma R, Maya-Silva J, et al. Development and validation of an objective scoring tool to evaluate surgical dissection: Dissection Assessment for Robotic Technique (DART). *Urol Pract*. 2021;8:596-604.
44. Goderstad JM, Sandvik L, Fosse E, Lieng M. Assessment of surgical competence: development and validation of rating scales used for laparoscopic supracervical hysterectomy. *J Surg Educ*. 2016;73:600-608.
45. Moloney K, Janda M, Frumovitz M, et al. Development of a surgical competency assessment tool for sentinel lymph node dissection by minimally invasive surgery for endometrial cancer. *Int J Gynecol Cancer*. 2021;31:647-655.
46. Larsen CR, Grantcharov T, Schouenborg L, Ottosen C, Soerensen JL, Ottesen B. Objective assessment of surgical competence in gynaecological laparoscopy: development and validation of a procedure-specific rating scale. *BJOG*. 2008;115:908-916.
47. Larsen CR, Soerensen JL, Grantcharov TP, et al. Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. *BMJ*. 2009;338:b1802.
48. Strandbygaard J, Bjerrum F, Maagaard M, Rifbjerg Larsen C, Ottesen B, Sorensen JL. A structured four-step curriculum in basic laparoscopy: development and validation. *Acta Obstet Gynecol Scand*. 2014;93:359-366.
49. Oestergaard J, Larsen CR, Maagaard M, Grantcharov T, Ottesen B, Sorensen JL. Can both residents and chief physicians assess surgical skills? *Surg Endosc*. 2012;26:2054-2060.
50. Tannenbaum E, Walker M, Sullivan H, Huszti E, Farrugia M, Sobel M. Effects of a resident's reputation on laparoscopic skills assessment. *Obstet Gynecol*. 2021;138:16-20.
51. Mayooran Z, Rombauts L, Brown TI, Tsaltas J, Fraser K, Healy DL. Reliability and validity of an objective assessment instrument of laparoscopic skill. *Fertil Steril*. 2004;82:976-978.
52. Elessawy M, Mabrouk M, Heilmann T, et al. Evaluation of laparoscopy virtual reality training on the improvement of trainees' surgical skills. *Medicina (Kaunas)*. 2021;57:130.
53. Awtrey C, Chellali A, Schwaitzberg S, De S, Jones D, Cao C. Validation of the VBLaST: a virtual peg transfer task in gynecologic surgeons. *J Minim Invasive Gynecol*. 2015;22:1271-1277.
54. Moore AK, Grow DR, Bush RW, Seymour NE. Novices outperform experienced laparoscopists on virtual reality laparoscopy simulator. *JSLS*. 2008;12:358-362.
55. Hovgaard LH, Andersen SAW, Konge L, Dalsgaard T, Larsen CR. Validity evidence for procedural competency in virtual reality robotic simulation, establishing a credible pass/fail standard for the vaginal cuff closure procedure. *Surg Endosc*. 2018;32:4200-4208.
56. Varras M, Loukas C, Nikiteas N, Varra VK, Varra FN, Georgiou E. Comparison of laparoscopic surgical skills acquired on a virtual reality simulator and a box trainer: an analysis for obstetrics-gynecology residents. *Clin Exp Obstet Gynecol*. 2020;47:755-763.
57. Leon MG, Dinh TA, Heckman MG, Weaver SE, Chase LA, DeStephano CC. Correcting the fundamentals of laparoscopic surgery "illusion of validity" in laparoscopic vaginal cuff suturing. *J Minim Invasive Gynecol*. 2021;28:1927-1934.
58. Supramaniam PR, Mittal M, Davies R, Lim LN, Arambage K. Didactic lectures versus simulation training: a randomised pilot evaluation of its impact on surgical skill. *Gynecol Surg*. 2018;15:21.
59. Norris S, Papillon-Smith J, Gagnon LH, Jacobson M, Sobel M, Shore EM. Effect of a surgical teaching video on resident performance of a laparoscopic salpingo-oophorectomy: a randomized controlled trial. *J Minim Invasive Gynecol*. 2020;27:1545-1551.
60. Crochet P, Netter A, Schmitt A, et al. Performance assessment for total laparoscopic hysterectomy in the operating room: validity evidence of a procedure-specific rating scale. *J Minim Invasive Gynecol*. 2021;28:1743-1750.
61. Frazzini Padilla PM, Farag S, Smith KA, Zimberg SE, Davila GW, Sprague ML. Development and validation of a simulation model for laparoscopic colpotomy. *Obstet Gynecol*. 2018;132(Suppl 1):19S-26S.
