Surg Endosc (2016) 30:832–844
DOI 10.1007/s00464-015-4285-8

REVIEW

A systematic review of performance assessment tools for laparoscopic cholecystectomy

Yusuke Watanabe[1,2] · Elif Bilgic[1] · Ekaterina Lebedeva[3] · Katherine M. McKendy[1] · Liane S. Feldman[1] · Gerald M. Fried[1] · Melina C. Vassiliou[1]

Received: 30 March 2015 / Accepted: 23 May 2015 / Published online: 20 June 2015

Abstract
Background Multiple tools are available to assess clinical performance of laparoscopic cholecystectomy (LC), but there are no guidelines on how best to implement and interpret them in educational settings. The purpose of this systematic review was to identify and critically appraise LC assessment tools and their measurement properties, in order to make recommendations for their implementation in surgical training.
Methods A systematic search (1989–2013) was conducted in MEDLINE, Embase, Scopus, Cochrane, and grey literature sources. Evidence for validity (content, response process, internal structure, relations to other variables, and consequences) and the conditions in which the evidence was obtained were evaluated.
Results A total of 54 articles were included for qualitative synthesis. Fifteen technical skills and two non-technical skills assessment tools were identified. The 17 tools were used for either: recorded procedures (nine tools, 60 %), direct observation (five tools, 30 %), or both (three tools, 18 %). Fourteen (82 %) tools reported inter-rater reliability and one reported a Generalizability Theory coefficient. Nine (53 %) had evidence for validity based on clinical experience and 11 (65 %) compared scores to other assessments. Consequences of scores, educational impact, applications to residency training, and how raters were trained were not clearly reported. No studies mentioned cost.
Conclusions The most commonly reported validity evidence was inter-rater reliability and relationships to other known variables. Consequences of assessments and rater training were not clearly reported. These data and the evidence for validity should be taken into consideration when deciding how to select and implement a tool to assess performance of LC, and especially how to interpret the results.

Keywords Cholecystectomy · Validity · Reliability · Clinical competence · Workplace-based assessment · Laparoscopy

Laparoscopic cholecystectomy (LC) is one of the most commonly performed procedures in surgical training. While many instruments purport to measure LC performance, it is not clear which assessment tool can best meet the needs of training programs, and under which conditions. Assessment can be used in various ways: formative assessments to provide useful feedback during training, and summative assessments to demonstrate evidence of competence with the goal of increasing patient safety [1]. If the purpose of the tool is to confirm competency at the end of training or for credentialing purposes, it is critical that robust evidence be available to support the validity and reliability of the assessment.

✉ Yusuke Watanabe
ywatanabe328@gmail.com

Melina C. Vassiliou
melina.vassiliou@mcgill.ca

[1] Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation, McGill University Health Centre, 1650, Cedar Avenue, L9. 316, Montreal, QC H3G 1A4, Canada

[2] Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Sapporo, Hokkaido, Japan

[3] The Henry K.M. De Kuyper Education Centre, McGill University Health Centre, Montreal, QC, Canada

A variety of LC performance assessment instruments have been developed and tested in different settings (e.g., bench-top models, animal models and in the operating room) and for different purposes (e.g., research outcome, formative feedback, competency assessment). However, evidence for validity under one set of conditions cannot necessarily be assumed when the assessment is used in another setting or for another indication. There is no systematic review of performance assessments available for LC that appraises the tools using a contemporary framework of validity [2]. The purpose of this review is to identify LC performance assessment tools and to provide critical appraisal of their measurement properties using the unitary framework of validity. This will ultimately support the informed selection and implementation of these tools in surgical training.

Materials and methods

Search strategy

We performed a systematic literature search of all full-text articles published between January 1989 and April 2013 according to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines(3). Search strategies were developed with the assistance of a health science librarian (E.L). A systematic search was completed in May 2014 in MEDLINE, Embase, Scopus, and Cochrane as well as in grey literature sources (LILACS, Scirrus, ProQuest Dissertations & Theses, Bandolier, Current Controlled Trials, Clinical Trials.gov, Thesis.com, and Google Scholar). No geographical or language limits were applied. Articles written in languages other than English were assessed through their English abstract only, if available. Reference lists were hand-searched to identify additional studies. The search terms used were "laparoscopic cholecystectomy" AND "clinical competence" OR "assessment" and thesaurus terms such as Medical Subject Headings (MeSH) terms and Emtree terms. To increase the sensitivity of the search strategy, we combined key words with thesaurus terms individually (key words AND thesaurus terms). A more detailed search strategy is provided in the "Appendix" and is available on request.

Study selection

Eligible studies described observational assessment tools used for LC in the operating room (OR). Studies using assessment tools for LC exclusively outside of the OR, such as in simulated settings, as well as reviews, meeting abstracts, editorials, and letters were excluded.

Data extraction

All studies were assessed independently by two reviewers (Y.W. and E.B.). Differences in data abstraction were resolved through consensus adjudication. Extracted information included study characteristics, characteristics of performance assessment tools using predefined criteria (Table 1), and validity evidence according to a contemporary framework of validity.

Validity

Validity is defined as appropriate interpretation of assessment results; a validation study is a process of collecting evidence to support the interpretations of assessment results [2, 3]. The five sources of validity (content, response process, internal structure, relations to other variables, and consequences) were evaluated according to the *Standards* established by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education [2, 4, 5]. The summary of data extraction is shown in Table 2.

Results

Study characteristics

The primary search identified 1762 studies. Three hundred and thirty-seven duplicates were removed, and the remaining 1425 titles and abstracts were screened for relevance. Of these 1425 articles, 68 titles underwent full-text review, of which 54 met our inclusion criteria and were included for qualitative synthesis (Fig. 1). Characteristics of the articles included can be seen in Table 3. We excluded eight studies from further analysis for the following reasons: a unique tool with no validity evidence ($n = 2$) [6, 7], modifications of original tools without additional validation studies ($n = 4$) [8–11], and unclear descriptions of the setting in which the data were acquired ($n = 2$) [12, 13].

