Maastricht University

Objective assessment of technical surgical skills

Citation for published version (APA):

van Hove, P. D., Tuijthof, G. J., Verdaasdonk, E. G., Stassen, L. P., & Dankelman, J. (2010). Objective assessment of technical surgical skills. *British Journal of Surgery*, *97*(7), 972-987. https://doi.org/10.1002/bjs.7115

Document status and date:
Published: 01/07/2010

DOI:
10.1002/bjs.7115

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 19 Apr. 2024

Review

Objective assessment of technical surgical skills

P. D. van Hove[1,2,4], G. J. M. Tuijthof[1,3], E. G. G. Verdaasdonk[2], L. P. S. Stassen[4] and J. Dankelman[1]

[1]Department of Biomechanical Engineering, Delft University of Technology, and [2]Department of Surgery, Reinier de Graaf Group, Delft, [3]Department of Orthopaedic Surgery, Amsterdam Medical Centre, Amsterdam, and [4]Department of Surgery, Maastricht University Medical Centre, Maastricht, The Netherlands

Correspondence to: Dr P. D. van Hove, Department of Biomechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, The Netherlands (e-mail: P.D.vanHove@tudelft.nl)

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Background: **Surgeons are increasingly being scrutinized for their performance and there is growing interest in objective assessment of technical skills. The purpose of this study was to review all evidence for these methods, in order to provide a guideline for use in clinical practice.**

Methods: **A systematic search was performed using PubMed and Web of Science for studies addressing the validity and reliability of methods for objective skills assessment within surgery and gynaecology only. The studies were assessed according to the Oxford Centre for Evidence-based Medicine levels of evidence.**

Results: **In total 104 studies were included, of which 20 (19·2 per cent) had a level of evidence 1b or 2b. In 28 studies (26·9 per cent), the assessment method was used in the operating room. Virtual reality simulators and Objective Structured Assessment of Technical Skills (OSATS) have been studied most. Although OSATS is seen as the standard for skills assessment, only seven studies, with a low level of evidence, addressed its use in the operating room.**

Conclusion: **Based on currently available evidence, most methods of skills assessment are valid for feedback or measuring progress of training, but few can be used for examination or credentialing. The purpose of the assessment determines the choice of method.**

Presented in part to the 21st International Conference of the Society for Medical Innovation and Technology, Sinaia, Romania, October 2009

Paper accepted 10 March 2010

Published online 14 May 2010 in Wiley InterScience (www.bjs.co.uk). **DOI:** 10.1002/bjs.7115

Introduction

Traditionally, surgical skills have been assessed in the operating room by supervision and feedback[1,2]. However, this method has been criticized for being too subjective and not representing the actual level of skills[3]. There is an increasing demand from society, governments and insurance companies for clear and transparent quality measurements in healthcare; surgeons and trainees are increasingly being scrutinized for their performance[4–6]. New techniques, such as minimal access surgery, require new skills, which have different learning curves and require different training methods outside the operating room[1,7,8]. These developments have resulted in an increased interest in objective assessment methods for surgical skills. They are currently used in surgical residency programmes for assessing the performance of trainees and to provide feedback on training. Moreover, these methods are needed as tools for examination in, for instance, different stages of training. Likewise, governments are planning to use assessment of competence of practising laparoscopic surgeons for credentialling and revalidation[1,3,4,9,10].

A number of different methods for objective assessment of surgical skills have been developed, and studies addressing their validity and reliability are abundant[5,9,11–14]. However, so far, methods for objective assessment have not been adopted widely into clinical practice. Reasons include lack of expertise, a proper infrastructure for implementation, and cost. It could also be that educators are hesitant to use these methods because it has not yet been fully defined how and where they can be used.

Previous reviews of methods for objective assessment tended only to sum up and describe different methods[9,11,13]. The purpose of this study was to provide

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

a critical review of the current evidence for objective assessment methods for technical surgical skills.

Methods

A systematic search of the literature was performed, using PubMed and Web of Science, for studies concerning validity or reliability of methods for assessment of technical surgical skills. The following query was used: '(surgical OR operative OR laparoscopic OR technical) AND (skills OR competence) AND assessment'. Studies were included that addressed assessment methods which are applicable in or outside the operating theatre and concerned open surgery or laparoscopy in the domain of general surgery and gynaecology. Studies concerning other domains were not included. Only English-language studies were included. Studies addressing the validity of specific bench models or simulator tasks were excluded. However, studies were included when they used non-validated methods for rating a bench task; such investigations contributed to validating the method as well as the bench task. Reviews and congress abstracts were excluded.

