# Tools for measuring technical skills during gynaecologic surgery: a scoping review

Louise Hennings[1], Jette Sorensen[2], and Jeannet Strandbygaard[2]

[1]Herlev Hospital
[2]Copenhagen University Hospital Rigshospitalet

June 1, 2020

**Abstract**

Standardised assessment is key to structured surgical training. Currently, there is no consensus on the effectiveness of surgical assessment tools in gynaecologic surgery. The purpose of this review is to identify and assess measurement characteristics for assessment tools for measuring technical skills in gynaecologic surgery. Eight studies out of the 544 identified fulfilled the inclusion criteria. The assessment tools were categorised as global rating scales, global and procedure rating scales combined, task-specific rating scales or as non-procedure-specific error assessment. A combination of global and task-specific assessment tools appears to be the most comprehensive solution for observational assessment of technical skills

**Funding**

None

**Tweetable abstract**

Global & task-specific assessment tools are good for assessment of technical skills in gynaecological surgery.

**Keywords**Assessment, assessment tool, gynaecology, surgery

## Introduction

Assessment can be used to establish a current proficiency level, track progress over time and facilitate a learning process. When given as constructive feedback, assessment motivates development, and standardised assessment aids structured surgical training. Therefore, there is a need for assessment of surgical competencies. Without assessment, important knowledge and the potential for progression can be lost. The value of assessing a surgeon's competencies is hence indisputable, but requires a trained assessor and an objective structured assessment tool.[1] Currently, there is no consensus on the effectiveness of surgical assessment tools in gynaecology.

Surgical training has been shown to improve surgical skills, and several assessment tools have been validated in both live surgical settings and in a simulated environment.[2] When assessing surgical or procedural skills, the choice of assessment tool depends on whether the performance is evaluated in a simulated or live surgical setting. More than 20 years ago Van der Vleuten explored and described five criteria (reliability, validity, impact on future behaviour, acceptability and costs) to take into consideration when choosing an assessment method in the clinical setting.[3] They remain highly relevant, but especially reliability and validity must be thoroughly tested when applying an assessment tool.

Both task-specific and global rating tools are widely used in a variety of specialties. The tools use various scoring systems, e.g. binary checklists or anchors, such as a Likert scale. In general surgery, a number of task-specific checklists exist, but recent reviews showed a lack of validity and reliability.[4,5]

Implementation of objective assessment in clinical practice is difficult due to challenges on many levels: lack of time, lack of resources and often also lack of knowledge on how and when to use an assessment tool. To overcome these challenges it is important that the chosen assessment tool is acceptable, feasible, valid, well-described and easy to apply in a surgical setting.[3]There is an ongoing debate on when assessment should be performed and which form of assessment should be used to evaluate a given performance.[1,6] Kane states that a procedure evaluated in simulation settings cannot be transferred to the operating room, indicating the importance of validating an assessment tool in the environment it will be used in.[7]

The aim was to conduct a systematic scoping review to identify assessment tools measuring technical skills during gynaecologic surgery and to assess the measurement characteristics of the tools used in a clinical setting in the operating room.

## Methods

We chose the scoping review methodology to characterises the quantity and quality of existing assessment scales.[8] Conducted in accordance with Arksey and O'Malley,[9] the review was designed to cover all available literature on the topic, to summarise existing knowledge and to identify research gaps in the current literature. The underlying methodological framework comprised five consecutively linked stages (Table S2). Levac et al., who further developed Arksy and O'Malley's approach in order to clarify and enhance the various stages,[10] recommend an optional sixth stage that involves consulting stakeholders.

Evaluation of each study focused especially on stage four of Arksey and O'Malley's methodology and involved a careful examination of the design, observation method and domains assessed in each of the included studies.

The review is reported using the principles laid out in Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR).[11] This method of evidence synthesis was selected to summarise and disseminate research findings, and the aim of this scoping review was to identify whether one or more validated assessment tools exist that can be applied to assess technical surgical skills at both trainee and specialist level in the operating room in gynaecologic surgery. We excluded studies assessing surgical performance on animals and ones that tested tools in a simulated setting, which means that only studies analysing the assessment tool in live surgery were included.

We applied Kane's validity argument, which comprises four inferences (Table S3) to evaluate the various assessment tools.[12]The four inferences link an observation to a score, which then estimates the performance in a test setting. This performance provides an estimate of performance in live surgery, which leads to an action/decision. Inspired by a recent systematic review by Hatala et al. that used Kane's validity argument to evaluate an assessment tool, we chose Kane's framework approach as it offers key phases or inferences when planning and evaluating the validity argument.[13]

In this review, the term assessment tool refers to a specific tool that assesses specific surgical competencies, whereas the word scale refers to a widely applicable assessment tool or a component of a specific tool.