62. Tunitsky-Bitton E, Propst K, Muffly T. Development and validation of a laparoscopic hysterectomy cuff closure simulation model for surgical training. *Am J Obstet Gynecol*. 2016;2016(214):392.e1-392.e6.
63. Tunitsky-Bitton E, King CR, Ridgeway B, et al. Development and validation of a laparoscopic sacrocolpopexy simulation model for surgical training. *J Minim Invasive Gynecol*. 2014;21:612-618.
64. King CR, Donnellan N, Guido R, Ecker A, Althouse AD, Mansuria S. Development and validation of a laparoscopic simulation model for suturing the vaginal cuff. *Obstet Gynecol*. 2015;126(Suppl 4):27S-35S.
65. Gala R, Orejuela F, Gerten K, et al. Effect of validated skills simulation on operating room performance in obstetrics and gynecology residents: a randomized controlled trial. *Obstet Gynecol*. 2013;121:578-584.
66. Balafoutas D, Joukhadar R, Kiesel M, et al. The role of deconstructive teaching in the training of laparoscopy. *JSLS*. 2019;23:e2019.00020.
67. Gheza F, Pinkard L, Grand A, Aguiluz-Cornejo G, Mangano A, Ladanyi A. Development of an affordable, immersive model for robotic vaginal cuff closure: a randomized trial. *J Robot Surg*. 2023;17:109-116.
68. Coleman RL, Muller CY. Effects of a laboratory-based skills curriculum on laparoscopic proficiency: a randomized trial. *Am J Obstet Gynecol*. 2002;186:836-842.
69. Schneyer RJ, Molina AL, Green IC, et al. Development and validation of a simulation model for laparoscopic myomectomy. *Am J Obstet Gynecol*. 2022;227:304.e1-304.e9.
70. Banks EH, Chudnoff S, Karmin I, Wang C, Pardanani S. Does a surgical simulator improve resident operative performance of laparoscopic tubal ligation? *Am J Obstet Gynecol*. 2007;197(541):e1-e5.
71. Chahine EB, Han CH, Auguste T. Construct validity of a simple laparoscopic ovarian cystectomy model using a validated objective structured assessment of technical skills. *J Minim Invasive Gynecol*. 2017;24:850-854.
72. Culligan P, Gurshumov E, Lewis C, Priestley J, Komar J, Salamon C. Predictive validity of a training protocol using a robotic surgery simulator. *Female Pelvic Med Reconstr Surg*. 2014;20:48-51.
73. Karaoglan T, Aydin S, Bilginer U. Development of a low-fidelity laparoscopic sacrocolpopexy simulation model and evaluation of curriculum. *Female Pelvic Med Reconstr Surg*. 2021;27:474-480.
74. Patel NR, Makai GE, Sloan NL, Della Badia CR. Traditional versus simulation resident surgical laparoscopic salpingectomy training: a randomized controlled trial. *J Minim Invasive Gynecol*. 2016;23:372-377.

75. Arora C, Menzies A, Han ES, et al. Comparing surgical experience and skill using a high-fidelity, total laparoscopic hysterectomy model. *Obstet Gynecol*. 2020;136:97-108.
76. Botchorishvili R, Rabischong B, Larrain D, et al. Educational value of an intensive and structured interval practice laparoscopic training course for residents in obstetrics and gynecology: a four-year prospective, multi-institutional recruitment study. *J Surg Educ*. 2012;69:173-179.
77. Tarr ME, Rivard C, Petzel AE, et al. Robotic objective structured assessment of technical skills: a randomized multicenter dry laboratory training pilot study. *Female Pelvic Med Reconstr Surg*. 2014;20:228-236.
78. Campo R, Reising C, van Belle Y, Nassif J, O'Donovan P, Molinas CR. A valid model for testing and training laparoscopic psychomotor skills. *Gynecol Surg*. 2010;7:133-141.
79. Sheth SS, Fader AN, Tergas AI, Kushnir CL, Green IC. Virtual reality robotic surgical simulation: an analysis of gynecology trainees. *J Surg Educ*. 2014;71:125-132.
80. Dioun SM, Fleming ND, Munsell MF, Lee J, Ramirez PT, Soliman PT. Setting benchmarks for the new user: training on the robotic simulator. *JSLS*. 2017;21:e2017.00059.
81. Akdemir A, Ergenoglu AM, Yeniel AO, Sendag F. Conventional box model training improves laparoscopic skills during salpingectomy on LapSim: a randomized trial. *J Turk Ger Gynecol Assoc*. 2013;14:157-162.
82. Larsen CR, Grantcharov T, Aggarwal R, et al. Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. *Surg Endosc*. 2006;20:1460-1466.
83. Lamblin G, Thiberville G, Druette L, et al. Virtual reality simulation to enhance laparoscopic salpingectomy skills. *J Gynecol Obstet Hum Reprod*. 2019;49:1-7.
