Assessment of Technical Skills Competence in the Operating Room: A Systematic and Scoping Review

Christine Fahim, MSc, Natalie Wagner, BSc, Markuu T. Nousiainen, MSc, MEd, MD, FRCSC, and Ranil Sonnadara, PhD, MSc

Abstract

Purpose

While academic accreditation bodies continue to promote competency-based medical education (CBME), the feasibility of conducting regular CBME assessments remains challenging. The purpose of this study was to identify evidence pertaining to the practical application of assessments that aim to measure technical competence for surgical trainees in a nonsimulated, operative setting.

Method

In August 2016, the authors systematically searched Medline, Embase, and the Cochrane Database of Systematic Reviews for English-language, peer-reviewed articles published in or after 1996. The title, abstract, and full text of identified articles were screened. Data regarding study characteristics, psychometric and measurement properties, implementation of assessment, competency definitions, and faculty training were extracted. The findings from the systematic review were supplemented by a scoping review to identify key strategies related to faculty uptake and implementation of CBME assessments.

Results

A total of 32 studies were included. The majority of studies reported reasonable scores of interrater reliability and internal consistency. Seven articles identified minimum scores required to establish competence. Twenty-five articles mentioned faculty training. Many of the faculty training interventions focused on timely completion of assessments or scale calibration.

Conclusions

There are a number of diverse tools used to assess competence for intraoperative technical skills and a lack of consensus regarding the definition of technical competence within and across surgical specialties. Further work is required to identify when and how often trainees should be assessed and to identify strategies to train faculty to ensure timely and accurate assessment.

Competency-based medical education (CBME) frameworks have become increasingly prevalent among surgical education accreditation bodies, including the Royal College of Physicians and Surgeons of Canada, the General Medical Council in the United Kingdom, and the Accreditation Council for Graduate Medical Education in the United States. Traditional residency training programs use a time-based model, wherein trainees are assumed to have achieved competence at the time of program completion via clinical exposure during their training years.[1] In comparison, trainees in CBME programs must demonstrate competence in a variety of activities and contexts throughout their programs to successfully complete their training.[1,2]

Please see the end of this article for information about the authors.

Correspondence should be addressed to Ranil Sonnadara, McMaster University, ABB131B, 1280 Main St. W., Hamilton, ON, L8S4L8; telephone: (905) 525-9140, ext. 24156; e-mail: info@skillslab.ca.

Acad Med. XXXX;XX:00–00.
First published online
doi: 10.1097/ACM.0000000000001902

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A482.

While academic accreditation bodies continue to promote CBME, the feasibility of conducting regular assessments of competence remains challenging. The assessment of intraoperative technical skills for surgical trainees is especially challenging. Trainees often have limited opportunities to demonstrate technical competence because of operating room time pressures, safety concerns, and the reluctance of supervising staff to allow trainees to operate independently.[3,4] Furthermore, the implementation of CBME has been limited by a lack of consensus regarding what constitutes a satisfactory demonstration of competence.

In this review, we focus on the literature pertaining to the implementation of intraoperative assessment tools available to measure technical competence in surgical trainees. While previous reviews have identified multiple assessment tools related to technical skill, they have also highlighted the lack of data on feasibility, acceptability, educational impact, and generalizability to other contexts.[5,6] The purpose of this study was to use a systematic and scoping review methodology to identify evidence pertaining to the practical application of assessments that aim to measure technical competence for surgical trainees in a nonsimulated, operative setting. The first objective was to identify all assessment tools that evaluate technical skills for surgical trainees and determine how the tools were used to define competence. The second objective was to identify key strategies for faculty training to ensure effective implementation of CBME assessments.

Method

To meet our first objective, we performed a systematic review to search for literature pertaining to (1) assessment of technical skills among surgical residents and (2) how assessment tools were used to define competence. Three databases were used to conduct this search: Medline and Embase hosted by Ovid, and the Cochrane Database of Systematic Reviews hosted by Wiley. Searches were performed on August 17, 2016, and used the following search terms to conduct this study: *surgery*, *resident*, *trainee*, *evaluation* or *evaluation study*, *assessment* or *assessment tool*, *clinical competence*, or *skill* (the full search strategy is shown in Supplemental Digital Appendix 1 at

http://links.lww.com/ACADMED/A482). MeSH terms were exploded for relevant terms that met the inclusion criteria, and Boolean terms were used to combine search terms. The searches were limited to English-language, peer-reviewed articles (see below) published between 1996 and August 17, 2016.

The specific inclusion criterion for the study was literature pertaining to the technical skill assessment of surgical residents in an operating room setting (or shortly after a case was completed). Articles pertaining to assessment during simulation exercises were excluded because of the lack of evidence supporting performance transfer from simulation settings to real clinical practice.[7] Articles that focused on medical students, fellows, or staff; assessed nontechnical skills; focused on instrument development rather than implementation; assessed fewer than 10 trainees; and were commentaries, reviews, and/or conference abstracts were also excluded.

Two independent reviewers (C.F., N.W.) first completed a title and abstract, and then a full-text, screening of the articles resulting from the systematic review; they resolved any disagreements using a consensus process. Phi coefficient was used to calculate agreement for full-text inclusion of articles. A data collection form was used to extract information pertaining to the study characteristics (location, study design, surgical specialty, and surgical procedure or skill) and psychometric and measurement properties of the included articles. Data pertaining to implementation of assessment (i.e., type of assessment tool, sample size [or number of assessments], time to completion, etc.), competency definitions, and faculty training were also noted. Because of the breadth of the review, which spanned various surgical sites and assessment tools, findings could not be meaningfully summarized using meta-analysis. Rather, we grouped the studies by surgical specialty to allow for a meaningful overview of the available assessment tools in each surgical field.

In addition to the systematic review, we also completed an independent scoping review to help meet the second objective of identifying key training strategies related to faculty uptake and implementation of CBME assessments. A scoping review is a recent approach to map existing literature on a topic that has not been reviewed in depth or that is complex in nature.[8] We deemed it appropriate to use this method to supplement the traditional systematic review for our second objective because of the lack of direct evidence and the inconsistent reporting on faculty engagement for CBME uptake. The following search terms were used to conduct the scoping review: *competency-based curriculum*, *competency-based education*, *evaluation*, *clinical competence*, *surgery*, *tool* or *skill*, *faculty*, or *training*. One reviewer (C.F.) screened the titles, abstracts, and full text of the results of the scoping review. Other elements of the scoping review method (e.g., databases, data range, date of search) were the same as in the systematic review.

Results

Study selection

The systematic search revealed 1,056 studies from Medline, 898 from Embase, and 7 from Cochrane, for a total of 1,961 studies. Of these, 532 duplicates were excluded. The titles and abstracts of the resulting 1,429 articles were screened (Figure 1).

A total of 103 articles were selected for full-text screening. Of those, 72 were excluded because they did not meet the inclusion criteria (18 involved simulation, 14 did not assess technical skills, 16 did not meet the minimum sample size, 6 were commentaries or reviews, 4 involved tool development, and 14 were considered out of scope [e.g., trainee self-assessment of skills]). The remaining 31 articles were included in the review. A hand search of the reference lists of the included articles revealed an additional article meeting the inclusion criteria, for a total of 32 articles (Figure 1). Phi coefficient for interrater reliability was 0.91.

