Review Article

Systematic Review of Intraoperative Assessment Tools in Minimally Invasive Gynecologic Surgery

J. Stuart Ferriss, MD, Anja S. Frost, MD, Alison Brooks Heinzman, MD, Rita Tsai, MD, Danielle Patterson, MD, Kristen Patzkowsky, MD, Jaime Blanck, MLIS, MPA, and Jessica L. Bienstock, MD

From the Kelly Gynecologic Oncology Service (Dr. Ferriss), Division of Urogynecology (Dr. Patterson), Division of Gynecologic Specialties (Dr. Patzkowsky), Department of Gynecology and Obstetrics (Drs. Frost and Bienstock), Welch Medical Library (Ms. Blanck), Johns Hopkins University School of Medicine, Baltimore, Maryland, and Department of Women's Health, Dell School of Medicine, University of Texas at Austin, Austin, Texas (Drs. Heinzman and Tsai)

ABSTRACT   **Objective:** To collect, summarize, and evaluate the currently available intraoperative rating tools used in abdominal minimally invasive gynecologic surgery (MIGS).
Data Sources: Medline, Embase, and Scopus databases from January 1, 2000, to May 12, 2020.
Methods of Study Selection: A systematic search strategy was designed and executed. Published studies evaluating an assessment tool in abdominal MIGS cases were included. Studies focused on simulation, reviews, and abstracts without a published manuscript were excluded. Risk of bias and methodological quality were assessed for each study.
Tabulation, Integration, and Results: Disparate study methods prevented quantitative synthesis of the data. Ten studies were included in the analysis. The tools were grouped into global (n = 4) and procedure-specific assessments (n = 6). Most studies evaluated small numbers of surgeons and lacked a comparison group to evaluate the effectiveness of the tool. All studies demonstrated content validity and at least 1 dimension of reliability, and 2 have external validity. The intraoperative procedure-specific tools have been more thoroughly evaluated than the global scales.
Conclusion: Procedure-specific intraoperative assessment tools for MIGS cases are more thoroughly evaluated than global tools; however, poor-quality studies and borderline reliability limit their use. Well-designed, controlled studies evaluating the effectiveness of intraoperative assessment tools in MIGS are needed. Journal of Minimally Invasive Gynecology (2021) 28, 692−697. © 2020 AAGL. All rights reserved.

Keywords: Surgical evaluation; Operative teaching; MIGS

A major focus of postgraduate education in gynecology is the attainment of specialized skills in pelvic surgery. The assessment of this core mission varies among programs and is often performed in a manner that is not standardized and subject to bias [1]. Residents and fellows seek to achieve procedural mastery and rely on repetitive practice (case volume), good surgical coaching, and feedback. With the effects of the global pandemic on health systems now evident, some institutions face fluctuating case volumes and pressures to conserve personal protective equipment—which can affect the number of learners allowed per case [2,3]. More than ever, surgical coaching and feedback are essential to allow learners to make each case count. Faculty and learners need tools to help make the development of surgical skill more efficient.

Objective evaluation tools in minimally invasive gynecologic surgery (MIGS) have been described for a variety of settings and procedures. These assessment scales can provide learners with useful, timely feedback that can be integrated rapidly into their practice. Most tools have been developed for simulations, with fewer used intraoperatively [1]. Simulation is a valuable and cost-effective way of providing skill acquisition and assessment in a low-stakes environment. However, a recent survey of obstetrics and gynecology residents noted that few found simulation exercises to be valuable to their learning [4]. Perioperative assessment tools may

prove more useful because they function as a catalyst for a meaningful debriefing discussion once the case is finished [5]. The purpose of this review was to systematically search, collect, summarize, and evaluate intraoperative assessment tools used in abdominal MIGS.

Methods

Registration and Search Strategy

This review was performed in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2009 statement [6]. Details of the protocol for this systematic review were registered on the International Prospective Register of Systematic Reviews (www.crd.york.ac. uk/PROSPERO/display_record.asp?ID=CRD42020191382). This review was exempt from institutional review board review. A comprehensive, systematic literature search was conducted using the PubMed interface for Medline, Embase, and Scopus databases using a combination of database-specific controlled vocabulary and keywords terms for each of the major concepts in the search. The search was designed and executed by a biomedical librarian with extensive experience working on systematic reviews. The search was limited to human studies published between January 1, 2000, and May 12, 2020, written in the English language. The search strategy can be accessed at https://www.crd.york.ac.uk/PROSPEROFILES/191382_STRATEGY_20200808.pdf.