Tool characteristics and appraised conditions for utilization

Of the 17 unique tools identified, 15 technical skills assessment tools and two non-technical assessment tools were identified. Technical skills assessment tools were grouped into three categories: generic skills assessment tools (GA; $n = 7$), procedure-specific assessment tools (PA; $n = 4$), and a hybrid of generic and specific assessment tools (HA; $n = 4$). The operative performance rating system (OPRS; HA) and global operative assessment of

Table 1 Extracted characteristics of included performance assessment tools

Skill set	Technical or non-technical skills assessment tool
Type of items	Generic, LC-specific, or hybrid[a]
Scoring matrix	Checklist, GRS, or error rating system
No. of items	Total number of items
Total score	Sum score or mean score
Setting	Direct observation or recorded performance
Rater	Experienced observer or reviewer, attending surgeon, or/self
Location	In the OR or outside of the OR (e.g., simulated environment)

LC laparoscopic cholecystectomy, *GRS* Global Rating Scale, *OR* operating room

[a] Combination of generic and LC-specific skills

Table 2 Five sources of validity

Validity sources	Examples of validity data collection strategies	Data extracted
Content	The relationship between the tool's content and the construct it intends to measure	Expert judgment including small group discussions
		Task analysis/cognitive task analysis/hierarchical task analysis
		Consensus development strategies including Delphi method, nominal group technique, or cross-sectional expert panel
Response process	The accuracy of scoring and score interpretation	Rater training
		Score interpretations and meaning attributed to score
Internal structure	The statistical or psychometric characteristics of assessments, or psychometric properties such as reproducibility and generalizability	Reliability: inter-rater reliability, Generalizability Theory
		Item analysis such as internal consistency, inter-test reliability, or item response theory
Relations to other variables	The comparison of scores with other known outcomes, performance assessment scores, or relevant variables	The comparisons with:
		Clinical data including postoperative patient outcomes, operative time, and estimated blood loss
		Training level such as post-graduate year or case experience
		Other performance assessment scores
		Scores on a simulator or using a motion analysis device
Consequences	The intended use and the impact of assessments	The intended use of assessments: summative or formative assessment
		The application of performance assessments to training programs
		The impact of assessment usage on trainees or patients

Adapted from Downing et al. [4] and Kogan et al. [5]

laparoscopic surgery (GOALS; GA) described intended use of an assessment tool for both summative and formative evaluation [14, 15]. The procedural-based assessment (PBA; HA) has been used for formative purposes only [16]. The appraised conditions setting in which each assessment tool was validated is summarized in Table 4. Out of 17 tools, 11 (65 %) tools used a Global Rating Scale, three (18 %) were categorized as checklists (two tools included error rating), and three (18 %) error ratings. Nine (53 %) tools were used by experienced surgeons or reviewers to assess recorded cases, and five (30 %) were used for assessment during direct observation; three (18 %) were used for both. OPRS, GRITS, and OpRate reported routine implementation of the assessment tools in surgery residency programs.

While OPRS, GOALS, Scott's objective structured assessment of technical skill (OSATS; GA) tools had evidence for either direct observation or recorded assessment, OPRS and Scott's OSATS were recommended for direct observation. OPRS and GOALS were used for assessment

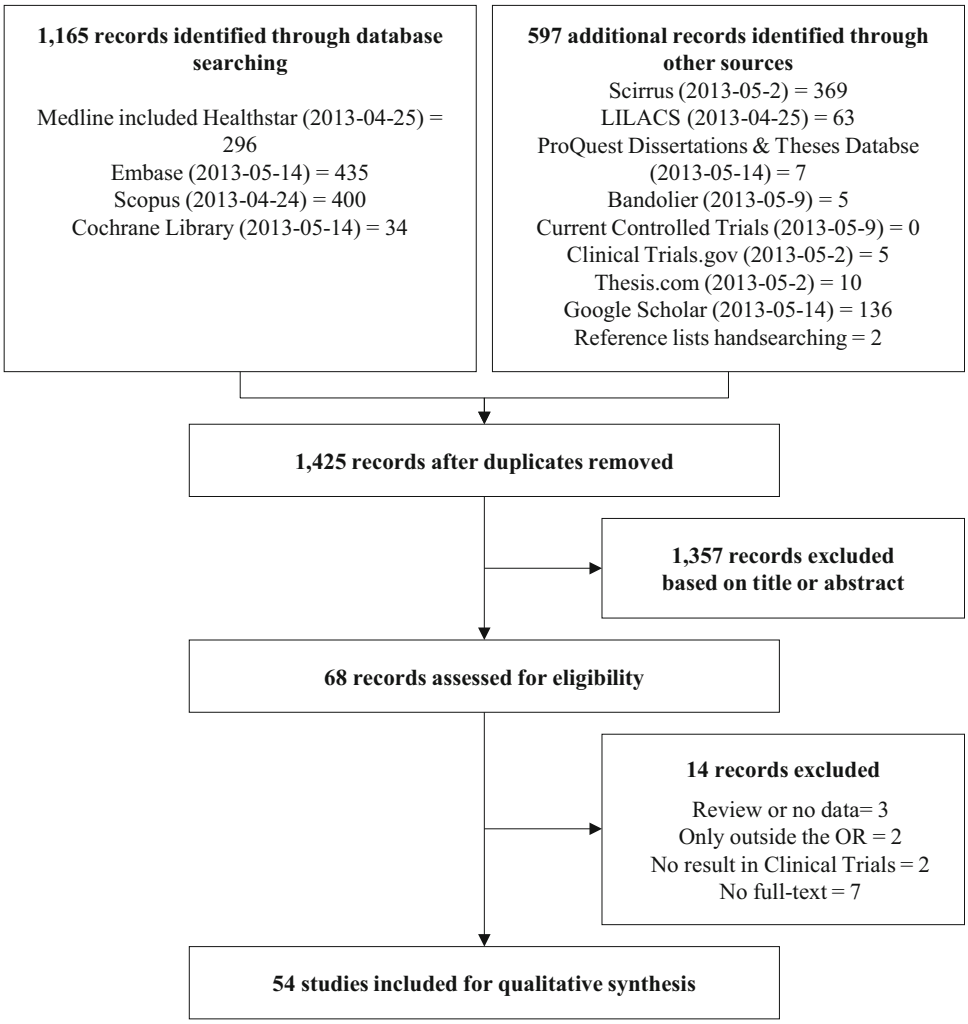

Fig. 1 Study identification and selection flow chart

Table 3 Characteristics of 54 studies describing performance assessment tools in laparoscopic cholecystectomy