Information on study selection for the scoping review is given below in the "Faculty training" section.

Study characteristics

The 32 included studies spanned a time frame of 1999 through 2015.[9–40] The majority (17/32) were from the United States. A variety of skills and procedures were assessed from the following specialties: general surgery, otolaryngology–head and neck surgery (OTL-HNS), orthopedic, ophthalmology, obstetrics and/or gynecology, urology, and microsurgery, as well as multiple specialties (Appendix 1). A prospective study design was used in 30/32 studies (Supplemental Digital Appendix 2 at http://links.lww.com/ACADMED/A482).[9–38] Seven studies involved assessment of a video-recorded procedure in the operating room.[11,12,15,24,25,32,38] Of these seven, two studies[24,32] were randomized and one study[38] used a qualitative design.

Type of assessment tool

An overview of the assessment tools from the included studies, organized by surgical specialty, is given in Appendix 1. Articles pertaining to general surgery procedures were most prevalent (11/32), followed by articles pertaining to OTL-HNS procedures (7/32). Assessments comprised global rating scales (GRSs), task-specific checklists (TSCs), or hybrid assessments (see below). A GRS assesses overall performance of a general task, typically using a five-point Likert scale. Studies varied in the amount of field-specific detail they included in their GRS item anchors (e.g., Saleh et al[32] anchored their general skills to the field of ophthalmology, while Goh et al[20] anchored their GRS items to robotic surgery; this level of detail is greater than that of other studies which used general anchors to assess surgical skills). In comparison, a TSC assesses each step of the task, using either Likert scales or binary criteria. Twelve studies used a GRS in isolation,[13,14,17,20,21,24,28,30,33,35,37,40] while 2 studies used only a TSC.[16,26] The remaining studies used hybrid approaches, comprising two or more components, such as TSCs, GRSs, pass/fail or competent/not competent scores, error lists, or entrustment criteria.

Fifteen articles used the original or a modified version of the objective structured assessment of technical skills (OSATS) tool. The global operative assessment of laparoscopic skills (GOALS) tool was also used in a number of studies (Appendix 1). Two studies compared the validity and reliability of the OSATS and GOALS.[24,37] Kramp et al[24] found the tools to be highly correlated ($r = 0.879$, $P = .021$) and used this correlation to establish validity of the

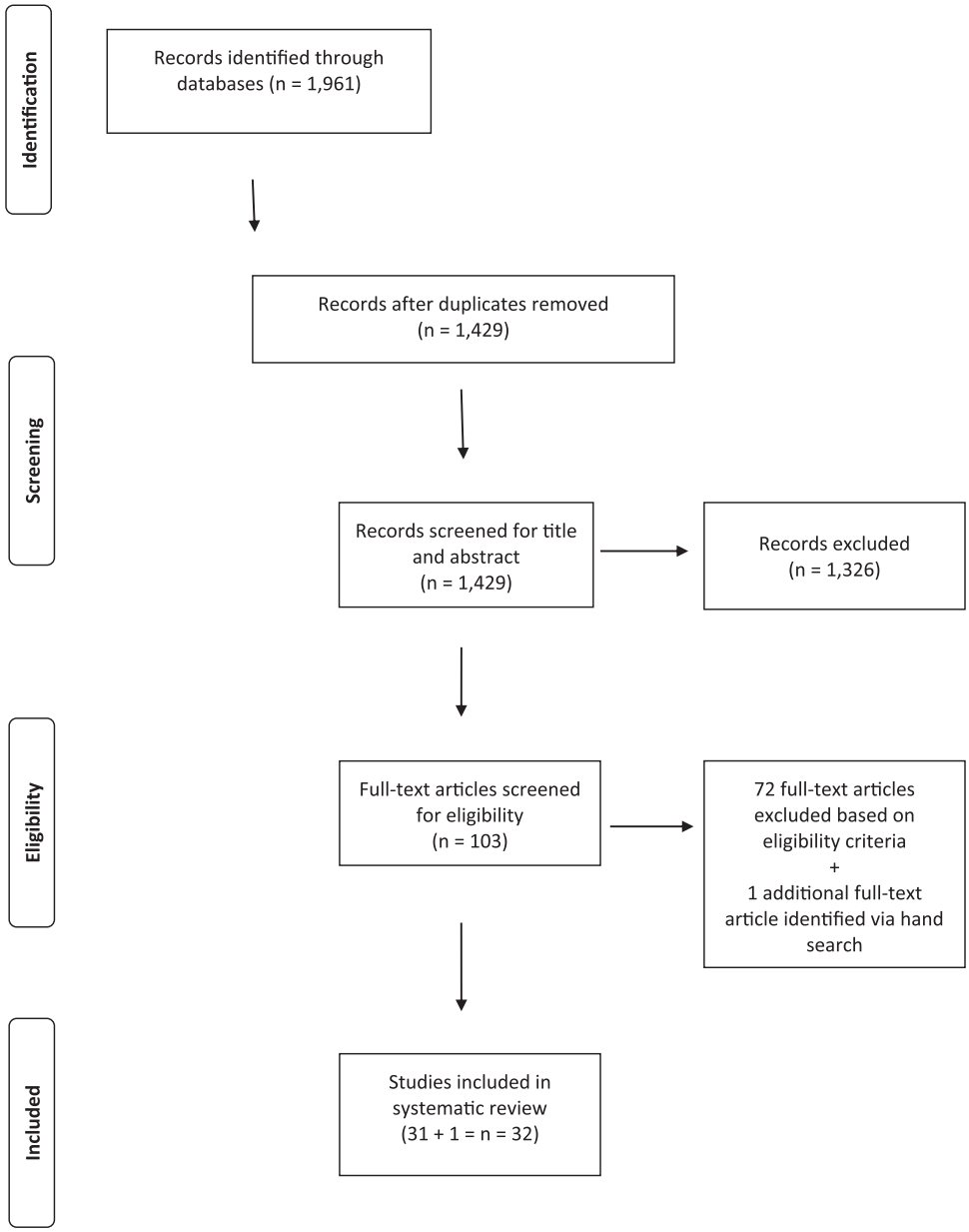

Figure 1 PRISMA flow diagram of the search and selection process in an August 2016 systematic review of literature to identify evidence pertaining to the practical application of assessments that aim to measure technical competence for surgical trainees in a nonsimulated, operative setting. Included articles were English language, peer reviewed, and published in or after 1996. Three databases were used: Medline, Embase, and the Cochrane Database of Systematic Reviews.

GOALS tool. In a direct comparison, Steigerwald et al[37] found a correlation of 0.975 ($P = .01$). Because of this near-perfect correlation, Kramp et al[24] and Steigerwald et al[37] questioned whether the GOALS tool adds value to the assessment of laparoscopic skills; instead, they suggested using the OSATS tool to maintain consistent nomenclature and standardization for surgical assessment.