### Search strategy

In accordance with stages one and two of Arksey and O'Malley's methodology, we used a broad search strategy guided by the aid of an information specialist (PP). We identified keywords and created a search string (Appendix S1). Two researchers (LIH and JS) worked independently searching the databases and then screened records by title/abstract and, finally, full texts articles were read. In the event of disagreement on inclusion of an article, it was discussed until consensus could be reached. Four databases were searched covering 1989-2000: PubMed, Medline, Embase and Cochrane.

Inclusion/exclusion criteria

The search covered articles from 1 January 1989 up to 15 January 2020 and was limited to literature published in English. Our review only included studies that analysed assessment tools in live gynaecologic surgery.

## Results

Figure S1 contains a flowchart of the reference search, and Table S1 presents an overview of study characteristics for the eight articles that met our inclusion criteria.

**Data synthesis**The measurement characteristics, i.e. validity, reliability and validation context, are summarised in the following for each type of assessment tool: 1) global rating scale, 2) global and procedure rating tools combined, 3) task-specific rating tools and 4) non-procedure-specific error assessment. Table S4 presents an overview of each assessment tool using Kane's validity argument.

**1) Global rating scales*Objective Structured Assessment of Technical Skills (OSATS)***

Currently, the most widely used and validated assessment scale is OSATS,[14] which originally consisted of a task-specific checklist and a global rating scale, the latter of which has been shown to have high reliability and validity and to be applicable at various trainee levels and for a variety of surgical procedures.[15]

Hiemstra et al.[16] present the use of an objective assessment tool as a way to establish learning curves and analyse the OSATS scores of nine trainees over a three-month period. Nineteen types of procedures were identified among the 319 they assessed.

The surgical procedures consisted of abdominal hysterectomy (39%), labioplasty (31%), a vaginal approach (20%) and hysteroscopies (10%).

The trainees were instructed to fill out an OSATS assessment sheet after every procedure. A consultant would then perform supervision, discuss the result with the trainee and provide constructive feedback. Within the six OSATS domains, scores range from 6 to 30 points, and a score of 24 was the selected threshold for good surgical performance.

To prove construct validity, the authors hypothesise that surgical performance improves over time, with increasing procedure-specific experience.[16] They found that performance improved 1.10 OSATS points per assessed procedure (p=0.008, 95% confidence interval (CI) 0.44–1.77) and that the learning curve for a specific procedure passed the threshold of 24 points at a caseload of five procedures. Furthermore, a performance plateau was reached after performing eight of the same procedures.

**2) Global and procedure-specific assessment tools combined**

Vaginal Surgical Skills Index (VSSI)

Chen et al.[17] introduced VSSI, which is a procedure-specific rating scale for evaluating surgeons while performing vaginal hysterectomies. VVSI is an expanded and modified form of the original seven-item Global Rating Scale (GRS). [14] They developed and adjusted GRS to contain items considered important for vaginal surgery, 13 in total, including: initial inspection, performance of an incision, maintenance of visibility, use of electro surgery, knot tying, maintenance of haemostasis, removing fluid and debris, completion of the procedure and forward planning.

Twenty-seven trainees performed 76 surgeries in the study period. The operating surgeon was evaluated using VSSI as soon as the hysterectomy was performed. VSSI was then compared with GRS scores and good construct validity was found. The procedure was videotaped, and to assess interrater reliability, a blinded reviewer evaluated the performing surgeon using VSSI, GRS and a 100-mm visual analogue scale, which was included as an additional measure to furnish the assessor with a global impression of the trainees' surgical skills.

To assess intrarater reliability, the supervising surgeon watched and re-evaluated the video after four weeks. Internal consistency for VSSI and GRS was high (Cronbach's alpha=0.95-0-97). Using intraclass correlation coefficient (ICC), the authors concluded that the VSSI interrater reliability (ICC=0.53) and intrarater reliability (ICC=0.82) were good.

Hopkins Assessment of Surgical Competency (HASC)

Chou et al. developed the Hopkins Assessment of Surgical Competency (HASC), which contains two assessment scales, General Surgical Skills and Case Specific Surgical Skills, to measure trainee surgical competency in gynaecologic surgery.[18]. With the exception of oncology cases, all levels of surgical complexity were covered, from hysterectomy to dilation of cervix to urogynaecology.

The assessment form contains seven items from OSATS[14]and four from an American Council on Resident Education in Obstetrics and Gynecology toolbox. Another four items were included from a rating system already in use at Chou et al-s' institution. After modifying all the items using factor analysis, the results were divided into two six-item scales, a case-specific one and a general surgical one, each item scored using a five-point Likert scale. Sixteen faculty physicians evaluated 16 trainees after the trainees had performed surgery. The trainees also performed self-evaluation.

Their study analysed 362 cases, and the authors find internal validity and reliability, demonstrated by high Cronbach's alpha (>.80) and high Pearson correlation coefficients (>0.80) for both scales. Discriminant validity was also significant ($p<0.001$) for both scales when comparing the performance of trainees in their second and fourth year of training.