Characteristics	No. (%)
Country	
US	20 (36)
UK	17 (32)
Canada	9 (17)
Netherlands	3 (6)
Others	5 (9)
Publication (year)	
1998–2005	16 (30)
2006–2013	38 (70)
Study design	
Development of tools and/or validation study	37 (69)
Utilization of tools for educational intervention study	11 (20)
Development or validation and use in education intervention study	6 (11)
Cost mentioned	0 (0)
Institutional review board approval mentioned	30 (57)

Table 4 Description of performance assessment tools and supportive evidence in different conditions

Tool	Type of items	No. of items	Total score	Setting				Location	
				Recorded	Direct observation				
				Reviewer	Observer	Attending	Self	In the OR	Outside the OR
Technical skills									
Global Rating Scale									
OPRS [14, 24, 33, 40, 41]	Hybrid[a]	10	Mean	×	×	×		×	
GOALS [15, 17, 27, 42–47]	Generic	5	25	×	×	×	×	×	×
OSATS									
Original [25, 28, 48, 49]	Generic	7	35	×				×	
Grantcharov's [12, 34, 48, 50, 51]	Generic	4	20	×				×	
Scott's [34, 48, 50]	Generic	8	0	×	×			×	
GRITS									
Original [35]	Generic	9	Mean			×		×	
Moldovanu's [18]	Generic	6	Mean	×				×	
OpRate [30, 52]	Generic	6	NR			×		×	
Sarker's GRS [20, 21, 51, 53]	Hybrid[a]	13	100	×				×	
Checklist									
PBA [16]	Hybrid[a]	15	30	×				×	×
Checklist + error rating									
Sarker's checklist [54, 55]	Hybrid[a]	27	NR	×				×	
Eubanks's checklist [8, 26, 56]	LC-specific	44	100	×				×	
Error rating									
Seymour's unnamed [57, 58]	LC-specific	8	NR	×				×	
OCHRA									
Original [19, 31, 59]	LC-specific	22	NR	×				×	
Misha's [60]	LC-specific	32	NR		×			×	
Non-technical skills									
Global Rating Scale									
NOTECS [22, 32, 60, 61]	Generic	4	16		×			×	
OTAS [23, 62]	Generic	5	30		×			×	

NR not report, *LC* laparoscopic cholecystectomy, *GRS* Global Rating Scale, *OPRS* operative performance rating system, *GOALS* global operative assessment of laparoscopic surgery, *OSATS* objective structured assessment of technical skill, *GRITS* global rating index for technical skills, *PBA* procedural-based assessment, *OCHRA* observation clinical human reliability assessment, *NOTECHS* non-technical skills, *OTAS* observational teamwork assessment for surgery

[a] Combination of generic and LC-specific skills

by direct observation by multiple raters. Only the PBA supported its use both in the OR and in the simulation setting (LapMentor virtual reality (VR) simulator, Simbionix, Ltd., Israel). Only GOALS was evaluated in the clinical and animal laboratory setting [16, 17]. None of the technical skills assessment tools were used in human cadaver training. To increase the sensitivity, Moldovanu et al. [18] used a global rating index for technical skill (GRITS; GA) to rate each procedural step (exposure of biliary region and adhesiolysis, dissection of the cystic pedicle and critical view, and dissection of gallbladder) as well as overall performance. The generic items of the technical assessment tools were based on either OSATS or GOALS, and the LC-specific tools were based on task analysis or hierarchical task analyses [19, 20].

Validity evidence

In this section, the results of the studies included are analyzed on the basis of the sources of validity evidence

specified in the unitary framework [2, 4]. The reported validity evidence of the performance assessment tools is summarized in Table 5.

Content

Within technical skills assessment tools, two hybrid tools developed by Sarker et al. (Sarker's Global Rating Scale and PBA) and the observation clinical human reliability assessment (OCHRA; PA) which is an error rating tool were developed based on task analyses using training manuals and the technical protocol of the operation [19, 21]. The other tools were developed by expert judgment including institutional expert panels. None of the tools used a comprehensive strategic method, which includes task analysis or cognitive task analysis, a cross-sectional expert panel, and the process of achieving consensus such as Delphi methodology or nominal group technique.

Response process

OPRS, PBA, non-technical skills (NOTECHS), and observational teamwork assessment for Surgery (OTAS) include user manuals for raters. NOTECHS is associated with concrete evidence of rater training [22]. Only two studies reported rater training clearly before the implementation of a tool [10, 23]. OPRS, GOALS, OpRate, and GRITS described orienting raters to the tool via informal techniques or preexisting institutional faculty meetings. OPRS used behavior anchors on overall scores: a rating of four or higher indicating technical proficiency and ability to perform operations independently. The anchor assumes that a resident consistently performs at this level and has met institutional benchmarks for achievement. All residents must be evaluated at least three times, by a minimum of two different raters, and with no ratings of three or less. PBA also has behavior anchors, for example, a satisfactory standard for certification level or development required.

Internal structure

Inter-rater reliability was reported for 12 technical skills assessment tools and two nontechnical assessment tools, and was the most commonly reported evidence for raters. However, there was no consistent way of calculating inter-rater reliability; techniques used included intraclass correlation coefficient, internal consistency (Cronbach's α), and Kohen κ coefficient. Four technical skills assessment tools reported item analyses; internal consistency was described for GOALS, GRITS, and OpRate; inter-item correlations were analyzed for OPRS; item-total correlations were described for GOALS. The reliability coefficient of Generalizability Theory was reported for OPRS, delineating the number of assessment scores per month that would be desirable in residency training, in order to achieve a valid assessment of performance by direct observation [24]. No studies reported data using item response theory.