Time to completion

Eight studies reported the average time required to complete the assessment form (see Supplemental Digital Appendix 3 at http://links.lww.com/ACADMED/A482).[9,12,14,18,27,29,34,35] Average completion times ranged from 2 to 18 minutes and were comparable between prospective studies. All studies with mean completion times of less than 5 minutes reported feasibility of the assessment tools and satisfaction by faculty assessors.[9,12,29] Glarner et al[18] found that, despite an average completion time of less than 2 minutes, assessors did not choose the same rating for all items, which would indicate response bias; rather, the assessments appeared to be thoughtfully and accurately completed. This finding demonstrates that short, simple assessments could feasibly be used regularly, without posing a significant barrier to staff. Studies that used video recordings to assess performance had conflicting reports regarding time to completion and ease of use.[12,25]

Competency definitions

Studies were further explored to determine whether the authors established a minimum competency score or criteria (Table 1). Twelve studies presented criteria for establishing competence. Seven of these studies outlined minimum numeric

Table 1

Criteria for Establishing Competence as Defined in 12 Studies Pertaining to Assessment of Technical Skills in the Operating Room[a]

Author[ref]	Specialty	Competence criteria
Marriott et al[27]	Multiple	Competence defined by achieving a score of Level 4 (out of 4 levels, where Level 4 = competent to perform the procedure unsupervised).
Kramp et al[24]	General surgery	Competence rated on a 10-point scale. Score required to be deemed competent is unclear.
Vassiliou et al[33]	General surgery	VAS (from "Was unable to complete the task with maximum guidance" to "Could perform the task safely and independently–fully competent") used to establish overall competence.
Hopmans et al[23]	General surgery	Competence defined using alphabetic summary scale (need to score D or "Competent to perform without supervision" to establish competence).
Laeeq et al[39]	OTL-HNS	Competence was defined by achieving a score of > 3 (out of 5) on every task on the scale (anchors differed for each item; refer to original tool).
Diaz Voss Varela et al[34]	OTL-HNS	Competence was defined by achieving 3 or higher (out of 5, where 3 = achieves competency) on the GRS. Faculty were also asked a binary question regarding competence.
Malik et al[26]	OTL-HNS	Competence established by scoring Level 3 or higher (out of 5 levels, Level 3 not defined).
Francis et al[16]	OTL-HNS	Competence established by scoring Level 3 or higher (out of 5 levels, where Level 3 = performs with minimal prompting).
Obeid et al[29]	OBGYN	Minimum pass level of 3 (out of 5) and full competence at 5 on GRS (where 3 = minimal acceptable pass and 5 = minimally acceptable level up to full competency).
Hodgins et al[22]	Orthopedic	Competence was defined by the performance limits on the LC-CUSUM graph. Competence was defined by a score of 40/50 on the GRS and 8/10 on the TSC. Adequate performance defined as a 10% failure rate and inadequate performance defined as a 30% failure rate. The minimum number of cases required to achieve competence were outlined.
Gofton et al[19]	Orthopedic	A binary global question was used to define competence to perform procedure independently. GRS scores (out of 5) used entrustment anchors for independence, where 5 = complete independence, understands risks and performs safely, practice ready.
Chan et al[11]	Microsurgery	Scores were anchored to reflect competence (score of 3 out of 5 needed, anchors differed for each item; refer to original tool) in terms of level of entrustment for independence.

Abbreviations: VAS indicates visual analog scale; OTL-HNS, otolaryngology–head and neck surgery; OBGYN, obstetrics and/or gynecology; GRS, global rating scale; LC-CUSUM, cumulative summation test for learning curve; TSC, task-specific checklist.

[a]From an August 2016 systematic review of literature to identify evidence pertaining to the practical application of assessments that aim to measure technical competence for surgical trainees in a nonsimulated, operative setting. Included articles were English language, peer reviewed, and published in or after 1996.

competence scores, which were typically established at a threshold of 3 or higher on a 5-point Likert scale.[11,16,26,27,29,34,39] One study by Hodgins et al[22] provided a statistical method to evaluate competence using the cumulative summation test for learning curve. The curve indicated that trainees should exhibit a score of 40/50 on the GRS and a score of 8/10 on the TSC to demonstrate competence.[22] Vassiliou et al[33] defined competence using a visual analog scale (from "Was unable to complete the task with maximum guidance" to "Could perform the task safely and independently—fully competent"). Hodgins et al[22] and Diaz Voss Varela et al[34] measured competence using a tool that had anchors that went beyond competence to excellence; for example, in Hodgins et al[22] a score of 1 (out of 5) was anchored as "poor" performance, 3 as "competent," and 5 as "clearly superior," while Diaz Voss Varela et al[34] anchored a score of 3 (out of 5) as "achieves competency" and a score of 4 as "achieves proficiency." Finally, Kramp et al[24] rated competence on a 10-point scale; however, it was unclear how a single score for competence was defined. While Chen et al,[12] Chou et al,[13] Goh et al,[20] and Larsen et al[25] did not establish minimum scores, they did highlight the need to do so as a next step in their work.

Entrustment criteria were used in two studies: Gofton et al[19] anchored the GRS in terms of how much the assessor trusted the trainee to perform independently (where 1 = I had to do and 5 = I did not need to be there), and Hopmans et al[23] used an alphabetic summary scale to assess competency and entrustment (i.e., from A = competent to *assist* adequately to E = competent to *supervise and educate* the operation).

Psychometric and measurement properties

All of the included studies assessed reliability, with the majority reporting reasonable scores of interrater reliability and internal consistency. Beard et al[10] and Fung Kee Fung et al[17] comprehensively evaluated reliability through the use of generalizability studies that evaluated the overall reliability of the tool, taking into consideration interrater reliability, internal consistency, and overall reliability.

Similarly, all included articles, with the exception of Sanfey et al's[38] qualitative study, demonstrated validation testing. Extreme groups comparison was the most common method used to establish construct validity. Some authors concluded that their assessment tools were psychometrically valid, despite very low reliability scores. For example, Qureshi and Ali[30] indicated that the unnamed tool in their study was found to be "reliable and valid for evaluating competence," despite reporting a very low interrater reliability score of 0.176. Similarly, Obeid et al[29] concluded that their tool was "valid, reliable, and

feasible," without having first established interrater reliability.

Psychometric properties could not be meaningfully compared because of significant variability in the study design, purpose, and evaluation metrics used in the included studies. The sample size for the number of evaluations assessed ranged from just 10 assessments to 1,635.[9–40]

Faculty training

Content that related to the importance or role of faculty training, identified barriers to faculty training or completion of assessment forms, or identified training procedures and outcomes was extracted. Of the included studies, 25/32 mentioned faculty training (Table 2). Many of the faculty training interventions focused on timely completion of assessments or scale calibration. Limitations to faculty completion of assessments included recall bias, halo bias, time pressures, lack of faculty buy-in, and assessor or evaluation fatigue. Fung Kee Fung et al[17] and Larsen et al[25] found that assessment scores were affected by a lack of faculty training, with Fung Kee Fung et al[17] reporting that faculty were more likely to globally assess the procedure and/or past performance of the resident than to uniquely score each individual item based on a single observed performance.