Objective Structured Assessment of Laparoscopic Salpingectomy (OSALS)

Larsen et al. [19] carried out a blinded prospective cohort study that developed OSALS, which is a method designed to conduct an objective structured assessment of technical surgical skills based on laparoscopic salpingectomy. Like the HASC, OSALS consists of a general rating scale based on OSATS and on a case-specific scale, but three of the task-specific items are directly related to the procedure evaluated: laparoscopic salpingectomy. Two independent observers used the OSALS chart for assessment of 21 unedited video recordings of 21 laparoscopic salpingectomies, performed by 21 different surgeons, grouped as either novices, intermediate or experts. The median score in each group showed that the OSALS tool was construct valid and able to discriminate between all groups ($p<0.03$).

The overall interrater agreement was 0.831, varying from 0.759 in the experienced group to 0.905 in the intermediate group. There was a wide performance range in the expert group and a narrow performance range in the novice group.

**3) Task-specific assessment tools**

Robotic Hysterectomy Assessment Score (RHAS)

An observational study by Frederick et al.[20] led to the development and validation of Robotic Hysterectomy Assessment Score (RHAS) based on the study and evaluation of live video-recorded procedures assessing six surgical domains on a five-point Likert scale. Each domain was subdivided into specific tasks and a maximum score of 80 was possible. Delphi methodology was used for content validation of the six surgical domains. The evaluation covered 25 expert, 20 beginner and 7 novice surgeons.

Interrater reliability was acceptable, and ICC varied from 0.605 to 0.748 for domains 1, 2, 3, 4 and 6. Total ICC was 0.600 ($p=0.001$).). RHAS demonstrated the ability to differentiate between experts, advanced beginners and novices, and the median overall scores for the three categories of surgeons were 75.25 for experts, 72.25 for advanced beginners and 70.25 for novices ($p=0.006$)

Competence Assessment Tool for Laparoscopic Supracervical Hysterectomy (CAT-LSH)

Goderstad et al.[21] developed CAT-LSH, a procedure-specific rating tool for laparoscopic supracervical hysterectomy and compared it with Global Operative Assessment of Laparoscopic Skills (GOALS) [22], a general rating scale. GOALS has been validated for laparoscopic ventral hernia repair, laparoscopic appendectomy, laparoscopic inguinal hernia and laparoscopic and open cholecystectomy.[4]

By defining the main steps of the hysterectomy procedure, CAT-LSH assesses ligament mobilisation, release of adnexa form uterus, division of uterine vessels and uterus amputation. Each step evaluates the use of instruments, tissue handling and errors, with a maximum of 16 points assigned per step for a total possible

score of 64. The procedure was recorded, and the performing surgeon was evaluated by both the operating assistant and by blinded reviewers on video footage. Twenty-one doctors performed 37 procedures eligible for blinded assessment.

The study found that GOALS allowed blinded observers to differentiate between inexperienced and intermediate experienced surgeons, but not between intermediate experienced surgeons compared to expert surgeons (p=0.085). When performed by the assistant surgeon, the GOALS assessment differed significantly between the three groups. When the assistant surgeon performed assessments using CAT-LSH, it was able to discriminate between inexperienced and intermediate experienced surgeons (p=0.001) and intermediate experienced surgeons and experts (p=0.001), demonstrating a significant discriminative validity. The interrater reliability comparing the mean scores for CAT-LSH and GOALS showed good agreement with an ICC>0.75. The study concluded that CAT-LSH has construct validity and is feasible in live surgical assessment with a significant discriminative validity.

A feasible rating scale for formative and summative feedback

Savran et al., who had nine experienced gynaecologists on their team, used Delphi methodology to develop the most recent procedure-specific rating scale, which is a feasible rating scale for formative and summative feedback.[23] The scale comprises 12 items evaluated on a five-point Likert scale. Messick's framework was used to measure the validity evidence. Grouped as beginners (had performed <10 procedures) or experienced surgeons (had performed >200 procedures), 16 surgeons performed 16 laparoscopic hysterectomies. The procedure was video recorded and analysed by two blinded reviewers.

The authors found internal consistency reliability with high Cronbach's alpha 0.95 (p<0.001) and high interrater reliability (ICC=0.996) for one rater and ICC=0.998 for two raters. The mean scores of the beginners versus the experienced surgeons were significantly different (p<0.001).

Savran et al. concluded that the tool is suitable for both formative and summative assessment.

**4) Non-procedure-specific Error assessment**

Generic Error Rating Tool (GERT)

Husslein et al.[24] were the first to test a non-procedure-specific error assessment tool. Called GERT, the tool, which uses a Likert scale with nine anchors, is designed to analyse technical errors and resulting events during laparoscopy. GERT is based on the inverse relationship between surgeon and skill, i.e. more skilled surgeons make fewer errors.

Technical errors are defined as "the failure of planned actions to achieve their desired goal" and an event as "an action that may require additional measures to avoid an adverse outcome".[24] The GERT technical error analysis comprises nine generic surgical tasks during which errors can occur. Each of these generic task groups is subdivided into four distinct error modes: 1) too much use of force or distance, 2) too little use of force or distance, 3) inadequate visualisation and 4) wrong orientation of instrument. To assess error distribution within different operative sub-steps, the procedures are divided into insertion of trocars, creation of bladder flap, colpotomy and vault closure.