Relations to other variables

None of the studies attempted to investigate the relationship between performance scores and patient outcomes. OSATS and Eubanks's checklist compared scores with operative time [25, 26]. Performance scores were compared across training levels in nine (53 %) tools, and all studies demonstrated improved scores with increasing levels of training. Comparison with other performance assessment scores was described for nine (53 %) tools. Comparison with simulation scores, written exams, and Objective Structured Clinical Examinations (OSCE) was less common than comparison to training levels: GOALS versus bench-top simulation scores (McGill Inanimate System for Training and Evaluation of Laparoscopic Skills: MISTELS) [27], original OSATS versus motion tracking data [25] or VR scores [28], Scott's OSATS versus American Board of Surgery In-Training Examination (ABSITE) or bench-top simulation scores (Southwestern Center for Minimally Invasive Surgery Guided Endoscopic Module: SCMIS GEM) [29], and OpRate versus VR scores [30], modified Eubanks's checklist versus motion tracking data [8], OCHRA versus OSCE [31] or NOTECHS scores [22, 32].

Consequences

Only OPRS and GOALS reported the intended use clearly which are for formative and summative assessments [14, 15]. OPRS, GRITS and OpRate reported routine implementation of the assessment tools in surgery residency programs, using scores to identify residents who required remediation, indicating that the intended use could be for summative assessment. The OPRS is used to establish benchmarks that residents should achieve prior to advancing to the next level of training [33]. There were no investigations determining pass/fail scores as a summative assessment or predicting patient outcomes from the assessment scores which may represent the quality of their performance. The educational impact of using the tools for providing feedback was reported for Grantcharov's OSATS and GRITS [34, 35]. Figure 2 proposes an algorithm for selecting and implementing LC performance assessment tools in residency training according to existing evidence.

Table 5 Validity evidence of performance assessment tools in laparoscopic cholecystectomy

	OPRS [14, 24, 33, 40, 41]	GOALS [15, 17, 27, 42–47]	OSATS			GRITS		OpRate [30, 52]	Sarker's GRS [20, 21, 51, 53]
			Original [25, 28, 48, 49]	Grantcharov's [12, 34, 48, 50, 51]	Scott's [34, 48, 50]	Original [35]	Moldovanu's [18]		
Content									
Expert judgment	+	+	+	+	+	+	+	+	
Task analysis/CTA/HTA									+
Consensus method[a]									
Response process									
Rater training									
Score interpretations and meaning attributed to score	+								
Internal structure									
Inter-rater reliability	+	+	+	+	+		+		+
Item analysis	+	+				+		+	
GT coefficient	+								
Relations to other variables									
Operative data			+						
Training level or case experience	+	+	+	+	+	+		+	+
Other performance assessment scores		+	+	+					+
Others[b]		+	+		+			+	
Consequences									
Applications to residency program	+					+		+	
Criterion-referenced score (benchmark score)	+								
Educational impact[c]				+		+			

	PBA [16]	Sarker's checklist [54, 55]	Eubanks's checklist [8, 26, 56]	Seynour's unnamed [57, 58]	OCHRA		NOTECHS [22, 32, 60, 61]	OTAS [23, 62]
					Original [19, 31, 59]	Misha's [60]		
Content								
Expert judgment			+	+			+	+
Task analysis/CTA/HTA	+	+			+	+		
Consensus method[a]								
Response process								
Rater training								+
Score interpretations and meaning attributed to score	+							
Internal structure								
Inter-rater reliability	+	+	+	+	+		+	+
Item analysis							+	
GT coefficient								

Table 5 continued

	PBA [16]	Sarker's checklist [54, 55]	Eubanks's checklist [8, 26, 56]	Seynour's unnamed [57, 58]	OCHRA		NOTECHS [22, 32, 60, 61]	OTAS [23, 62]
					Original [19, 31, 59]	Misha's [60]		
Relations to other variables								
Operative data			+					
Training level or case experience	+	+	+					
Other performance assessment scores			+		+	+	+	+
Others[b]			+					
Consequences								
Applications to residency program								
Criterion-referenced score (benchmark score)								
Educational impact[c]								

OPRS operative performance rating system, *GOALS* global operative assessment of laparoscopic surgery; *OSATS* objective structured assessment of technical skill, *GRITS* global rating index for technical skills, *PBA* procedural-based assessment, *NOTECHS* non-technical skills, *OTAS* observational teamwork assessment for surgery, *CTA* cognitive task analysis, *HTA* hierarchical task analysis, *GT* generalizability theory

[a] Consensus method such as Delphi method, nominal group technique, or cross-sectional expert panel

[b] Others include video trainer scores, virtual reality simulation scores, or data of motion analysis

[c] The impact of assessment usage on trainees' learning

Discussion

Our results provide a summary of LC performance assessment tools including the conditions for their implementation in training and their validity evidence based on a contemporary validity framework using a systematic approach. From the validity evidence framework, our systematic review reports that the validity evidence for the internal structure and relations to the other variables are more commonly demonstrated. However, the validity evidence for the content, response process, and consequences aspects are limited. To apply LC assessment tools in surgical training, there may be a need to acquire additional validity evidence, depending on the intended use and consequences of the results.

Assessment of surgical competence has historically emphasized the need to adopt careful scientific methodology in order to establish validity evidence. Until recently, the methodology usually applied in surgical education has been based on an outdated validity framework, which includes concepts such as construct, content, and criterion-related validity. The most recently accepted framework of validity is based on identifying evidence from multiple sources including content, response process, internal structure, relations to other variables, and consequences of assessment. Validity states "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests." [2]. Validity evidence should be gathered for the intended use of performance assessments and not as a property of the assessment tool itself. The quality of the validity evidence should therefore be analyzed and interpreted according to its intended use and the conditions and environment in which the evidence was obtained. As a result, the commonly used term "validated assessment tool" is often inaccurate, as it refers to the tool itself. The evidence for validity under one set of conditions is often expanded and applied incorrectly to a new setting when implementing and interpreting the results. For example, a tool used by trained evaluators to measure technical skills during an OSCE in a simulated environment may not perform in the same way if used by untrained surgeons to assess laparoscopic skills in the clinical environment. Conditions and the interpretation of scores in a given study would have to be reproduced in order to be implemented in surgical training for a different purpose or under different conditions. As another example, some instruments have had evidence for direct observation but were used for blinded assessment of recorded procedures, while others have been described only in the OR but are applied in simulated environments.