Despite these potential biases, there was some evidence to demonstrate the efficacy of faculty training. Vassiliou et al[33] reported intraclass correlation coefficients (ICCs) for trained versus untrained raters using the GOALS tool. Those who had been trained demonstrated higher reliability scores compared with those who were untrained (ICC = 0.76 and 0.39, respectively).[33]

Interventions for faculty training included education regarding the importance of timely and thorough completion of assessments. Ahmed et al[9] reported a 66% assessment completion rate using this strategy, while Obeid et al[29] and Laeeq et al[39] reported a 90% compliance rate. Some authors highlighted the importance of creating a "cultural shift" to reduce the prevalence of failure and right-shifting (the phenomenon of scoring trainees with a grade higher than what is believed to be deserved)[19,35]; however, these authors did not provide details regarding how to practically facilitate such education on timeliness or cultural shifts. Sanfey et al,[38] who used video recordings to assess trainees, stressed the importance of rater confidentiality and assured raters that their individual comments would remain confidential to promote honest assessments. Finally, a number of included studies mentioned faculty training in the context of the study trial (e.g., "do not intervene in the operating room unless patient safety is compromised" and "consider case difficulty and circumstances in completion of assessment forms").

The scoping review on faculty training revealed a total of 1,712 articles, of which 28 were selected for full-text screening. These 28 articles provided general descriptions of the benefits of CBME but were largely focused on how to establish the psychometric properties of individual scales, whereas the focus of this review was to identify pragmatic strategies to train faculty to effectively use existing competency curricula and assessment tools. The scoping review, therefore, did not reveal any additional articles pertaining to key strategies to train faculty to objectively assess technical skills using a competency framework.

Table 2

Barriers and Strategies Pertaining to Faculty Training as Mentioned in 25 Studies[a]

Author[ref]	Barriers to faculty training/buy-in	Training procedures and outcomes
Ahmed et al[9]	Recall bias (late completion of assessment results in recall bias) and Hawthorne effect (limits in having faculty observe residents).	Faculty training focused on timely completion. This resulted in a 66% completion rate within six days.
Beard et al[10]	Reliability was not affected by training; however, training is required for high-quality supervision and feedback.	Faculty were trained for study purposes. Authors asked faculty to only prompt or intervene if patient care became compromised. Authors stressed the importance of faculty training for intervention success.
Diaz Voss Varela et al[34]	Halo effect (phenomenon where the resident is judged based on overall performance) due to faculty bias.	Continuous professional training needed to combat halo bias. Experts were involved in instrument development. Assessors were asked to provide feedback regarding assessment process.
Doyle et al[14]	None listed.	Faculty was asked to compare trainees with a trained surgeon (training was provided to faculty to calibrate the scale).
Fung Kee Fung et al[17]	Raters account for significant variance in assessment scores. It appears that raters used the tool as a single GRS, demonstrating a halo effect.	Trained faculty to calibrate the scale, where 5 out of 5 demonstrates readiness for independent practice. Authors recommended rewriting or reanchoring items for clarity and training all faculty. However, it appears that faculty training may be futile if raters treat scale as a single GRS.
Glarner et al[18]	Bias introduced by assessor knowledge of resident.	None listed.
Gofton et al[19]	None listed.	There were faculty training sessions to create a cultural shift and to use the full scale.
Goh et al[20]	None listed.	Trained observers were used for the study. All observers were residents who were given grading instructions. Observers were asked to consider case difficulty when rating.
Gumbs et al[21]	Lack of interrater and intrarater reliability.	None listed.

(*Table continues*)

Table 2
(Continued)

Author[ref]	Barriers to faculty training/buy-in	Training procedures and outcomes
Hodgins et al[22]	Bias introduced by assessor knowledge of resident.	None listed.
Hopmans et al[23]	Bias introduced by assessor knowledge of resident. Authors noted having difficulty finding blinded raters.	Residents and staff were briefed on study purpose and trained regarding use of the scale. Faculty were instructed to allow residents to lead surgery. Completion of assessments was required immediately post operation, and instruction was given that assessment was not to be affected by previous experiences with the resident.
Kramp et al[24]	Scores may be affected by assessor fatigue and/or time pressure.	Teach-the-teacher trainings previously held at the institution.
Laeeq et al[39]	Educators trained as assessors may be limited in their ability to effectively assess surgical skills. Lack of faculty buy-in due to time was a limitation to implementation.	Faculty were instructed to complete assessment immediately post operation, with resident present, to ensure high-quality formative feedback and avoid recall bias. Involving faculty early on in assessment process resulted in a 90% compliance rate.
Larsen et al[25]	The usability of some scale items was limited by a lack of faculty training.	None listed.
Malik et al[26]	Plateau in scores for initial milestones may be due to faculty bias (i.e., faculty believe competence has been achieved in initial milestones and are less likely to assess these milestones in future procedures).	None listed.
Marriott et al[27]	Compliance by raters was difficult to attain because of organizational issues and time limitations.	"Achieving PBA reliability may not require rigorous training of clinical supervisors.… The form is intuitive."[27] Training was required to understand purpose and process of assessment. Assessors and residents were trained on the use of PBAs. Assessors were trained to allow the resident to lead and only intervene if patient care were to become compromised, and were instructed that assessment was not to be affected based on previous experiences with the resident.
Niitsu et al[28]	None listed.	All assessors were asked to watch a number of video-recorded procedures, and scores were calibrated to ensure objectivity of assessment.
Obeid et al[29]	Faculty bias (trainees evaluated by single nonblinded faculty member). Tool demonstrated interitem reliability, but interrater reliability not established; this places limit on validity scores.	Faculty were trained to complete assessment directly after operation. There were a total of 175/195 (90%) complete (no missing data) assessments.
Qureshi and Ali[30]	None listed.	Importance of training was recognized. Assessors were asked not to intervene during operation and to only provide feedback following the procedure and submission of the assessment form.
Sanfey et al[38]	Performance errors were not always noted by a single assessor and were only witnessed once the video was discussed at an assessment meeting.	Raters were assured that comments would remain confidential, allowing them to express critical opinions. Authors recommended that raters be given clear frameworks for operative procedures to improve consistency and accuracy of technical assessment, and requiring multiple assessors.
Steigerwald et al[37]	Availability of trained assessors in the operating room was limited because of logistical constraints.	Assessors were trained in the use of the assessment tools.
Vassiliou et al[33]	None listed.	Observers were trained to use the full range of scores on the assessment tool. Observers were asked to consider case difficulty and special circumstances that may have affected resident performance.
Wagner et al[35]	None listed.	Opinion leader (program director) was used to encourage timely completion of assessments and provision of feedback to residents. Hospital culture facilitated participation.
Williams et al[36]	None listed.	Authors recommended that faculty be encouraged to complete assessments immediately following procedure, as assessments completed > 3 days post operation lack clarity and detail.
Wohaibi et al[40]	No rigorous training program for raters available at time of retrospective review.	A structured training program that will be responsive to rater behaviors is under development as a next step in the authors' research.