Two blinded reviewers analysed twenty video recordings of total laparoscopic hysterectomies, and correlation analyses were performed between GERT and OSATS. Scores from the latter were used to establish a measure of technical skills and to divide surgeons into two groups as either high or low performers. The results showed a significant negative correlation between OSATS and GERT scores (rater 1: Spearman = -0.76, p<0.001; rater 2 = -0.88, p< 0.001). Group comparison showed that high performers made significantly fewer technical errors than low performers.

Interrater reliability was high (CCI>0.95) for the total number of errors and events. Within the nine anchors (task groups) ICC was >0.8 in all groups except for cutting, transection and stapling; clipping; and use of suction. Intrarater reliability was high (ICC>0.95) for total number of errors and events. By analysing the

different operative sub-steps, the study was able to detect procedures more prone to technical errors, e.g. vault closure.

## Discussion

### Main findings

There is a need for robust validated tools across different measurement properties in order to aid surgical educators in selecting the appropriate tool for assessment. This systematic review identified eight technical assessment tools validated during gynaecologic surgery. The studies, which have different validity strengths according to Kane's framework, present a variety of challenges.

Hiemstra et al.'s[16] study tested the OSATS intraoperatively to establish learning curves for each trainee either using direct supervision or self-assessment. As expected, learning curves were established but the authors identified enormous variation in assessors' OSATS scores, and the trainees reported a lack of objectivity in the assessment tool. This important limitation of the OSATS, according to Kane's validity argument, it does not meet the extrapolation criteria.

VSSI [17] was developed as a procedure-specific rating tool to assess surgeons while performing vaginal hysterectomies. Interestingly, the 13 items in the Likert scale are not procedure specific and can be applied to laparoscopic surgery in general. A limitation is that this transfer of general competencies to a specific rating tool did not prove to be appropriate. Importantly, the authors focus on case mix, where a specific (patient) characteristic is known to potentially effect (surgical) outcome. A recent review on case-mix variables and predictors for outcomes of laparoscopic hysterectomy showed that body mass index, previous operations, adhesions and age were predominate case-mix characteristics.[25]This knowledge on case mix is important when choosing a surgical case for assessment.

Chou et al. modified an existing global rating scale by adding procedure specific items to develop HASC [18], which targets gynaecologic trainees and aims to evaluate all surgical competencies in gynaecologic surgery. This procedure-specific rating tool is applicable to all types of laparoscopic surgery. The generalizability and lack of a task-specific checklist makes HASC applicable to other gynaecologic programmes. To our knowledge, this applicability has not been demonstrated in other validated studies. The study was not blinded, only trainees were tested and data were collected for all types of surgical procedures, lowering the strength of the study.

The OSALS rating tools is incorporated in the Danish curriculum for assessment of OBGYN trainees. [1926].It comprises five general and five task-specific items and was developed and validated in a blinded study[19]. There was a wide performance range in the expert group and a narrow performance range in the novice group, which could be explained by case mix and by the fact that categorising surgeons as intermediate or expert can be difficult. The study is limited by a small sample size.

Arguably, a disadvantage of video evaluation is that it is time-consuming, but Larsen et al. underline it as a strength for the objective assessment, an assertion that Langermann et al. support, arguing that video recording in the operating theatre enhances and supports surgical training and can be performed equally good by doctors with different expeince.[27] [28]

Six of the included studies used video recording and blinded observers when evaluating the surgeon's performance. All of the studies found significant discriminative validity, demonstrating that the assessor can differentiate between novices, advanced beginners and experts. This indicates that video-recorded assessment is a good choice when validating an assessment tool, but as it is time-consuming, it may not be an obvious choice for implementation in daily clinical practice.[29]

### Strengths and limitations

Even though the development of content validity for procedure-specific assessment tools requires using Delphi methodology, which is a consensus-based approach,[30] of the eight studies in our scoping review, only Frederick et al. did so.[20]They discussed the potentially confounding variable of the attending physician providing direct

supervision and guidance when evaluating a novice surgeon. This may account for why novice surgeons' RHAS[20] scores did not differ more relative to their more experienced colleagues. Case mix may also explain this lack of difference in scores.

RHAS [20] demonstrated both construct and discriminative validity and appears to be feasible. It is argued that many of the skills RHAS measures can also be applied to hysterectomies performed either laparoscopically or abdominally, as the basic steps in the procedure are identical. The study's intent is to facilitate surgical training by tracking progress over time and to give immediate and constructive feedback to trainees.

The procedure-specific rating tool CAT-LSH was superior in terms of discriminative validity compared to the validated tool Global Operative Assessment of Laparoscopic Skills (GOALS) [22] used for laparoscopy. Goderstad et al. asserted that this is the case because CAT-LSH is more detailed for each step of the procedure compared to GOALS, a finding supported by Frederick. [20] The study, which uncovered another challenge when assessment is done by non-blinded observers, showed that the operating assistant gave a higher total score than the blinded reviewer, both in terms of GOALS and CAT-LSH in all three groups. A reasonable explanation is a cognitive bias, e.g. confirmation bias or stereotype bias.