With the lack of a definitive consensus regarding the desirable conditions of performance assessments, the following factors should be considered when applying performance assessment instruments to residency training: the

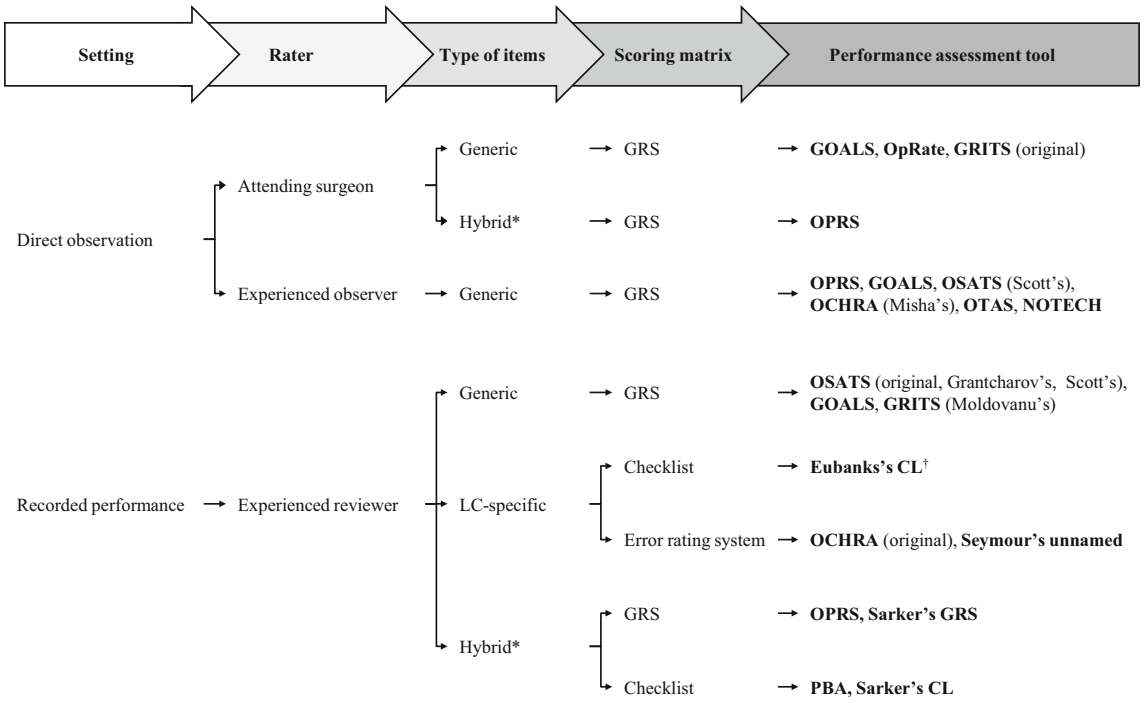

Fig. 2 The selection of an assessment tool in laparoscopic cholecystectomy. *LC* laparoscopic cholecystectomy, *GRS* Global Rating Scale, *CL* checklist, *OPRS* operative performance rating system, *GOALS* global operative assessment of laparoscopic surgery, *OSATS* objective structured assessment of technical skill, *GRITS* global rating index for technical skills, *PBA* procedural-based assessment, *NOTECHS* non-technical skills, *OTAS* observational teamwork assessment for surgery, *OCHRA* observation clinical human reliability assessment. *Asterisk* combination of generic and LC-specific items; *dagger symbol* Eubanks's checklist includes error rating

purpose, the rater and other conditions. Surgical performance is classically assessed in two ways: by direct observation or by assessment of a recorded procedure. Direct observation by an attending surgeon would be practical for formative evaluations in regular training curricula because it is practical, immediate and requires little extra time or equipment. Assessment of a recorded case may have value for credentialing purposes and offer the benefit of blinding the rater so as to reduce potential bias related to the relationship between the rater and the trainee. The OSATS-derived performance assessment systems are well known, although most have evidence for blinded assessment of recorded LCs. Thus, additional evidence might be required to use this tool for assessing performance by direct observation, such as for formative assessment. OPRS and GOALS have evidence for both direct observation and videotaped evaluations. The OPRS, which is a hybrid tool, is recommended to be used during direct observation by attending surgeons, or could be used in combination with audio of the OR team for videotaped assessments. GOALS was also designed for direct observation, and it can be used not only for self-assessment but also for videotaped evaluations using only the laparoscopic view without audio recordings. Most generic items were composed of OSATS, GOALS, or a combination of parts of both.

GOALS can be scored by direct observation or video of the laparoscopic view, while using OSATS for videotaped assessment might be challenging since it includes items that cannot be assessed just by the laparoscopic video such as knowledge of instruments, knowledge of anatomy, knowledge of specific procedure, and use of assistants. Error rating tools such as OCHRA have been used based on videotaped evaluations, though the feasibility of this approach for routine implementation has been questioned due to the limited resources of trained raters and time. PBA can be used in the VR setting (LapMentor, Simbionix, Ltd., Israel), and GOALS can be used in the porcine model in addition to the OR. Therefore, when applying other tools in simulated environments, further investigations are required.

When selecting a LC assessment tool, what is being assessed and how the results will be used in your training program are essential. Although hybrid or procedure-specific tools are preferred because the trainee can obtain more specific feedback, the role of the trainee during a LC in your program should also be considered. If more than two trainees have different roles based on their training levels to perform a LC, for example, the senior resident dissects the Calot's triangle and then the junior resident removes the gallbladder from liver bed, it might be challenging to use procedure-specific or hybrid tools for one

resident. Although the generic assessment tools are flexible and suitable to assess trainees' performance in these situations, the concrete goals of each step of the procedural are not described or assessed specifically.

Evidence for previously established construct, face, content, and criterion-related validity, based on more dated frameworks of validity, was abundant in the studies that were reviewed. For content validity, all tools for LC performance assessment were developed using either local experts or task analysis, or both. To have more reliable assessment tools, comprehensive item development strategies should be used which could include cognitive task analysis, cross-sectional expert panels, or consensus development methods such as Delphi method or nominal group technique.