Study Selection

The search results were loaded into Covidence (Covidence, Melbourne, Victoria, Australia) for processing, and duplicates were removed [7]. The software was configured to present abstracts to all reviewers until each abstract had been screened by 2 authors for inclusion or exclusion. Conflicting results were resolved by 1 author (J.S.F.). Articles describing the use of an evaluation tool focused on a minimally invasive abdominal gynecologic procedure in the intraoperative or perioperative setting were included. Reviews, commentaries, abstracts without a manuscript, and articles focused on simulation or minor procedures (e.g., hysteroscopy) were excluded from the review. After abstract screening, the full text of each remaining study was obtained for further review. Full-text screening was completed in a similar manner, with each paper screened by 2 authors for inclusion and any resulting conflicts resolved by 1 author (J.S.F.).

Data Collection and Analysis

Data were collected from the included studies and recorded in an electronic spreadsheet. The collected data included name of the first author, year of publication, whether the tool was a global evaluation or procedure-specific, the name of the evaluation tool, the type of learner, the number of learners evaluated, and any reported measures of validity and reliability. Intraclass correlation (ICC) of 0.80 was used as an acceptable threshold for reliability because values above this number are unlikely to be significantly different [8]. Differences in opinion regarding the extracted data were resolved by consensus. The quality of the data and the risk of bias for each included study were evaluated with the Newcastle-Ottawa Quality Assessment Scale for cohort studies [9]. Owing to differences in study design and methodology, meta-analyses of the data were not possible. However, data were grouped and summarized for tools with overlapping characteristics.

Results

The search returned 3016 unique citations, and 2967 were rejected on the basis of the title and abstract screening (Fig. 1). Full text was obtained for 51 citations, including 2 additional citations that were found through hand searching references of the included papers. Forty-one full-text papers were excluded for the reasons listed in Fig. 1. After review and consensus, 10 papers met our inclusion criteria and were included in the analysis (Table 1). Four of the studies used a global assessment tool, and the remaining 6 used a tool evaluated during a specific minimally invasive procedure: salpingectomy [2], supracervical hysterectomy, total laparoscopic hysterectomy [2], and robotic hysterectomy. The risk of bias and quality assessment using the Newcastle-Ottawa Scale found that all study designs had a risk of bias and were of poor quality, given the lack of meaningful comparison groups.

Global Assessment Tools

Connolly et al [10] reported on the use of myTIPreport (myTIPreport, Richmond, VA) as a platform for surgical skills feedback. This smartphone application was used immediately after operative procedures and included both a learner self-assessment and a faculty assessment. Data from 14 different institutions and a combination of 883 resident- and fellow-level learners were included. This tool has content validity that is based on expert consensus. Construct validity was demonstrated on the basis of the ability to distinguish between resident and fellow performance in a variety of settings. External validity was demonstrated across different institutions. Learner self-assessments and faculty assessments were noted to have a high degree of correlation (Spearman correlation coefficient 0.89, p <.001). Reliability data were not reported.

Fung et al [11] evaluated an interactive voice response instrument to assess resident laparoscopic surgical skills. Twenty-nine residents and 13 faculty raters participated in a total of 809 laparoscopic procedures. Faculty and residents were instructed to call a toll-free number after each laparoscopic case and were prompted to respond to 3 questions using a 5-point Likert scale. Data analysis indicated that the 3 questions had a high degree of correlation and were essentially the same global performance measurement. Nonetheless, the tool demonstrated construct validity