Abbreviations: GRS indicates global rating scale; PBA, procedure-based assessment.
[a]From an August 2016 systematic and scoping review of literature to identify evidence pertaining to the practical application of assessments that aim to measure technical competence for surgical trainees in a nonsimulated, operative setting. Included articles were English language, peer reviewed, and published in or after 1996.

Discussion

Use of assessment tools to establish competence

Defining competence. On the basis of this review, it appears that a clearer definition for competence needs to be established. It is unclear whether study authors defined competence as the minimum skill required to safely and independently practice or as a complete mastery of the procedure. For example, the anchoring system used by Hodgins et al,[22] which had a maximum rating of clearly superior, leads readers to believe there is a level of surgical excellence that surpasses competence.[41] Of 32 included studies related to technical skills assessment, only 12 presented criteria for establishing competence. About half of these studies chose to define specific numerical cutoff points to establish competence in performing a specific surgical procedure (e.g., a trainee must demonstrate a score of 3 on a 5-point Likert scale), while others, albeit just a few, used entrustment criteria (i.e., assessors trusted the trainee to be able to complete the procedure independently). The latter approach has garnered traction among researchers such as Gofton et al,[19] who hypothesize that entrustment scores will reduce the rate of right-shifted numerical scores by clearly delineating the meaning of competence. While the concept of entrustment is promising, it is still important to recognize that the success of entrustment scales hinges on faculty willingness and ability to allow trainees to complete portions of procedures independently, and to complete regular, objective assessments despite existing time and resource limitations.

Number of assessments required to establish competence. Although this review did look at data on the number of assessments needed to determine competence, to the best of our knowledge, guidelines recommending the minimum number of assessments or raters needed to establish reliability in CBME assessments do not exist. However, the measurement literature suggests that reliability is increased by multiple observations from a variety of contexts.[36,42,43] As such, the establishment of guidelines regarding *when* and *how often* to evaluate trainees may be useful for future studies to consider.[44]

One potential solution to determine how many assessments are necessary to determine competence may come from learning curve data from individual technical procedures. Once established, these data points may be beneficial by identifying critical periods during which faculty should assess trainees. For example, Malik et al[26] reported that program directors believed trainees demonstrated competency in a mastoidectomy procedure after 8 to 10 procedures, which was typically achieved in postgraduate year 4. Similarly, Ahmed et al[9] showed that second-year trainees experienced the steepest learning curve in obtaining technical skills for tonsillectomies and suggested that assessments for this procedure should correspond with this time in a trainee's career. Establishment of learning curves will highlight these trends and may allow assessors to focus their energies on critical time points in residents' training. Such focus may create a more efficient system of assessment that would benefit both assessors and trainees.

We found no consistency regarding how often to evaluate residents. However, the use of generalizability studies could prove useful. As one example, Fung Kee Fung et al[17] use a generalizability study to reveal that a minimum of 12 ratings are required to demonstrate a variance score of 0.80 and obtain reliable scores on general laparoscopic skills competence.

Need for psychometrically sound scales. This review revealed a number of diverse tools used to assess competence for intraoperative technical skills, with some authors concluding that their assessment tools were psychometrically valid despite very low reliability scores. The ability to establish technical competence using assessment tools hinges on the reliability and validity of the instruments. Because reliability places an upper limit on validity, it is difficult to conclude that the presented tools hold construct validity for the assessed groups.[43] Furthermore, as assessment instruments are often designed to assess a specific task, validity in one context does not necessarily ensure validity in another. As such, measurement experts warn against the overgeneralizability of assessment tools and caution educators to look beyond whether an assessment tool is valid and consider the appropriateness of the scale for the construct being measured.[42]

Faculty training

Beard,[45(p282)] lead researcher in the field of surgical education and competency, states that "the key to reliable assessment and constructive feedback is well-trained trainers." While more than half of the studies in our review highlight the importance of faculty training, very few provide any details on how to practically facilitate faculty education. The most common strategy found in the literature is to recommend that assessments be completed within two to three days of the observed procedure.[35,46] In support of this recommendation, Williams et al[36] found that the amount of feedback provided on assessments was greatest when the assessment was completed within two to three days of the procedure, and feedback was found to be most useful when provided within one to three days following surgery.

One barrier to timely completion is limited time resources. If assessors are overburdened with the number of assessments they are required to complete, it can lead to what is known as "evaluation fatigue."[47,48] This fatigue results in poorly completed or missing assessment data. We postulate that unless evaluation fatigue is mitigated, faculty training strategies to objectively and accurately complete multiple trainee assessments will be largely unsuccessful.

Doyle et al[14] and Laeeq et al[39] proposed an alternative strategy of assessment, which placed the onus of assessment on trainees rather than faculty. Such strategies may reduce evaluation fatigue while promoting a culture of timely completion. Doyle et al[14] also suggest that trainees ensure that all assessments were completed immediately following surgery, a strategy that resulted in a compliance rate of 90% among faculty in Obeid et al[29] and Laeeq et al.[39] As the number of assessments continues to rise in CBME, further work is required to identify effective strategies to train faculty to ensure timely and accurate assessment.

Limitations

This review has several limitations. First, our search strategy was limited to peer-reviewed articles published in the English language. As such, we may have overlooked relevant data published in non-English articles or the gray literature. Additionally, we elected to conduct a

scoping review on faculty uptake and implementation of CBME assessments. While we believe the scoping review was sufficient to meet our second objective, another full systematic review may have resulted in additional data pertaining to key training strategies for faculty.

Conclusion

A reduction of trainee duty hours, increasing concerns for patient safety, and lack of operating time, coupled with the need to train residents, has pushed many accreditation bodies to implement CBME curricula for surgical trainees. Despite this, practical guidelines regarding CBME implementation and assessment are lacking. Surgical programs are often provided with general competency targets for their specialty, yet there remains significant inconsistency in the processes and methods by which competence is assessed. This review revealed a number of diverse tools used to assess competence for intraoperative technical skills. A large number of the included studies built on the OSATS or GOALS assessment tools. Despite this, psychometric limitations and gaps in faculty training continue to pose threats to the establishment of a reliable and valid way to assess technical competence. Use of the OSATS, or a modified version of the OSATS, may help to standardize assessments across specialties and provide consistency in the manner by which surgical trainees are assessed. TSCs could be used in conjunction with GRS items (such as the OSATS) to provide meaningful, formative feedback to residents. However, further research is required to demonstrate whether the use of TSCs will truly benefit residents' learning and progression to competence. Additional research is also required to determine procedure-specific learning curves, cutoff scores for defining competence, and the number of assessments required to establish competence. Assessments requiring less than five minutes to complete are considered acceptable by faculty assessors, suggesting that efficient assessment of technical skills in an operative setting is feasible. Future researchers should aim to address the gaps identified in this review and develop a comprehensive faculty training strategy that will ensure the use of valid and reliable tools for the consistent assessment of competence among residents in surgical fields.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: Reported as not applicable.

C. Fahim is a PhD candidate, Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada.

N. Wagner is a PhD candidate, Department of Psychology, Neuroscience and Behaviour, McMaster University, Hamilton, Ontario, Canada.

M.T. Nousiainen is orthopedic surgeon and assistant professor, Sunnybrook Hospital, Department of Surgery, University of Toronto, Toronto, Ontario, Canada.