Even though GOALS [22] is used as a comparison in the CAT-LSH study [21], the general rating scale has never been tested and validated in a gynaecologic surgical setting. Interestingly, that is also the case for the most widely used global assessment scale, OSATS. A comprehensive study by Hatala et al.[13] thoroughly analysed the validity evidence for OSATS in a simulating setting, but the global rating scale must still be validated in a real-life clinical operating room in gynaecologic surgery.

Husslein et al. examined GERT [24], which was able to significantly discriminate between low and high performers by analysing errors. The study also identified procedures more prone to technical errors, which is important knowledge when determining the focus of a procedure-specific assessment tool and how detailed each procedural step should be evaluated. The study is limited by a small sample size and the fact that the videos were retained from a previous study.

**Interpretation**

A systematic review by Ahmed et al.[31] concluded that a combination of global and task-specific assessment tools appears to be the most comprehensive solution for observational assessment of technical skills. This is supported by findings in the RHAS, CAT-LSH and OSALS, tools which all consist of a general and procedure-specific checklist and are validated in studies with relatively strong methodology. It has been shown in a simulation setting that evaluation of a clinical competence solely using a procedure-specific checklist does not preclude incompetence in terms of technical ability and safety.[32] Identifying safety issues requires the inclusion of assessment using a global rating scale. By adding GERT the operative substeps prone to errors, can be identified.

Savran et al.[23] asserted that their assessment tool meets the criteria for summative assessment, using the contrasting group method to set a pass/fail score. Similar to most studies in our scoping review, the authors grouped the surgeons according to surgical load, with experienced or expert surgeons defined according to the number of cases performed, even though this is not an objective measure of competency, just as a pre-set standard must exist to establish summative assessment.[1]

Focused on formative feedback, high-stakes assessment and programme evaluation, Hatala et al.[13] used Kane's framework to evaluate OSATS and found reasonable evidence in terms of scoring and extrapolation for formative and high-stakes assessment. For programme assessment, there was validity evidence for generalisation and extrapolation but a complete lack of evidence regarding implications and decisions based on OSATS scores. This calls for more research.

**Conclusion**

We identified eight tools measuring technical skills during gynaecologic surgery, all of which depend on user context, with varying validity frameworks. A combination of global and task-specific assessment tools

with a focus on operative substeps prone to errors appears to be the most adequate way to assess surgical competencies in gynaecology. Our systematic review can serve as a guide for surgical educators who wish to evaluate surgical assessment. When choosing a tool it must be determined whether an assessment is for formative or summative assessment, just as it much have strong construct validity tested in the gynaecologic operating room.

**Acknowledgements**Pernille Pless (PP), information specialist, who guided the search strategy.**Disclosure of interests**

None**Contribution to authorship**

Louise Inkeri Hennings (LIH), Jette Led Sørensen (JLS) and Jeanett Strandbygaard (JS) contributed to the authorship. LIH conceived the study. LIH and JS constructed the scoping review with input from JLS. LIH and JS performed searches, screening and data extraction. LIH and JS analysed the data. LIH was responsible for writing the first draft and all authors contributed to finalising the manuscript.**Details of ethics approval**

Ethics approval not compulsory

**Funding**None

**References**

1. Wass V, Van Der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945-949.

2. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality. Acad Med. 2013;88(6):872-883.

3. Vleuten C Van Der. The assessment of professional competence: developments, research and practical implications. Adv Heal Sci Educ. 1996;1:41-67.

4. Ghaderi I, Manji F, Soo Park Y, Dorthea Juul, Ott, Ilene Harris et al. Technical skills assessment toolbox a review using the unitary framework of validity. Ann Surg. 2015;261(2):251-262.

5. Jelovsek JE, Kow N, Diwadkar GB. Tools for the direct observation and assessment of psychomotor skills in medical trainees: A systematic review. Med Educ. 2013;47(7):650-673

6. Epstein RM. Medical education - Assessment in medical education. N Engl J Med. 2007;356(4):387-396.

7. Kane MT. The assessment of professionel competence. Evaluation & the Health Professions. 1992;15:163-182

8. Grant MJ, Booth A. A typology of reviews: An analysis of 14 review types and associated methodologies. Health Info Libr J. 2009;26(2):91-108

9. Arksey H, O 'malley L. Scoping Studies: Towards a Methodological Framework. Int J Soc Res Methodol. 2005;8(1):19-32

10. Levac D, Colquhoun H, O'Brien KK. Scoping studes: advancing the methodology. Implement Sciene. 2010;5(69):1-9

11. Tricco AC, Lillie E, Zarin W, O'Brien KK , Heather Colquhoun, Danielle Levac et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-473

12. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49:560-575

13. Hatala R, Cook DA, Brydges R, Hawkins R. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. Adv Heal Sci Educ. 2015;20:1149-1175

14. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghanet J, Hutchison C et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84(2):273-278

15. Szasz P, Louridas M, Harris KA, Aggarwal R, Grantcharov TP. Assessing technical competence in surgical trainees: A systematic review. Ann Surg. 2015;261(6):1046-1055.