Fig. 1

Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram.
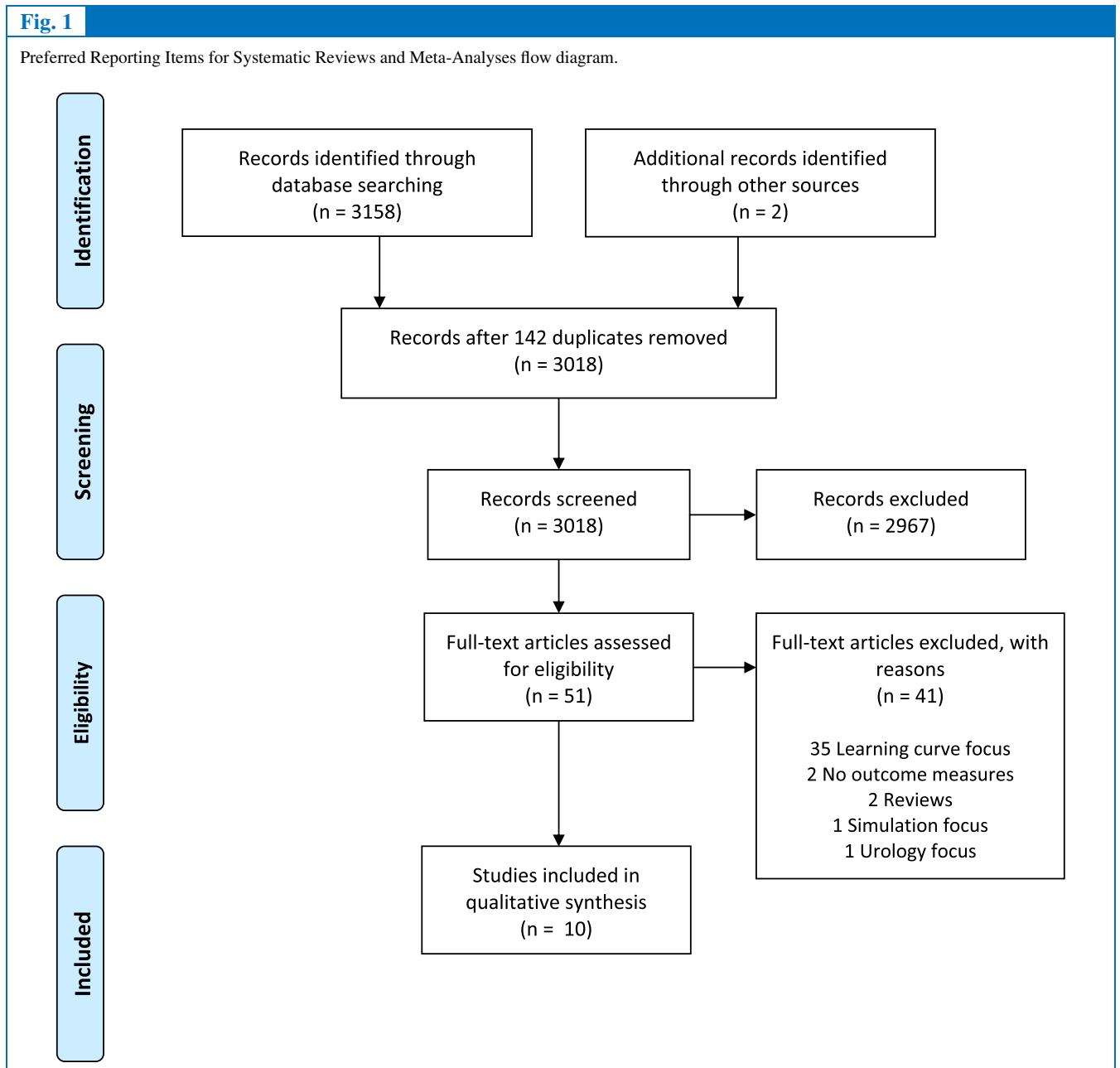
distinguishing among different levels of residents. Inter-rater reliability was acceptable if 12 ratings were used (ICC 0.80). Intra-rater reliability was not reported.

Kilani [12] reported on the Global Rating Index of Technical Skills in gynecologic laparoscopy. Eight surgeons participated by recording their surgeries, and 2 blinded experts reviewed their edited videos and provided a score. Content validity was assessed by expert review. Construct validity was not reported. Inter-rater reliability was very good (ICC 0.96), and intra-rater reliability was not reported.

Shime et al [13] developed and evaluated the Laparoscopic Skills Index as an objective measurement. Twenty laparoscopic surgeries were recorded, half of them performed by residents and the rest by faculty members. These

videos were then scored by 4 blinded reviewers. Content validity was determined by expert review. Construct validity was not reported. Inter-rater reliability approached threshold (ICC 0.77), and intra-rater reliability was not reported.

In summary, 4 intraoperative global assessment tools in MIGS were identified and evaluated. Construct validity was demonstrated in 2 studies, and only 1 of these reported reliability data. Most of the studies are limited by small sample sizes, with a median of 25 (range 8−440) participants. The myTIPreport study used the largest sample of learners and demonstrated external validity. Nevertheless, all the study designs were judged to be of poor quality and susceptible to bias. With these limitations in mind, no individual global assessment tool can be recommended for use over the others.