All studies were divided into separate categories based on the type of assessment method. Some studies discussed more than one assessment method and were therefore included in more than one category. The following categories were defined: procedure-specific checklists, global rating scales, motion analysis, virtual reality (VR) simulators, video assessment and miscellaneous. Extra categories were defined for Objective Structured Assessment of Technical Skills (OSATS) and Fundamentals of Laparoscopic Skills (FLS) manual skills test, because these two methods have both been studied extensively and are used in clinical practice.

All studies were rated according to the Oxford Centre for Evidence-based Medicine levels of evidence[15] using the category for 'diagnostic studies', as validating studies are best compared with diagnostic studies. Results and evidence for each category are summarized in a separate table, and the most important findings are discussed in separate sections.

Validity, reliability and types of assessment

Validity is defined as 'the property of being true, correct and in conformity with reality'[16] and is subdivided into different levels: face validity, content validity, construct validity, concurrent validity and predictive validity. Face validity addresses users' opinion about the functionality and realism of a test. Content validity refers to whether the content of a test is suited to measure what it is supposed to measure. Construct validity refers to whether a test indeed measures the trait it is supposed to measure, in this case technical surgical skill. Discriminant validity is a variant of construct validity and requires a test to discriminate even more specifically, for instance between different experts. Concurrent validity is an expression of the comparison of a test to a standard or another test that measures the same trait. Predictive validity refers to the extent to which a test predicts future performance[16,17].

Reliability refers to whether a test is consistent in its outcome. Evidently, this also affects the validity of a test. Frequently used items for reliability are internal consistency, inter-rater reliability and intertest (test–retest) reliability. Internal consistency reflects the correlation between different items of a test and how these items contribute to the outcome of the test. Inter-rater reliability refers to the agreement of the scores of two or more raters testing the same subject. This is best tested with raters who are unaware of the subject's training level and identity (blinded raters). Intertest reliability refers to the agreement of scores when the same test is taken twice[17]. Reliability is represented by a reliability coefficient, which ranges from 0 to 1·0. Generally, 0·8 is accepted as a threshold for good reliability[18].

Finally, assessment can be either formative or summative. Formative assessment aims at development by monitoring a trainee's progress and giving structured feedback. When an assessment method is to be used for formative assessment, it should be able to identify different levels of performance (construct validity). A summative assessment is used for selection and therefore needs predefined levels of outcome. For instance, an examination can be passed or failed and there is a preset threshold that has to be reached. Summative assessment would be required for credentialling. Higher standards for construct validity and reliability are required with this form of assessment than with formative assessment. Moreover, clear cut-off values have to be defined adherent to the predefined consequences and, ideally, the sensitivity and specificity of these values should be tested.

Results

The search identified 931 unique studies, which were assessed for relevance. After this, 257 studies were selected, and their abstracts were assessed for inclusion criteria by two authors. Disagreement was solved by discussion. This left 104 studies for further analysis. Twenty-two were excluded after reading the full text and 22 new relevant articles were identified from reference lists. In total, 104