16. Hiemstra E, Kolkman W, Wolterbeek R, Trimbos B, Jansen FW. Value of an objective assessment tool in the operating room. Can J Surg. 2011;54(2):116-122.

17. Chen CCG, Korn A, Klingele C, Barber MD, Paraiso MFR, Walters MD et al. Objective assessment of vaginal surgical skills. Am J Obstet Gynecol. 2010; 203:79.e1-8

18. Chou B, Bowen CW, Handa VL. Evaluating the competency of gynecology residents in the operating room: validation of a new assessment tool. Am J Obstet Gynecol. 2008; 199:571.e1-571.e5

19. 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 An Int J Obstet Gynaecol. 2008;115(7):908-916

20. Frederick PJ, Szender JB, Hussein AA, Kesterson JP, Shelton JA, Anderson TL et al. Surgical Competency for Robot-Assisted Hysterectomy: Development and Validation of a Robotic Hysterectomy Assessment Score (RHAS). Journal of Minimally Invasive Gynecology; 2017;24:55-61.

21. 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(4):600-608.

22. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré Ket, Stanbrigde D al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg. 2005;190(1):107-113.

23. 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.

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(6):1259-1265.

25. Driessen SRC, Sandberg EM, Chapelle CF, Twijnstra ARH, Rhemrev JPT, Jansen FW. Case-Mix Variables and Predictors for Outcomes of Laparoscopic Hysterectomy : A Systematic Review. J Minim Invasive Gynecol. 2016;23(3):317-330.

26. 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(4):359-366.

27. Langerman A, Grantcharov TP. Are We Ready for Our Close-up? Ann Surg. 2017;266(6):934-936.

28. 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(7):2054-2060.

29. Strandbygaard J, Scheele F, Sorensen JL. Twelve tips for assessing surgical performance and use of technical assessment scales. Med Teach. 2017;39(1):32-37.

30. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and Nominal Group in medical education research*. Med Teach. 2017;39(1):14-19. doi:10.1080/0142159X.2017.1245856

31. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: A systematic review. Am J Surg. 2011;202:469-480.

32. Ma IWY, Zalunardo N, Pachev G, Beran T, Brown M, Hatala R et al. Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. Adv Heal Sci Educ. 2012;17(4):457-470.

**Table S1** Overview of studies identified for inclusion.

| Author and year | Assessment tool | Study design and assessment method | Domains assessed |
|---|---|---|---|
| **Hiemstra et al.** [21] 2011 | **OSATS Objective Structured Assessment of Technical Skills** | *Observational study* Self-assessment and peri- and postoperative assessment by supervisor | *Generic scale* **1)** Respect for tissue **2)** Time and motion **3)** Knowledge and handling of instrument **4)** Flow of operation **5)** Use of assistants **6)** Knowledge of specific procedure |
| **Chen et al.**[1 8] 2010 | **VSSI** Vaginal Surgical Skills Index | *Observational study* Assessment by supervisor and blinded reviewer of video recording | *Generic and procedure-specific scale* **1)** Initial inspection **2)** Incision **3)** Maintenance of visibility **4)** Use of assistants **5)** Knowledge of instruments **6)** Tissue and instrument handling **7)** Electro surgery **8)** Knot tying **9)** Haemostasis **10)** Procedure completion **11)** Time and motion **12)** Flow of operation and forward planning **13)** Knowledge of specific procedure |

| Author and year | Assessment tool | Study design and assessment method | Domains assessed |
|---|---|---|---|
| **Chou B et al.**[19] 2008 | **HASC** Hopkins Assessment of Surgical Competency | *Observational study* Self-assessment and assessment by supervisor | *Generic and procedure-specific scale General surgical skills*: **1)** Knowledge/avoidance of potential complications, **2)** Respected tissue, **3)** Instrument Handling, **4)** Time and motion/moves not wasted, **5)** Bleeding controlled, **6)** Flow of operation *Specific surgical skills*: **1)** Knowledge of patient history/surgical indication, **2)** Knowledge of anatomy, **3)** Patient properly positioned on table/in stirrups, **4)** Proper placement of retractors, **5)** Proper assembly equipment, **6)** Proper positioning of lights |
| **Larsen CR et al.**[20] 2008 | **OSALS** Objective Structured Assessment of Laparoscopic Salpingectomy | *Prospective cohort study* Blinded video assessment by two observers | *Generic and procedure-specific scale OSALS general skills* **1)** Economy of movement, **2)** Confidence of movement, **3)** Economy of time, **4)** Errors; respect for tissue, **5)** Flow of operation/operative technique *OSALS specific skills:* **1)** Presentation of anatomy, **2)** Use of diathermy, **3)** Dissection of fallopian tube, **4)** Care for ovary, ovarian artery and pelvic wall, **5)** Extraction of fallopian tube |