R. Sonnadara is director of education science and associate professor, Department of Surgery, McMaster University, Hamilton, Ontario, Canada, and associate professor, Department of Surgery, University of Toronto, Toronto, Ontario, Canada; ORCID: http://orcid.org/0000-0001-8318-5714.

References

1 Hodges B. Assessment in the post-psychometric era: Learning to love the subjective and collective. Med Teach. 2013;35:564–568.
2 Tsue TT, Dugan JW, Burkey B. Assessment of surgical competency. Otolaryngol Clin North Am. 2007;40:1237–1259, vii.
3 Reznick RK, MacRae H. Teaching surgical skills—Changes in the wind. N Engl J Med. 2006;355:2664–2669.
4 Sonnadara RR, Mui C, McQueen S, et al. Reflections on competency-based education and training for surgical residents. J Surg Educ. 2007;71:151–158.
5 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.e6.
6 Torsney KM, Cocker DM, Slesser AA. The modern surgeon and competency assessment: Are the workplace-based assessments evidence-based? World J Surg. 2015;39:623–633.
7 Buckley CE, Kavanagh DO, Traynor O, Neary PC. Is the skillset obtained in surgical simulation transferable to the operating theatre? Am J Surg. 2014;207:146–157.
8 Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA. A scoping review of scoping reviews: Advancing the approach and enhancing the consistency. Res Synth Methods. 2014;5:371–385.
9 Ahmed A, Ishman SL, Laeeq K, Bhatti NI. Assessment of improvement of trainee surgical skills in the operating room for tonsillectomy. Laryngoscope. 2013;123:1639–1644.
10 Beard JD, Marriott J, Purdie H, Crossley J. Assessing the surgical skills of trainees in the operating theatre: A prospective observational study of the methodology. Health Technol Assess. 2011;15:i–xxi, 1.
11 Chan W, Niranjan N, Ramakrishnan V. Structured assessment of microsurgery skills in the clinical setting. J Plast Reconstr Aesthet Surg. 2010;63:1329–1334.
12 Chen CC, Korn A, Klingele C, et al. Objective assessment of vaginal surgical skills. Am J Obstet Gynecol. 2010;203:79.e1–79.e8.
13 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.
14 Doyle JD, Webber EM, Sidhu RS. A universal global rating scale for the evaluation of technical skills in the operating room. Am J Surg. 2007;193:551–555.
15 Eubanks TR, Clements RH, Pohl D, et al. An objective scoring system for laparoscopic cholecystectomy. J Am Coll Surg. 1999;189:566–574.
16 Francis HW, Masood H, Laeeq K, Bhatti NI. Defining milestones toward competency in mastoidectomy using a skills assessment paradigm. Laryngoscope. 2010;120:1417–1421.
17 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.
18 Glarner CE, McDonald RJ, Smith AB, et al. Utilizing a novel tool for the comprehensive assessment of resident operative performance. J Surg Educ. 2013;70:813–820.
19 Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–1407.
20 Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills. J Urol. 2012;187:247–252.
21 Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. J Am Coll Surg. 2007;204:308–313.
22 Hodgins JL, Veillette C, Biau D, Sonnadara R. The knee arthroscopy learning curve: Quantitative assessment of surgical skills. Arthroscopy. 2014;30:613–621.
23 Hopmans CJ, den Hoed PT, van der Laan L, et al. Assessment of surgery residents' operative skills in the operating theater using a modified Objective Structured Assessment of Technical Skills (OSATS): A prospective multicenter study. Surgery. 2014;156:1078–1088.
24 Kramp KH, van Det MJ, Hoff C, Lamme B, Veeger NJ, Pierie JP. Validity and reliability of global operative assessment of laparoscopic skills (GOALS) in novice trainees performing a laparoscopic cholecystectomy. J Surg Educ. 2015;72:351–358.
25 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.
26 Malik MU, Varela DA, Park E, et al. Determinants of resident competence in mastoidectomy: Role of interest and deliberate practice. Laryngoscope. 2013;123:3162–3167.
27 Marriott J, Purdie H, Crossley J, Beard JD. Evaluation of procedure-based assessment for assessing trainees' skills in the operating theatre. Br J Surg. 2011;98:450–457.
28 Niitsu H, Hirabayashi N, Yoshimitsu M, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. Surg Today. 2013;43:271–275.

29 Obeid AA, Al-Qahtani KH, Ashraf M, Alghamdi FR, Marglani O, Alherabi A. Development and testing for an operative competency assessment tool for nasal septoplasty surgery. Am J Rhinol Allergy. 2014;28:e163–e167.
30 Qureshi RN, Ali SK. Assessment of competence for caesarean section with global rating scale. J Pak Med Assoc. 2013;63:1003–1007.
31 Roberson DW, Kentala E, Forbes P. Development and validation of an objective instrument to measure surgical performance at tonsillectomy. Laryngoscope. 2005;115:2127–2137.
32 Saleh GM, Gauba V, Mitra A, Litwin AS, Chung AK, Benjamin L. Objective structured assessment of cataract surgical skill. Arch Ophthalmol. 2007;125:363–366.
33 Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg. 2005;190:107–113.
34 Diaz Voss Varela DA, Malik MU, Thompson CB, Cummings CW, Bhatti NI, Tufano RP. Comprehensive assessment of thyroidectomy skills development: A pilot project. Laryngoscope. 2012;122:103–109.
35 Wagner JP, Chen DC, Donahue TR, et al. Assessment of resident operative performance using a real-time mobile Web system: Preparing for the milestone age. J Surg Educ. 2014;71:e41–e46.
36 Williams RG, Chen XP, Sanfey H, Markwell SJ, Mellinger JD, Dunnington GL. The measured effect of delay in completing operative performance ratings on clarity and detail of ratings assigned. J Surg Educ. 2014;71:e132–e138.
37 Steigerwald SN, Park J, Hardy KM, Gillman L, Vergis AS. Establishing the concurrent validity of general and technique-specific skills assessments in surgical education. Am J Surg. 2016;211:268–273.
38 Sanfey H, Williams RG, Chen X, Dunnington GL. Evaluating resident operative performance: A qualitative analysis of expert opinions. Surgery. 2011;150:759–770.
39 Laeeq K, Waseem R, Weatherly RA, et al. In-training assessment and predictors of competency in endoscopic sinus surgery. Laryngoscope. 2010;120:2540–2545.
40 Wohaibi EM, Earle DB, Ansanitis FE, Wait RB, Fernandez G, Seymour NE. A new Web-based operative skills assessment tool effectively tracks progression in surgical resident performance. J Surg Educ. 2007;64:333–341.
41 Ten Cate O, Billett S. Competency-based medical education: Origins, perspectives and potentialities. Med Educ. 2014;48:325–332.
42 Schoenherr JR, Hamstra SJ. Psychometrics and its discontents: An historical perspective on the discourse of the measurement tradition. Adv Health Sci Educ Theory Pract. 2016;21:719–729.
43 Streiner DL, Norman GR, Cairney J. Health Measurement Scales: A Practical Guide to Their Development and Use. 5th ed. Oxford, UK: Oxford University Press; 2015.
44 Moonen-van Loon JM, Overeem K, Donkers HH, van der Vleuten CP, Driessen EW. Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. Adv Health Sci Educ Theory Pract. 2013;18:1087–1102.
45 Beard JD. Assessment of surgical skills of trainees in the UK. Ann R Coll Surg Engl. 2008;90:282–285.
46 Kim MJ, Williams RG, Boehler ML, Ketchum JK, Dunnington GL. Refining the evaluation of operating room performance. J Surg Educ. 2009;66:352–356.
47 Fahim C, Bhandari M, Yang I, Sonnadara R. Development and early piloting of a CanMEDS competency-based feedback tool for surgical grand rounds. J Surg Educ. 2016;73:409–415.
48 McQueen S. Assessment and Feedback in Surgical Training [MSc dissertation]. Hamilton, Ontario, Canada: McMaster University; 2015.