Table 1

Summary of included studies

Author [citation], year	Assessment context	Name of tool	Subjects (n)	Validity			Reliability		NOS quality assessment (score/total possible)
				Content	Construct	External	Inter-rater (ICC)	Intra-rater (ICC)	
Connolly [10], 2017	Global	myTIPreport	883	Expert consensus	Yes	Yes	NR	NR	Selection 2/4 Comparability 0/2 Outcome 3/3
Fung [11], 2003	Global	IVR	29	NR	Yes	No	0.80	NR	Selection 2/4 Comparability 0/2 Outcome 2/3
Kilani [12], 2018	Global	GRITS	8	Expert consensus	NR	No	>0.90	NR	Selection 2/4 Comparability 0/2 Outcome 3/3
Shime [13], 2002	Global	LSI	20	Expert consensus	NR	No	0.77	NR	Selection 2/4 Comparability 0/2 Outcome 3/3
Larsen [14], 2008	Procedure-specific, salpingectomy	OSA-LS	21	Expert consensus	Yes	No	0.83	NR	Selection 2/4 Comparability 0/2 Outcome 3/3
Oestergaard [15], 2011	Procedure-specific, LS	OSA-LS	20	Previously validated	Yes	Yes	NR	NR	Selection 2/4 Comparability 1/2 Outcome 3/3
Husslein [16], 2015	Procedure-specific, TLH	GERT	14	Expert consensus (general surgery)	Yes	Yes (general surgery)	>0.95	>0.95	Selection 2/4 Comparability 0/2 Outcome 3/3
Savran [17], 2019	Procedure-specific, TLH	OSA-TLH	16	Modified Delphi	Yes	No	0.99	NR	Selection 2/4 Comparability 0/2 Outcome 3/3
Goderstad [18], 2016	Procedure-specific, SCH	CAT-LSH	21	Expert consensus	Yes	No	Intraop rater, 0.75 Blinded rater, 0.85	NR	Selection 2/4 Comparability 0/2 Outcome 3/3
Frederick [19], 2016	Procedure-specific, RH	RHAS	52	Delphi	Yes	No	0.28−0.75	NR	Selection 3/4 Comparability 0/2 Outcome 3/3

CAT-LSH = competence assessment tool for laparoscopic supracervical hysterectomy; GERT = Generic Error Rating Tool; GRITS = Global Rating Index of Technical Skills; ICC = interaclass correlation; Intraop = intraoperative; IVR = interactive voice response; LS = laparoscopic salpingectomy; LSI = Laparoscopic Skills Index; NOS = Newcastle-Ottawa Scale; NR = not reported; OSA-LS = Objective Structured Assessment of Laparoscopic Salpingectomy; OSA-TLH = Objective Structured Assessment of Total Laparoscopic Hysterectomy; RH = robotic hysterectomy; RHAS = robotic hysterectomy assessment score; SCH = supracervical hysterectomy; TLH = total laparoscopic hysterectomy.

Procedure-Specific Tools

Larsen et al [14] refined and evaluated the Objective Structured Assessment of Laparoscopic Salpingectomy (OSA-LS) tool. After a pilot lead-in, 21 consecutive right-sided salpingectomies were performed by surgeons of varying skill (novice, intermediate, and expert), recorded, and then scored by 2 blinded observers. Content validity was evaluated by expert consensus, and construct validity was demonstrated as the ability to discriminate among the 3 types of surgeons. Inter-rater reliability was acceptable (ICC 0.83), and intra-rater reliability was not reported. Oestergaard et al [15] also evaluated the OSA-LS tool seeking to determine if it could be used by surgeons with different levels of expertise; the authors were not evaluating the tool per se. They asked 10 faculty surgeons, 8 residents, and 2 experts to each review 3 blinded videos of a right-sided salpingectomy performed by a novice, intermediate, and expert surgeon, respectively. Construct validity was confirmed, regardless of the level of the assessor; however, only the expert reviewers could consistently distinguish the novice and expert videos from the intermediate one. External validity was also demonstrated by successfully applying the tool using a new population. Reliability statistics were not reported.