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

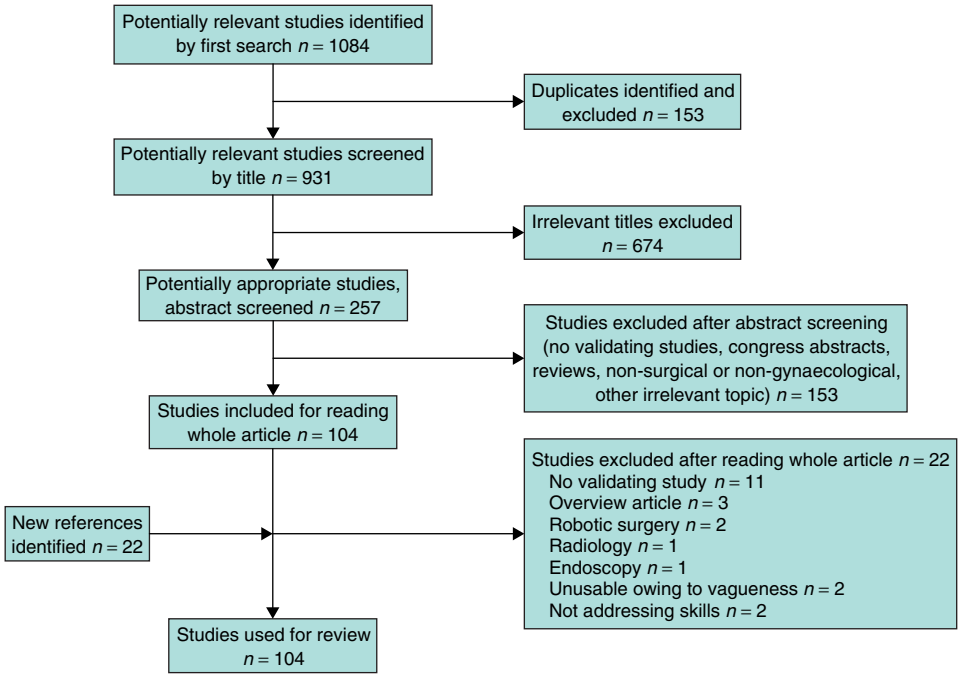

Fig. 1 Flow diagram for selection of articles

studies were left for review (*Fig. 1*). Twenty (19·2 per cent) of these studies offered level 1b or 2b evidence. In 28 studies (26·9 per cent) the assessment method was used in the operating room.

Procedure-specific checklists

Procedure-specific checklists are specifically designed for each procedure and usually follow subsequent steps of an intervention, which are scored. Nine studies concerned eight procedure-specific checklists[19–27]. Levels of evidence ranged from 2b to 4. Five checklists were used in the operating theatre[19–22,24,27], of which three were designed for laparoscopic cholecystectomy[19,21,22,24]. Four of these five checklists were used in combination with video registration (*Table 1*).

The only studies with a high level of evidence (2b) were by Sarker and colleagues[21] and Eubanks and co-workers[19]. Two checklists for assessment of laparoscopic cholecystectomy were designed by Sarker and both showed construct validity. Inter-rater reliability was above the cut-off value of 0·8, meaning that their reliability was good. Another checklist for laparoscopic cholecystectomy was designed by Eubanks *et al.*[19], which showed moderate correlation with experience and reasonable to good inter-rater reliability. The same checklist was studied by Aggarwal and co-workers[24], but with inferior results. All other studies had lower levels of evidence, owing to either non-consecutive cohorts (level 3b), which could imply bias in selection of participants, or unblinded raters (level 4).

Global rating scales

Global rating scales are used to rate more general skills; they are applicable to all surgical procedures and are thus not procedure specific. Eleven studies were identified concerning eight different global rating scales[7,28–37]. Two scales were studied in a laboratory setting and the other six in the operating theatre. Only two studies contained level 1b or 2b evidence[28,37]; all other studies consisted of level 4 evidence (*Table 2*).

Except for two scales[28,34], all were used to assess live operations. Every global rating scale was studied with different operations, except those of Sidhu and colleagues[34] and the Global Operative Assessment of Laparoscopic Skills (GOALS)[33,35–37]. These were respectively tested for laparoscopic colectomy in a porcine model, and laparoscopic cholecystectomy and appendicectomy in humans.

An investigation by Bramson and co-workers[28] was one of the two studies offering a high level of evidence. Development of a global rating scale for use with small

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Table 1 Specifications and most important results for included studies addressing procedure-specific checklists

Reference	Setting	Mode	Checklist	Level	*n*	Construct validity	Reliability	
							Internal consistency	Inter-rater reliability
19	OR	Video	Checklist score and error score for LC	2b	30	0·50*	NA	0·74–0·96
24	OR	Video	Checklist score and error score for LC	3b	47	No	NA	0·58
21	OR	Video and live†	Technical and technological skills for LC	2b	100	Yes/No‡	NA	> 0·8
22	OR	Video	Generic and specific skills for LC	3b	50	Yes	NA	> 0·8
24	OR	Video	Generic and specific skills for LC	3b	47	No	NA	0·62
23	Lab	Video	Checklist for ten stations	3b	21	Yes§	NA	0·78
25	Lab	Video	Checklist for intracorporeal suturing	3b	26	Yes	NA	0·90
26	Lab	Video	Rating for low anterior resection and Nissen fundoplication on a pig	3b	29	No	NA	0·73
27	OR	Video	Checklist for tubal ligation	4	23	No	NA	0·007–0·88
20	OR	Live	Rating different key procedures	4	300	Yes	> 0·90	NA

*Coefficient for correlation of checklist score with experience. †Technical skills were assessed from recorded video; technological operating room (OR) equipment skills were assessed live. ‡Technical skills were divided into generic and specific; construct validity was not established for the specific technical skills checklist. §Construct validity for six of ten workstations. LC, laparoscopic cholecystectomy; NA, not addressed; Lab, laboratory.