References cited in Appendix 1 only

49 Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. Am J Surg. 1997;173:226–230.
50 Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997;84:273–278.

Appendix 1

Assessment Properties of Included Studies (n = 32) in an August 2016 Systematic Review of Literature to Identify Evidence Pertaining to the Practical Application of Assessments That Aim to Measure Technical Competence for Surgical Trainees in a Nonsimulated, Operative Setting[a]

Author[ref]	Assessment tool	Procedure or skill	Assessment type (number of assessments)	Areas of assessment (number of assessments)	Comments
General surgery					
Glarner et al[18]	Unnamed	Laparoscopic segmental colon	TSC (8) GRS (7)	TSC: Difficulty of case (1) OR setup and positioning (1) Procedure-specific items (6) GRS: OSATS[b] (7)	Also evaluated nontechnical performance (i.e., situational awareness, decision making, communication and teamwork, leadership)
Doyle et al[14]	Global Rating Index for Technical Skills	Multiple	GRS (9)	Respect for tissue (1) Time and motion (1) Instrument handling (1) Flow of operation (1) Knowledge of specific procedure (1) Use of assistants (1) Communication skills (1) Depth perception (1) Bimanual dexterity (1)	Based on the OSATS and GOALS
Wohaibi et al[40]	OpRate	Multiple	GRS (13)	Preoperation evaluation (4) OR skills (7) Summative overall performance (2)	Completed electronically
Wagner et al[35]	Southern Illinois University Operative Performance Rating Scale	Multiple	GRS (7)	Case difficulty (1) Degree of prompting (1) Instrument handling (1) Respect for tissue (1) Time and motion (1) Operation flow (1) Overall performance (1)	Based on the OSATS
Gumbs et al[21]	GOALS	Laparoscopic cholecystectomy and appendectomy	GRS (5)	GOALS[c] (5)	
Kramp et al[24]	GOALS	Laparoscopic cholecystectomy	GRS (6)	GOALS[c] (5) Overall case difficulty (1)	Compared with the OSATS to determine reliability and validity of the GOALS for laparoscopic cholecystectomy
Vassiliou et al[33]	GOALS	Laparoscopic cholecystectomy	GRS (5)	GOALS[c] (5)	• Compared with a TSC for dissection of gallbladder and two VASs for overall competence and case difficulty to evaluate construct validity • The GOALS found to be superior to a TSC or VAS

(*Appendix continues*)

Appendix 1
(Continued)

Author[ref]	Assessment tool	Procedure or skill	Assessment type (number of assessments)	Areas of assessment (number of assessments)	Comments
Eubanks et al[15]	Unnamed	Laparoscopic cholecystectomy	Point sheet (23 items, 100 points) Error points, graded based on severity of error (21 items)	Point sheet (points given based on procedure-specific items): Initial exposure (4 items, 10 points) Initial dissection (3 items, 15 points) Cystic duct dissection (4 items, 17 points) Cystic duct cannulation (4 items, 20 points) Cystic artery dissection (6 items, 24 points) Gallbladder fossa dissection (2 items, 14 points) Error sheet (points deducted if errors occur): Gallbladder (3 items) Liver (4 items) Cystic duct (7 items) Cystic artery (5 items) Miscellaneous (2 items)	
Sanfey et al[38]	Not applicable (qualitative study)	Multiple			
Hopmans et al[23]	Modified OSATS	Multiple	GRS (8) Overall performance scale (2) Alphabetic summary scale (1)	GRS: Indication for surgery (1) Respect for tissue (1) Time and motion (1) Knowledge and handling of instruments (1) Use of assistants (1) Flow of operation (1) Knowledge of specific procedure (1) Perioperative management (1) Overall performance scale (2) Alphabetic summary scale (1)	Adapted the OSATS (combined one domain from original OSATS, added two non-OSATS domains to the GRS)
Williams et al[36]	Unnamed	Multiple	TSC (3–5) Overall performance item (1)	TSC: Varied by procedure[d] Overall performance item (1)	• Forms available from authors • Part of the tool based on the OSATS
Otolaryngology–head and neck surgery					
Ahmed et al[9]	Unnamed	Tonsillectomy	TSC (11) GRS (11) P/F/C (2)	TSC: Procedure-specific items (11) GRS: Understanding indications for surgery (1) Communication with anesthesiologist (1) Instrument handling (1) Respect for tissue (1) Time and motion (1) Amount of tension (1) Direction of tension (1) Cautery technique (1) Knowledge of specific procedure (1) Flow of operation (1) Overall surgical performance (1) P/F/C: Feedback and competence (2)	GRS includes some overlap with the OSATS

(*Appendix continues*)

Appendix 1
(Continued)

Author[ref]	Assessment tool	Procedure or skill	Assessment type (number of assessments)	Areas of assessment (number of assessments)	Comments
Roberson et al[31]	Unnamed	Tonsillectomy	GRS (12) TSC[d]	GRS: Respect for tissue (1) Appropriate tension (1) Instrument handling (1) Flow of operation (1) Pace of operation (1) Unnecessary moves (1) Use of assistant (1) Cautery technique (1) Knowledge of procedure (1) Response to stress/complications (1) Teamwork and leadership (1) Patient care (1) TSC[d]: "Specific steps" items Cautery items Patient care items	• Modeled after Reznick et al[49] • Some of the items relate to nontechnical skills that may affect technical performance (i.e., response to stress, teamwork and leadership) • The TSC portion is available from authors ("too complex for routine use for resident evaluation")
Laeeq et al[39]	Endoscopic Sinus Surgery Assessment Tool	Endoscopic sinus surgery	GRS (10) TSC (21)	GRS: Understanding objectives of surgery (1) Use of radiographs or image guidance (1) Use of endoscopes (1) Knowledge of instruments (1) Instrument handling (1) Respect for tissue (1) Time and motion (1) Knowledge of specific procedure (1) Flow of operation (1) Overall surgical performance (1) TSC: Procedure-specific items (21)	Based on the OSATS
Diaz Voss Varela et al[34]	OSATS	Thyroidectomy	TSC (10) GRS (9) P/F/C (2)	TSC: Procedure-specific items (10) GRS: Understanding objectives of surgery (1) Use of intraoperative nerve monitor and placement of endotracheal tube (1) Knowledge of instruments (1) Instrument handling (1) Respect of tissue (1) Time and motion (1) Knowledge of specific procedure (1) Flow of operation (1) Overall surgical performance (1) P/F/C: Feedback and competence (2)	GRS based on the OSATS
Malik et al[26]	Mastoidectomy task-based checklist	Cortical mastoidectomy	TSC (5)	Procedure-specific items (5)	
Francis et al[16]	Unnamed	Mastoidectomy	TSC (17)	Procedure-specific items (17)	