Husslein et al [16] evaluated a tool measuring technical errors while performing a total laparoscopic hysterectomy, the Generic Error Rating Tool (GERT). They asked 2 blinded reviewers to evaluate recorded cases from 14 surgeons (who had varying levels of training) using both the GERT and the Objective Structured Assessment of Technical Skills (OSATS) tool. The evaluators were trained on the tool before receiving the videos. Using the OSATS scores to cluster the videos into low or high performers, the GERT was able to discriminate between the 2 groups, demonstrating a measure of construct validity. The GERT demonstrated very good inter-rater and intra-rater reliability scores (ICC >0.95). There was a significant negative correlation between increasing number of errors and total OSATS scores.

Savran et al [17] developed and evaluated a formative assessment tool for total laparoscopic hysterectomy that was based on OSATS. A modified Delphi process ensured content validity. Video recordings of total laparoscopic hysterectomy performed by 8 beginner surgeons and 8 experienced surgeons (n = 16) were reviewed by 2 blinded raters. The rating scale was able to distinguish between the 2 groups, demonstrating construct validity. Inter-rater reliability was very high (ICC 0.99), and intra-rater reliability was not reported. Furthermore, the authors were able to describe a cut point to be used in a pass/fail evaluation.

Goderstad et al [18] developed and evaluated a competence assessment tool for use in laparoscopic supracervical hysterectomy. Content validity was established by expert consensus. The participants were divided into 3 groups: inexperienced, intermediate experience, and expert. An intraoperative observer scored each surgeon, and 2 blinded observers also scored a video of the surgery. Thirty-seven procedures by 21 individual surgeons were evaluated. The competence assessment tool for use in laparoscopic supracervical hysterectomy score successfully discriminated among the 3 groups of surgeons (construct validity), regardless of observer (intraoperative or blinded). However, the score of the intraoperative observer was consistently higher than that of the blinded ones. Inter-rater reliability was acceptable (ICC 0.85) for the 2 blinded raters, but below threshold for the intraoperative rater (ICC 0.75). Intra-rater reliability was not reported.

Frederick et al [19] developed and evaluated the Robotic Hysterectomy Assessment Score. Delphi methodology was used to arrive at consensus and ensure content validity. The participating surgeons were divided into 3 groups: novice, advanced beginners, and experts. Fifty-two surgeon videos were created from 26 surgeries (1 surgeon per side) and were scored by an expert panel. The rating tool demonstrated construct validity by successfully discriminating among the 3 groups of surgeons. Inter-rater reliability was evaluated for each element of the tool, and all fell below threshold (ICC range 0.28−0.75). Intra-rater reliability was not reported.

In summary, 6 studies evaluating 5 different intraoperative assessment tools during specific MIGS procedures were identified and evaluated. Most studies were small with a median of 21 (range 14−52) participants. The study designs were noted to be susceptible to bias and were judged to be of poor quality, given the lack of comparison groups. The study by Oestergaard et al [15] is a notable exception because this study included a comparison group despite the small sample size. In addition, this study applied the OSA-LS tool in a setting unrelated to prior work, providing evidence of external validity. Currently, the OSA-LS intraoperative assessment tool has the most data supporting its use among those evaluated.

Discussion

Surgical coaching with feedback to the learner is a necessary component in the development of surgical expertise [20]. Although surgical coaching has a long tradition in medical education, the use of objective assessment tools as a means to help accomplish this goal is relatively new and has not been universally adopted in the teaching of minimally invasive gynecologic procedures. This review focused on intraoperative assessment tools in abdominal MIGS cases. These instruments promise to provide immediate and standardized feedback to learners. This type of feedback is known to be essential to a successful surgical training program [20]. To fulfill their promise, these tools need to be feasible (not time intensive or complicated to use), have validity (content, construct, and external), and also have acceptable levels of reliability for the same rater (intra-rater) and among different raters (inter-rater) [1].

Furthermore, if they are going to be used in the educational setting, there should be evidence that these tools are both effective as a teaching aid and acceptable to learners.

Tools to assess surgical skills that are designed to be used intraoperatively or immediately after a surgical procedure are part of a larger framework of surgical training. A structured curriculum with simulation, coaching, and immediate feedback with debriefing are well-recognized pillars of surgical education [20−22]. Although emerging data supporting learner assessments presented here and elsewhere are promising, surgical coaching remains an area in need of systematic study and objective evaluation [23].