| Author and year | Assessment tool | Study design and assessment method | Domains assessed |
|---|---|---|---|
| **Peter J. Frederick et al.[15]** 2016 | **RHAS** Robot Hysterectomy Assessment Score | *Observational study* Blinded video assessment by expert reviewers | *Procedure-specific scale* **1)** Handling of the round ligament, **2)** Developing the bladder flap, **3)** Isolating and securing the infundibulopelvic ligament (or utero-ovarian ligament if the ovaries were retained), **4)** Securing the uterine vessels, **5)** Performing the colpotomy and **6)** Closing the vaginal cuff |
| **Jeanne Goderstad et al.[16]** 2016 | **CAT-LSH** Competence Assessment for Laparoscopic Supracervical Hysterectomy | *Prospective interobserver study* Blinded video assessment by expert reviewers | *Procedure-specific scale* **1)** Ligament mobilisation, **2)** Release of adnexa form uterus, **3)** Division of uterine vessels, **4)** Uterus amputation |
| **Savren et al.[23]** 2019 | Feasible rating scale for formative and summative feedback | *Prospective cohort study* Blinded video assessment by two observers | *Procedure-specific scale* **1)** Division of fallopian tube and uteroovarian OR division of the infundibulopelvix ligament **2)** Dividing the round ligament **3)** Care for the ureter **4)** Opening the utero-vesicale peritoneum **5)** Identification and skeletonising **6)** Presentation and ligation of uterine arteries **7)** Opening of the vagina **8)** Suturing (catching the needle) **9)** Driving the needle through tissue, **10)** Placement and depth of sutures in the vaginal cuff, **11)** Suturing of the vagina and tying the knot |

| Author and year | Assessment tool | Study design and assessment method | Domains assessed |
|---|---|---|---|
| **Heinrich Husslein et al.** [24] 2015 | **GERT** Generic Error Rating Tool | *Prospective observational study* Blinded video assessment by expert reviewers | *Error assessment - generic and procedure-specific scale* **1)** Abdominal access and removal of instruments or trocars, **2)** Use of retractors, **3)** Use of energy, **4)** Grasping and dissection, **5)** Cutting, transection and stapling, **6)** Clipping, **7)** Suturing, **8)** Use of suction, **9)** Other *Each generic task subdivided into four distinct error modes:* **(1)** Too much use of force or distance, **2)** Too little use of force or distance, **3)** Inadequate visualisation, **4)** Wrong orientation of instrument *Procedure subdivided into:* **1)** Insertion of trocars, **2)** Creation of bladder flap, **3)** Colpotomy **4)** Vault closure |
| **Jeanne Goderstad et al.**[16] 2016 | **CAT-LSH** Competence Assessment for Laparoscopic Supracervical Hysterectomy | *Prospective interobserver study* Blinded video assessment by expert reviewers | *Procedure-specific scale* **1)** Ligament mobilisation, **2)** Release of adnexa form uterus, **3)** Division of uterine vessels, **4)** Uterus amputation |

| Author and year | Assessment tool | Study design and assessment method | Domains assessed |
|---|---|---|---|
| **Savren et al.[23]** 2019 | Feasible rating scale for formative and summative feedback | *Prospective cohort study* Blinded video assessment by two observers | *Procedure-specific scale* **1)** Division of fallopian tube and uteroovarian OR division of the infundibulopelvix ligament **2)** Dividing the round ligament **3)** Care for the ureter **4)** Opening the utero-vesicale peritoneum **5)** Identification and skeletonising **6)** Presentation and ligation of uterine arteries **7)** Opening of the vagina **8)** Suturing (catching the needle) **9)** Driving the needle through tissue, **10)** Placement and depth of sutures in the vaginal cuff, **11)** Suturing of the vagina and tying the knot |

| Author and year | Assessment tool | Study design and assessment method | Domains assessed |
|---|---|---|---|
| **Heinrich Husslein et al.** [24] 2015 | **GERT** Generic Error Rating Tool | *Prospective observational study* Blinded video assessment by expert reviewers | *Error assessment - generic and procedure-specific scale* **1)** Abdominal access and removal of instruments or trocars, **2)** Use of retractors, **3)** Use of energy, **4)** Grasping and dissection, **5)** Cutting, transection and stapling, **6)** Clipping, **7)** Suturing, **8)** Use of suction, **9)** Other *Each generic task subdivided into four distinct error modes:* **(1)** Too much use of force or distance, **2)** Too little use of force or distance, **3)** Inadequate visualisation, **4)** Wrong orientation of instrument *Procedure subdivided into:* **1)** Insertion of trocars, **2)** Creation of bladder flap, **3)** Colpotomy **4)** Vault closure |