(*Appendix continues*)

Appendix 1
(Continued)

Author[ref]	Assessment tool	Procedure or skill	Assessment type (number of assessments)	Areas of assessment (number of assessments)	Comments
Obeid et al[29]	Septoplasty TSC	Septoplasty	GRS (7) TSC (8) P/F/C (4)	GRS: Understanding of indications and objectives of surgery (1) Respect for tissue (1) Time and motion (1) Instrument handling (1) Knowledge of instruments (1) Flow of operation and forward planning (1) Knowledge of specific procedure (1) TSC: Procedure-specific items (8) P/F/C: Feedback and competence (4)	Based on the OSATS
Obstetrics and/or gynecology					
Larsen et al[25]	Objective Structured Assessment of Laparoscopic Salpingectomy	Laparoscopic salpingectomy	TSC (5) GRS (5)	TSC: Procedure-specific items (5) GRS: Economy of movements (1) Confidence of movements (1) Economy of time (1) Errors and respect for tissue (1) Flow of operation or operative technique (1)	Based on the OSATS
Qureshi and Ali[30]	Unnamed	Septoplasty	GRS (10) P/F/C (1)	Not provided	
Chen et al[12]	Vaginal Surgical Skills Index	Vaginal hysterectomy	GRS (13)	Initial inspection (1) Incision (1) Maintenance of visibility (1) Use of assistants (1) Knowledge of instruments (1) Tissue and instrument handling (1) Electrosurgery (1) Knot tying (1) Hemostasis (1) Procedure completion (1) Time and motion (1) Flow of operation and forward planning (1) Knowledge of specific procedure (1)	
Fung Kee Fung et al[17]	Interactive Voice Response Assessment System	Laparoscopy skills	GRS (3)	Knowledge and handling of instruments (1) Ability to plan and perform operative moves (1) Knowledge of anatomy and interpretation of operative findings (1)	

(*Appendix continues*)

Appendix 1
(Continued)

Author[ref]	Assessment tool	Procedure or skill	Assessment type (number of assessments)	Areas of assessment (number of assessments)	Comments
Chou et al[13]	Hopkins Assessment of Surgical Competency	General surgical skills	GRS, general surgical skills (6) GRS, case-specific skills (6)	GRS, general surgical skills: Knowledge of patient history or surgical indication (1) Respected tissue (1) Instrument handling (1) Time and motion or moves not wasted (1) Bleeding controlled (1) Flow of operation (1) GRS, case-specific skills: Knowledge of patient history or surgical indication (1) Knowledge of anatomy (1) Patient properly positioned on table or in stirrups (1) Proper placement of retractors (1) Proper assembly of equipment (1) Proper positioning of lights (1)	
Orthopedic					
Hodgins et al[22]	Basic Arthroscopic Knee Skill Scoring System	Knee arthroscopy	TSC (10) GRS (10)	TSC: Procedure-specific items (10) GRS: Dissection (1) Instrument handling (1) Depth perception (1) Bimanual dexterity (1) Flow of operation and forward planning (1) Knowledge of instruments (1) Efficiency (1) Knowledge of specific procedure (1) Autonomy (1) Quality of final product (1)	
Gofton et al[19]	Ottawa Surgical Competency Operating Room Evaluation	Multiple	GRS (8) P/F/C (3)	GRS: Preprocedure plan (1) Case preparation (1) Knowledge of specific procedural steps (1) Technical performance (1) Visuospatial skills (1) Postprocedure plan (1) Efficiency and flow (1) Communication (1) P/F/C: Competence (1) Feedback (2)	
Urology					
Goh et al[20]	Global Evaluative Assessment of Robotic Skills	Robotic prostatectomy	GRS (6)	Depth perception (1) Bimanual dexterity (1) Efficiency (1) Force sensitivity (1) Autonomy (1) Robotic control (1)	Based on the GOALS

(*Appendix continues*)

Appendix 1
(Continued)

Author[ref]	Assessment tool	Procedure or skill	Assessment type (number of assessments)	Areas of assessment (number of assessments)	Comments
Ophthalmology					
Saleh et al[32]	Objective Structured Assessment of Cataract Surgical Skill	Cataract surgery	TSC (14) GRS (6)	TSC: Procedure-specific items (14) GRS: Wound neutrality and minimizing eye rolling and corneal distortion (1) Eye positioned centrally within microscope view (1) Conjunctival and corneal tissue handling (1) Capsule protection (1) Iris protection (1) Overall speed and fluidity of procedure (1)	
Microsurgery					
Chan et al[11]	Structured Assessment of Microsurgery Skills	Microvascular anastomoses	GRS (12) Errors list (25) P/F/C (2)	GRS: Dexterity (3) Visuospatial ability (3) Operative flow (3) Judgement (3) Errors list: Planning (4) Dexterity (6) Visuospatial ability (6) Operative flow (3) Judgment (6) P/F/C: Overall performance (1) Indicative skill (1)	
Multiple specialties					
Beard et al[10]	OSATS (modified by specialty), PBA (varied by specialty)	Multiple	Varied[d]	TSC[d]: Consent Preoperative planning Preoperative preparation Exposure and closure Intraoperative technique Postoperative management P/F/C[d]	Also compared scores with nontechnical skills for surgeons
Niitsu et al[28]	OSATS	Multiple	GRS (7)	OSATS (7)	
Steigerwald et al[37]	OSATS, GOALS	Laparoscopic cholecystectomy	GRS (12)	OSATS (7) GOALS (5)	Purpose was to compare validity of the two scales
Marriott et al[27]	PBA	Multiple	PBA (3) Global summary (1)	PBA: Preoperative preparation (1) Exposure and closure (1) Intraoperative technique (1) Global summary (1)	

Abbreviations: TSC indicates task-specific checklist; GRS, global rating scale; OR, operating room; OSATS, objective structured assessment of technical skills; GOALS, global objective assessment of laparoscopic skills; VAS, visual analog scale; P/F/C, pass/fail or competent/not competent; PBA, procedure-based assessment.

[a]Included articles were English language, peer reviewed, and published in or after 1996.

[b]In this instance, the areas of assessment (number of assessments) for the OSATS were respect for tissue (1), time and motion (1), instrument handling (1), knowledge of instruments (1), flow of operation (1), use of assistants (1), and knowledge of specific procedure (1), and also included a TSC and overall pass/fail score (based on Martin et al[50]).

[c]In this instance, the areas of assessment (number of assessments) for the GOALS were depth perception (1), bimanual dexterity (1), efficiency (1), tissue handling (1), and autonomy (1).

[d]Please refer to original text for a full list of items.