The limitations of this review could arise from incomplete search results or errors in the screening and review process. Steps were taken to ensure the broadest search terms possible, and 1 of the authors (J.B.) is a librarian with extensive experience in this type of work. Adherence to best practices such as requiring more than 1 reviewer for every abstract and full-text review helps insure against user error. This review is also limited by the quality of evidence from the included studies. Finally, the heterogenous methods used in these studies prevented quantitative synthesis of the data.

To conclude, on the basis of this review, the current menu of available intraoperative assessment tools is small, and the quality of evidence supporting their use is poor. The quality of the data is a reflection of the early evolution of these tools, more so than any shortcomings of the included studies. Each assessment tool must undergo a series of evaluations before it can be declared both useful and effective. Therefore, additional work in larger populations is needed to further characterize and refine the assessments discussed here. Most notably, well-designed studies with appropriate comparison groups are currently lacking. This work will be necessary before these assessments can be used in high-stakes (pass/fail) evaluations either within training programs or for the purpose of board certification. Interesting new avenues of investigation include emerging data that indicate a possible role for automated skill assessment or crowdsourcing surgical skill assessment, either of which could accelerate the pace of innovation in this area [24,25].

References

1. Clark NV, Pepin KJ, Einarsson JI. Surgical skills assessment tools in gynecology. *Curr Opin Obstet Gynecol*. 2018;30:331–336.
2. Russell SW, Ahuja N, Patel A, O'Rourke P, Desai SV, Garibaldi BT. Peabody's paradox: balancing patient care and medical education in a pandemic. *J Grad Med Educ*. 2020;12:264–268.
3. Nakayama J, El-Nashar SA, Waggoner S, Traughber B, Kesterson J. Adjusting to the new reality: evaluation of early practice pattern adaptations to the COVID-19 pandemic. *Gynecol Oncol*. 2020;158:256–261.
4. Stairs J, Bergey BW, Maguire F, Scott S. Motivation to access laparoscopic skills training: results of a Canadian survey of obstetrics and gynecology residents. *PLoS One*. 2020;15:e0230931.
5. Connolly A, Hansen D, Schuler K, Galvin SL, Wolfe H. Immediate surgical skills feedback in the operating room using "SurF" cards. *J Grad Med Educ*. 2014;6:774–778.
6. 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.
7. Covidence. Covidence systematic review software: Veritas health innovation, Melbourne, Australia. Available at: Www.Covidence.org Accessed March 1, 2020.
8. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *J Strength Cond Res*. 2005;19:231–240.
9. Wells GA, Shea B, O'connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses 2012. Available at: http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp Accessed April 30, 2020.
10. Connolly A, Blanchard A, Goepfert A, et al. Surgical skills feedback and myTIPreport: is there construct validity? *Obstet Gynecol*. 2017;130(Suppl 1):17S–23S.
11. Fung Kee Fung K, Fung Kee Fung M, Bordage G, Norman G. Interactive voice response to assess residents' laparoscopic skills: an instrument validation study. *Am J Obstet Gynecol*. 2003;189:674–678.
12. Kilani R. Comparing self-assessment of laparoscopic technical skills with expert opinion for gynecological surgeons in an operative setting. *Gynecol Surg*. 2018;15:16.
13. Shime J, Pittini R, Szalai JP. Reliability study of the laparoscopic skills index (LSI): a new measure of gynaecologic laparoscopic surgical skills. *J Obstet Gynaecol Can*. 2003;25:186–194.
14. 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.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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.
20. Fagotti A, Petrillo M, Rossitto C, Scambia G. Standardized training programmes for advanced laparoscopic gynaecological surgery. *Curr Opin Obstet Gynecol*. 2013;25:327–331.
21. Aggarwal R, Moorthy K, Darzi A. Laparoscopic skills training and assessment. *Br J Surg*. 2004;91:1549–1558.
22. Goff BA, Lentz GM, Lee D, Houmard B, Mandel LS. Development of an objective structured assessment of technical skills for obstetric and gynecology residents. *Obstet Gynecol*. 2000;96:146–150.
23. Sampene KC, Littleton EB, Kanter SL, Sutkin G. Preventing error in the operating room: five teaching strategies for high-stakes learning. *J Surg Res*. 2019;236:12–21.
24. Baghdadi A, Hussein AA, Ahmed Y, Cavuoto LA, Guru KA. A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos. *Int J Comput Assist Radiol Surg*. 2019;14:697–707.
25. Lendvay TS, White L, Kowalewski T. Crowdsourcing to assess surgical skill. *JAMA Surg*. 2015;150:1086–1087.