**Table S2** Six consecutively linked stages of underlying methodological framework

| Scoping review methodology based on Arksey and O'Malley | Arksey and O'Malley's framework applied in scoping review |
|---|---|
| Review stage | |
| 1. Identify research question | Studies relevant to research question; broad search strategy c |
| 2. Identify relevant studies | Information specialist assisted in the design and execution of |
| 3. Study selection | Broad inclusion criteria but studies performed in simulation s |
| 4. Chart data | Overview noted author, assessment tool, study design, observ |
| 5. Collate, summarise and report results | Aim of a scoping review is to present an overview of all availa |
| 6. Consultation (optional) | Consulting with educational and gynaecological experts can p |

**Table S3**

Kane's validity argument

| | Meaning |
|---|---|
| **Scoring** | **Observed performance on score or rating scale** |

| | Meaning |
|---|---|
| **Generalisation** | **Reflection of performance in test setting** |
| **Extrapolation** | **Use of scores to reflect on real-world performance** |
| **Implication/decision** | **Application of scores to make a decision or take action** |

**Table 5**Overview of studies analysed using Kane's validity argument and of their strengths and weaknesses.

| **Assessment tool** | **Scoring** | **Generalisation** | **Extrapolation** | **Strength** | **Weakness** |
|---|---|---|---|---|---|
| Objective Structured Assessment of technical Skills (OSATS). [21] | Comparing OSATS scores over time | | Construct validity was demonstrated as a significant rise in score with increasing caseload as 1.10 OSATS point per assessed procedure (p=0.008, 95% CI 0.44–1.77) | Creating learning curves to identify residents in need of more guidance | No blinded assessment and self-evaluation; small sample size; high interrater variation; lack of objectivity; not adjusted for case mix |
| Vaginal Surgical Skills Index (VSSI). [1 8] | Comparing GRS with VSSI and adding a visual analogue scale for overall performance | Interrater reliability was 0.53 and intrarater reliability was 0.82 | Able to discriminate training levels for VSSI scores | 27 surgeons from two institutions; multiple expert reviewers; focus on case-mix | Assessment items not procedure specific and can be applied to laparoscopic surgery in general |
| Hopkins Assessment of Surgical Competency (HASC). [19] | Surgeons rated by supervisors on general surgical skills and case-specific surgical skills | Internal consistency reliability of the items using high Cronbach's alpha = 0.80 (p<0.001) | Discriminative validity for inexperienced vs intermediate surgeons (p<0.001) | 362 surgical cases were evaluated | No blinded assessment and self-evaluation; many different procedures evaluated; not adjusted for case mix |
| Objective Structured Assessment of Laparoscopic Salpingectomy (OSA-LS). [20] | Surgeons rated by OSA-LS | Interrater reliability =0.831 | Discriminative validity for inexperienced vs intermediate surgeon's vs experienced surgeons (p< 0.03) | Blinded | Small sample size; not adjusted for case mix |

| Assessment tool | Scoring | Generalisation | Extrapolation | Strength | Weakness |
|---|---|---|---|---|---|
| Robotic Hysterectomy Assessment Score (RHAS). [15] | Surgeons rated by expert viewers using RHAS | Interrater reliability for total domain score (p>0.006; p<0.001) | Differences demonstrated between experts, advanced beginners and novice in all domains except vaginal cuff closure | 52 blinded video recording; multiple expert reviewers | Confounding variable when assessing novice surgeons is the presence of an attending physician providing direct feedback; not adjusted for case mix |
| Competence Assessment for Laparoscopic Supracervical Hysterectomy (CAT-LSH). [16] | Comparing GOALS and CAT-LSH | Interrater reliability = 0.75 | Discriminative validity for inexperienced vs intermediate (p<0.006 and intermediate vs experts (p<0.001) | Video recording and blinded expert reviewers | Small sample size; not adjusted for case mix |
| Feasible rating scale for formative and summative feedback. [23] | Surgeons rated by expert viewers using 12-item procedure-specific checklist | Internal consistency reliability of the items Cronbach's alpha =0.95 (p<0.001) Interrater reliability =0.996 for one rater and 0.0998 for two raters | Discriminative validity for beginners and experienced surgeons (p=<0.001) | Video recording and blinded expert reviewers | Small sample size; not adjusted for case mix |
| GERT = Generic Error Rating Tool. [24] | OSATS scores used to establish and measure technical skills, to group surgeons as high or low performers and to correlate scores with GERT in an inverse relationship (more skilled surgeons make fewer errors) | Interrater reliability high (>0.95) Intrarater reliability significant (>0.95) | Significant negative correlation between OSATS and GERT scores | Video recording and blinded expert reviewers; analysis of operative substeps more prone to technical errors; captures near misses (events that may result in injury but did not, either by chance or timely intervention) | Although interrater reliability was high, not every error was rated identically by the two reviewers; not adjusted for case mix |

### Hosted file

`Figure S1.pptx` available at https://authorea.com/users/328486/articles/455736-tools-for-measuring-technical-skills-during-gynaecologic-surgery-a-scoping-review