Rater training, included in the response process of validity, was minimally described but crucial for reliable assessments. Although raters need training to rate learners' performance reliably and discriminate between performance levels, it might be a challenge to implement rater training due to perceived cost, time constraints, or unawareness of the importance of rater training. Rater training also includes rater knowledge of the meaning of assessment scores and the consequences of the scores on the trainee. Surgical residents can be given instructional resources, such as videos, demonstrating expected LC performance. For example, Ahlberg et al. [10] used the modified Seymour error rating tool to measure the effect of virtual reality simulation training on LC performance. All subjects and attending surgeons viewed an instructional video including all defined errors. The two raters were trained to reach predefined inter-rater reliability before the study to improve validity of response process, but inter-rater reliability was not reported. The impact of response process on assessment scores is unclear, but should be considered.

There was abundant evidence in terms of inter-rater reliability as internal structure of validity. The other reliabilities such as internal consistency, and inter-item/item-total correlations, are important to evaluate whether each item is measuring skills required to perform LC. However, these reliabilities cannot assess if the reliability of an instrument is affected by other factors such as different procedures, the quality of supervision, or the difficulty of the procedure. They also cannot evaluate the desired number of items, cases, and raters necessary in certain conditions. Generalizability Theory calculates the independent variability attributed to these factors and therefore can assess how these factors may affect reliability [36, 37]. Additionally, the scores of each item could have various meanings and have different impacts on the entire performance score, so item response theory could help clarify these aspects and weight each item by its difficulty and discriminative power [38].

In many studies, the comparison between assessment scores and experience level (post-graduate level, case experience), other instruments, and scores on simulators were described heterogeneously in order to demonstrate relations to the other variables. To investigate this component of validity evidence, a consensus about what data are meaningful might help to provide a common language for this type of research, thus allowing comparisons between different performance assessment tools. Although it is no longer feasible to compare scores with LC-related complications as they are infrequent and resident performance is usually supervised by a senior surgeon in residency training, whether the scores of these assessment tools are associated with patient outcomes remains an area for future research [39].

The consequences component of validity refers to the impact of the assessment, decisions, and outcomes, as well as the impact of assessments on teaching and learning. In other words, the intended purpose of the assessment tool's use and how to interpret the scores are very important. Although this aspect of validity is solidly embodied in the current *Standards*, it is relatively unstudied and reported ambiguously. It could have a profound impact on the identification of trainees who need remediation or for a certification process, or increasing learners' motivation when used for formative purposes. Assessment is important for both summative and formative purposes. Summative evaluations are completed at the end of a training period and play a role in determining whether an individual has achieved expected levels to move on to the next step of training or, perhaps to perform procedures independently, or be considered competent. Formative assessments are used at regular intervals to track progress and to provide constructive feedback with the goal of helping the learner improve. Within the validation process, different types and amounts of validity evidence are needed depending on the intended uses and consequences associated with assessment tools. For instance, a formative assessment might require a different amount of validity evidence, but not necessarily less rigorous, from a summative assessment which might be used to decide whether an individual is competent or should have privileges to perform a procedure.

It is tempting to pay attention to characteristics and contents of the assessment tool, but caution must be exercised in interpreting the results of tools used in conditions other than those for which evidence for its validity exist. If one desires to use an assessment tool under conditions other than those for which the tool has validity evidence, then, depending on the intended purpose, the tool should ideally be validated for the applied setting before the application of the tool.

In conclusion, this study provides a review of the assessment instruments available for laparoscopic cholecystectomy

and the validity evidence associated with each based on the most current framework. We also provide recommendations about how to select the tool that best meets your training needs. In the end, the goal is to try and provide assessments of trainees performing laparoscopic cholecystectomy that represents their true skill level as much as possible. This will increase the efficiency of education and hopefully have a positive impact on patient safety.

Disclosures The Steinberg-Bernstein Centre for Minimally Invasive Surgery and Innovation received an unrestricted educational grant from Covidien Canada. Y. Watanabe, E. Bilgic, E. Lebedeva, K.M. McKendy, L.S. Feldman, G.M. Fried, M.C. Vassiliou have no relevant conflicts of interests or financial ties to disclose.

Appendix: Search strategy of MEDLINE

1 exp cholecystectomy/or cholecystectomy, laparoscopic/
2 exp professional competence/or exp clinical competence/
3 (Task Performance and Analysis).mp. [mp = title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]
4 exp study characteristics/or exp evaluation studies/
5 (Internship and Residency).mp. [mp = title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]
6 1 and 2 and 3
7 exp Curriculum/ed, mt, sn [Education, Methods, Statistics & Numerical Data]
8 exp validation studies/
9 exp Laparoscopy/ed, mt, st [Education, Methods, Standards]
10 1 and 2 and 3 and 8
11 1 and 2 and 8
12 limit 1 to systematic reviews
13 2 and 12
14 4 and 8
15 1 and 2 and 14
16 3 and 15
17 exp self concept/or self-assessment/
18 exp methods/or exp observation/or exp research design/
19 (decision making and clinical competence$ and skill$).ab.
20 1 and 2 and 18 and 19
21 Educational Measurement/and "Internship and Residency"/
22 1 and 2 and 21
23 1 and 9 and 17
24 1 and 2 and 5
25 evaluation studies/or evaluation studies as topic/or program evaluation/or validation studies as topic/or Intervention Studies/or (effectiveness or (pre- adj5 post-)).ti,ab. or (program* adj3 evaluat*).ti,ab. or intervention*.ti,ab.
26 1 and 2 and 25
27 exp Clinical Trial/or double-blind method/or (clinical trial* or randomized controlled trial or multicenter study).pt. or exp Clinical Trials as Topic/or ((randomi?ed adj7 trial*) or (controlled adj3 trial*) or (clinical adj2 trial*) or ((single or doubl* or tripl* or treb*) and (blind* or mask*))).ti,ab.
28 1 and 2 and 27
29 (exp methods/or exp observation/or exp research design/) and #1.mp. and #9.mp. [mp = title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]
30 1 and 17 and 29
31 17 not patients.mp. [mp = title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]
32 1 and 9 and 31
33 (Task Performance and Analysis).tw
34 1 and 2 and 33
35 1 and 2 and 4 and 33
36 1 and 2 and 4 and 5
37 (self concept or self-assessment).tw
38 ((self concept or self-assessment) not patients).tw
39 1 and 2 and 4 and 5 and 38
40 1 and 2 and 38
41 1 and 33 and 38
42 1 and 2 and 5 and 7
43 curriculum.tw
44 1 and 5 and 43 (
45 validation studies.tw
46 1 and 2 and 29 and 45
47 1 and 5 and 18
48 1 and 5 and 25
49 1 and 5 and 27
50 1 and 17 and 25
51 1 and 21 and 25
52 1 and 5 and 9 and 21 and 25
53 decision$making.tw
54 (Educational adj2 assessment).tw
55 (General surgery adj2 training).mp
56 (objective adj2 assessment).mp
57 Non$technical skill$.mp
58 ((performance adj2 assessment) or (performance adj2 evaluation)).tw
59 (surgical adj2 assessment tool$).mp
60 (surgical adj2 skill$).mp
61 Technical error$.tw
62 (Resident adj2 evaluation).tw
63 (simulator adj2 training).tw
64 ((mental adj2 training) and (mental adj2 practice)).tw
65 (motor adj2 skill$).tw
66 (Intraoperative adj2 performance).tw
67 human error$.tw