84. Aggarwal R, Tully A, Grantcharov T, et al. Virtual reality simulation training can improve technical skills during laparoscopic salpingectomy for ectopic pregnancy. *BJOG*. 2006;113:1382-1387.
85. Goderstad JM, Sandvik L, Fosse E, Lieng M. Development and validation of a general and easy assessable scoring system for laparoscopic skills using a virtual reality simulator. *Eur J Obstet Gynecol Reprod Biol X*. 2019;4:100092.
86. Janssens S, Beckmann M, Bonney D. Introducing a laparoscopic simulation training and credentialing program in gynaecology: an observational study. *Aust N Z J Obstet Gynaecol*. 2015;55:374-378.
87. Mohtashami F, von Dadelszen P, Allaire C. A surgical virtual reality simulator distinguishes between expert gynecologic laparoscopic surgeons and perinatologists. *JSLS*. 2011;15:365-372.
88. Borahay MA, Haver MC, Eastham B, Patel PR, Kilic GS. Modular comparison of laparoscopic and robotic simulation platforms in residency training: a randomized trial. *J Minim Invasive Gynecol*. 2013;20:871-879.
89. Bowring J, Shepherd JH, Ind TE. Assessment of an in vitro model for laparoscopic pelvic lymphadenectomy. *BJOG*. 2007;114:964-969.
90. Bharathan R, Vali S, Setchell T, Miskry T, Darzi A, Aggarwal R. Psychomotor skills and cognitive load training on a virtual reality laparoscopic simulator for tubal surgery is effective. *Eur J Obstet Gynecol Reprod Biol*. 2013;169:347-352.
91. Dawe SR, Pena GN, Windsor JA, et al. Systematic review of skills transfer after surgical simulation-based training. *Br J Surg*. 2014;101:1063-1076.
92. de Win G, van Bruwaene S, Allen C, de Ridder D. Design and implementation of a proficiency-based, structured endoscopy course for medical students applying for a surgical specialty. *Adv Med Educ Pract*. 2013;4:103-115.
93. Ellis R, Cleland J, Scrimgeour DS, et al. Establishing the predictive validity of the intercollegiate membership of the Royal Colleges of surgeons written examination: MRCS part B. *Surgeon*. 2023;21:278-284.
94. Curtis NJ, Foster JD, Miskovic D, et al. Association of surgical skill assessment with clinical outcomes in cancer surgery. *JAMA Surg*. 2020;155:590-598.
95. *How Specialty Trainees in O&G are Assessed*. Accessed November 17, 2023. https://www.rcog.org.uk/for-the-public/about-your-doctor/how-specialty-trainees-in-og-are-assessed
96. Ferriss JS, Frost AS, Heinzman AB, et al. Systematic review of intraoperative assessment tools in minimally invasive gynecologic surgery. *J Minim Invasive Gynecol*. 2021;28:692-697.
97. Lyman WB, Passeri MJ, Murphy K, et al. An objective approach to evaluate novice robotic surgeons using a combination of kinematics and stepwise cumulative sum (CUSUM) analyses. *Surg Endosc*. 2021;35:2765-2772.
98. Mascagni P, Vardazaryan A, Alapatt D, et al. Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Ann Surg*. 2022;275:955-961.
99. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imaging*. 2017;36:86-97.
100. van Amsterdam B, Funke I, Edwards E, et al. Gesture recognition in robotic surgery with multimodal attention. *IEEE Trans Med Imaging*. 2022;41:1677-1687.
101. Gor M, McCloy R, Stone R, Smith A. Virtual reality laparoscopic simulator for assessment in 656 gynaecology. *BJOG*. 2003;110:181-187.
102. Akdemir A, Zeybek B, Ergenoglu AM, Yeniel AO, Sendag F. Effect of spaced training with a box trainer on the acquisition and retention of basic laparoscopic skills. *Int J Gynaecol Obstet*. 2014;127:309-313.
103. Howick J, Chalmers I, Glasziou P, et al. The 2011 Oxford CEBM Levels of Evidence (Introductory Document). Oxford Centre for Evidence-Based Medicine. Accessed July 15, 2023. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence
104. Paquette J, Lemyre M, Vachon-Marceau C, Bujold E, Maheux-Lacroix S. Virtual laparoscopy simulation: a promising pedagogic tool in gynaecology. *JSLS*. 2017;21:e2017.00048.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Tesfai FM, Nagi J, Morrison I, et al. Objective assessment tools in laparoscopic or robotic-assisted gynecological surgery: A systematic review. *Acta Obstet Gynecol Scand*. 2024;00:1-18. doi:10.1111/aogs.14840