68 direct observation.tw
69 (acquisition adj2 skil$).tw
70 or/53–69
71 1 and 70
72 feedback.tw
73 expert testimony.tw
74 Confidence Intervals.tw
75 video recording.tw
76 operating rooms.tw
77 simulation.tw
78 or/72–77
79 1 and 70 and 78

References

1. Vassiliou MC, Feldman LS (2011) Objective assessment, selection, and certification in surgery. Surg Oncol 20:140–145
2. The American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (1999) Standard for educational and psychological testing. American Educational Research Association, Washington
3. Ghaderi I, Manji F, Park YS, Juul D, Ott M, Harris I, Farrell TM (2014) Technical skills assessment toolbox: a review using the unitary framework of validity. Ann Surg. doi:10.1097/SLA.0000000000000520
4. Downing SM (2003) Validity: on meaningful interpretation of assessment data. Med Educ 37:830–837
5. Kogan JR, Holmboe ES, Hauer KE (2009) Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA 302:1316–1326
6. Schijven MP, Jakimowicz JJ, Broeders IAMJ, Tseng LNL (2005) The Eindhoven laparoscopic cholecystectomy training course–improving operating room performance using virtual reality training: results from the first E.A.E.S. accredited virtual reality trainings curriculum. Surg Endosc 19:1220–1226
7. van Det MJ, Meijerink WJHJ, Hoff C, Middel B, Pierie JPEN (2013) Effective and efficient learning in the operating theater with intraoperative video-enhanced surgical procedure training. Surg Endosc 27:2947–2954
8. Hwang H, Lim J, Kinnaird C, Nagy AG, Panton ONM, Hodgson AJ, Qayumi KA (2006) Correlating motor performance with surgical error in laparoscopic cholecystectomy. Surg Endosc 20:651–655
9. van Det MJ, Meijerink WJHJ, Hoff C, Middel LJ, Koopal SA, Pierie JPEN (2011) The learning effect of intraoperative video-enhanced surgical procedure training. Surg Endosc 25:2261–2267
10. Ahlberg G, Enochsson L, Gallagher AG, Hedman L, Hogman C, McClusky DA, Ramel S, Smith CD, Arvidsson D (2007) Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. Am J Surg 193:797–804
11. Gauger PG, Hauge LS, Andreatta PB, Hamstra SJ, Hillard ML, Arble EP, Kasten SJ, Mullan PB, Cederna PS, Minter RM (2010) Laparoscopic simulation training with proficiency targets improves practice and performance of novice surgeons. Am J Surg 199:72–80
12. Mohan P, Chaudhry R (2009) Laparoscopic simulators: are they useful! Med J Arm Forces India 324:1073–1078
13. Hamilton EC, Scott DJ, Fleming JB, Rege RV, Laycock R, Bergen PC, Tesfay ST, Jones DB (2002) Comparison of video trainer and virtual reality training systems on acquisition of laparoscopic skills. Surg Endosc 16:406–411
14. Williams RG, Sanfey H, Chen XP, Dunnington GL (2012) A controlled study to determine measurement conditions necessary for a reliable and valid operative performance assessment: a controlled prospective observational study. Ann Surg 256:177–187
15. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D, Fried GM (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190:107–113
16. Sarker SK, Maciocco M, Zaman A, Kumar I (2010) Operative performance in laparoscopic cholecystectomy using the procedural-based assessment tool. Am J Surg 200:334–340
17. Hogle NJ, Chang L, Strong VEM, Welcome AOU, Sinaan M, Bailey R, Fowler DL (2009) Validation of laparoscopic surgical skills training outside the operating room: a long road. Surg Endosc 23:1476–1482
18. Moldovanu R, Târcoveanu E, Dimofte G, Lupaşcu C, Bradea C (2011) Preoperative warm-up using a virtual reality simulator. J Soc Laparoendosc Surg 15:533–538
19. Joice P, Hanna GB, Cuschieri A (1998) Errors enacted during endoscopic surgery—a human reliability analysis. Appl Ergon 29:409–414
20. Sarker SK, Chang A, Vincent C, Darzi SAW (2006) Development of assessing generic and specific technical skills in laparoscopic surgery. Am J Surg 191:238–244
21. Sarker SK, Chang A, Albrani T, Vincent C (2008) Constructing hierarchical task analysis in surgery. Surg Endosc 22:107–111
22. Mishra A, Catchpole K, McCulloch P (2009) The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Qual Saf Health Care 18:104–108
23. Russ S, Hull L, Rout S, Vincent C, Darzi A, Sevdalis N (2012) Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. Ann Surg 255:804–809
24. Williams RG, Verhulst S, Colliver JA, Sanfey H, Chen X, Dunnington GL (2012) A template for reliable assessment of resident operative performance: assessment intervals, numbers of cases and raters. Surgery 152:517–527
25. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Papasavas P, Dosis A, Bello F, Darzi A (2007) An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Ann Surg 245:992–999
26. Eubanks TR, Clements RH, Pohl D, Williams N, Schaad DC, Horgan S, Pellegrini C (1999) An objective scoring system for laparoscopic cholecystectomy. J Am Coll Surg 189:566–574
27. Fried GM, Feldman LS, Vassiliou MC, Fraser SA, Stanbridge D, Ghitulescu G, Andrew CG (2004) Proving the value of simulation in laparoscopic surgery. Ann Surg 240:518–528
28. Kundhal PS, Grantcharov TP (2009) Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. Surg Endosc 23:645–649
29. Scott DJ, Valentine RJ, Bergen PC, Rege RV, Laycock R, Tesfay ST, Jones DB (2000) Evaluating surgical competency with the American Board of Surgery In-Training Examination, skill testing, and intraoperative assessment. Surgery 128:613–622
30. Wohaibi EM, Bush RW, Earle DB, Seymour NE (2010) Surgical resident performance on a virtual reality simulator correlates with operating room performance. J Surg Res 160:67–72
31. Tang B, Hanna GB, Carter F, Adamson GD, Martindale JP, Cuschieri A (2006) Competence assessment of laparoscopic operative and cognitive skills: objective structured clinical

examination (OSCE) or observational clinical human reliability assessment (OCHRA). World J Surg 30:527–534
32. Catchpole K, Mishra A, Handa A, McCulloch P (2008) Teamwork and error in the operating room: analysis of skills and roles. Ann Surg 247:699–706
33. Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL (2005) Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. Surgery 138:640–649
34. Grantcharov TP, Schulze S, Kristiansen VB (2007) The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. Surg Endosc 21:2240–2243
35. Doyle JD, Webber EM, Sidhu RS (2007) A universal Global Rating Scale for the evaluation of technical skills in the operating room. Am J Surg 193:551–555
36. Crossley J, Davies H, Humphris G, Jolly B (2002) Generalisability: a key to unlock professional assessment. Med Educ 36:972–978
37. Bloch R, Norman G (2012) Generalizability theory for the perplexed: a practical introduction and guide: AMEE guide no. 68. Med Teach 34:960–992
38. Downing SM (2003) Item response theory: applications of modern test theory in medical education. Med Educ 37:739–745
39. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJO, Michigan Bariatric Surgery Collaborative (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369: 1434–1442
40. Sanfey H, Williams RG, Chen X, Dunnington GL (2011) Evaluating resident operative performance: a qualitative analysis of expert opinions. Surgery 150:759–770
41. Kim MJ, Williams RG, Boehler ML, Ketchum JK, Dunnington GL (2009) Refining the evaluation of operating room performance. J Surg Educ 66:352–356
42. Gumbs AA, Hogle NJ, Fowler DL (2007) Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. J Am Coll Surg 204:308–313
43. Vassiliou MC, Feldman LS, Fraser SA, Charlebois P, Chaudhury P, Stanbridge DD, Fried GM (2007) Evaluating intraoperative laparoscopic skill: direct observation versus blinded videotaped performances. Surg Innov 14:211–216
44. Chang L, Hogle NJ, Moore BB, Graham MJ, Sinanan MN, Bailey R, Fowler DL (2007) Reliable assessment of laparoscopic performance in the operating room using videotape analysis. Surg Innov 14:122–126
45. Choy I, Fecso A, Kwong J, Jackson T, Okrainec A (2013) Remote evaluation of laparoscopic performance using the global operative assessment of laparoscopic skills. Surg Endosc 27:378–383
46. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayez R, Fried GM (2010) Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room-a randomized controlled trial. Am J Surg 199:115–120
47. Beyer L, Troyer JD, Mancini J, Bladou F, Berdah SV, Karsenty G (2011) Impact of laparoscopy simulator training on the technical skills of future surgeons in the operating room: a prospective study. Am J Surg 202:265–272
48. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A (2008) Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. Ann Surg 247:372–379
49. Calatayud D, Arora S, Aggarwal R, Kruglikova I, Schulze S, Funch-Jensen P, Grantcharov T (2010) Warm-up in a virtual reality environment improves performance in the operating room. Ann Surg 251:1181–1185
50. Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P (2004) Randomized clinical trial of virtual reality simulation for laparoscopic skills training. Br J Surg 91:146–150
51. Palter VN, Orzech N, Reznick RK, Grantcharov TP (2013) Validation of a structured training and assessment curriculum for technical skill acquisition in minimally invasive surgery: a randomized controlled trial. Ann Surg 257:224–230
52. Wohaibi EM, Earle DB, Ansanitis FE, Wait RB, Fernandez G, Seymour NE (2007) A new web-based operative skills assessment tool effectively tracks progression in surgical resident performance. J Surg Educ 64:333–341
53. Sarker SK, Hutchinson R, Chang A, Vincent C, Darzi AW (2006) Self-appraisal hierarchical task analysis of laparoscopic surgery performed by expert surgeons. Surg Endosc 20:636–640
54. Sarker SK, Chang A, Vincent C, Darzi AW (2005) Technical skills errors in laparoscopic cholecystectomy by expert surgeons. Surg Endosc 19:832–835
55. Sarker SK, Chang A, Vincent C (2006) Technical and technological skills assessment in laparoscopic surgery. J Soc Laparoendosc Surg 10:284–292
56. Guerlain S, Adams RB, Turrentine FB, Shin T, Guo H, Collins SR, Calland JF (2005) Assessing team performance in the operating room: development and use of a "black-box" recorder and other tools for the intraoperative environment. ACS 200:29–37
57. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, Satava RM (2002) Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 236:458–464
58. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Andersen DK, Satava RM (2004) Analysis of errors in laparoscopic surgical procedures. Surg Endosc 18:592–595
59. Tang B, Hanna GB, Joice P, Cuschieri A (2004) Identification and categorization of technical errors by observational clinical human reliability assessment (OCHRA) during laparoscopic cholecystectomy. Arch Surg 139:1215–1220
60. Mishra A, Catchpole K, Dale T, McCulloch P (2008) The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. Surg Endosc 22:68–73
61. McCulloch P, Mishra A, Handa A, Dale T, Hirst G, Catchpole K (2009) The effects of aviation-style non-technical skills training on technical performance and outcome in the operating theatre. Qual Saf Health Care 18:109–115
62. Hull L, Arora S, Kassab E, Kneebone R, Sevdalis N (2011) Observational teamwork assessment for surgery: content validation and tool refinement. J Am Coll Surg 212:234–245