Table 2 Specifications and most important results for included studies addressing global rating scales

Reference	Setting	Mode	Global rating scale	Level	*n*	Construct validity	Reliability	
							Internal consistency	Inter-rater reliability
28	Lab	Live	Ten-item rating scale for basic skills	2b	65	$r = 0·86$*	0·84	0·83
29	OR	Live	General and case-specific skills scale	4	362	Yes	> 0·80	
7	OR	Live	Rating scale for five competencies	4	40	$r = 0·96$†		> 0·96
30	OR	Live	Rating scale for three competencies	4	4	NA		0·82
35	OR	Live	GOALS	4	21	Yes	0·91–0·94	0·82–0·89
33	OR	Live	GOALS	4	94	Yes		
36	OR	Live	GOALS	4	40	Yes		
37	OR	Video	GOALS	1b	2	Yes‡		0·87–0·93
31	OR	Live	Modification of OSATS and GOALS	4	7	Yes $r = 0·943$§	> 0·91	
34	Lab	Video	Modification of OSATS and GOALS	4	22	NA	> 0·88	0·76
32	OR	Live	Telephone rating scale	4	993	Yes		0·28

*Correlation with surgical skills. †Correlation with faculty ratings. ‡For four of five domains. §Correlation with postgraduate year. Lab, laboratory; OR, operating room; NA, not addressed; GOALS, Global Operative Assessment of Lasparoscopic Skills; OSATS, Objective Structured Assessment of Technical Skills.

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

tasks on animal tissue laboratory models was described; there was good correlation with surgical skills (estimated by questionnaire) and reliability was above 0·8.

The only global rating scale tested in multiple studies is GOALS. Four studies addressed use of this scale for laparoscopy[33,35–37]. It was developed by Vassiliou *et al.*[35], who applied it to the dissection phase of laparoscopic cholecystectomy. It appeared to be highly reliable and construct validity was established for all separate domains. Next, Gumbs and colleagues[33] investigated whether GOALS would also be applicable to total laparoscopic cholecystectomy, and other laparoscopic operations. Ratings of 94 residents performing laparoscopic cholecystectomy or laparoscopic appendicectomy were obtained and construct validity was established. In another study, by McCluney and co-workers[36], predictive and concurrent validity were established by comparing the GOALS score with the score on the FLS simulator (correlation 0·77). Finally, Chang and colleagues[37] studied ten blinded observers who rated videos of a novice and an expert performing laparoscopic cholecystectomy. Construct validity was established for four of five domains and high inter-rater reliability was found with level of evidence 1b; however, only two videotapes were rated. GOALS appeared reliable and might be used for video assessment. Although all four studies showed consistently good results for GOALS, three provided only level 4 evidence, because the raters were not blinded.

Objective Structured Assessment of Technical Skills

OSATS was one of the first methods designed for objective skills assessment. It is also the instrument that has been studied most extensively and is one of the few actually used in clinical practice. It consists of a global rating scale and a procedure-specific checklist. Originally, it was designed for use in laboratory settings, but it is now also used in the operating theatre.

Twenty-six studies were identified that addressed OSATS[26,38–62]. Nineteen covered OSATS in the laboratory setting[26,38,41–47,49–54,57–59,61] and seven in the operating theatre[39,40,48,55,56,60,62] (*Table 3*). In total, construct validity was established in 18 studies, internal consistency was above 0·8 in 12 studies and inter-rater reliability was above 0·8 in ten studies (*Table 3*). For use in a laboratory setting, four studies had a level of evidence 1b or 2b[42,47,53,54]. These studies showed construct validity, high internal consistency and variable inter-rater reliability for OSATS used with gynaecology bench tasks. Other studies had a level of evidence 3b or 4, but showed similar results. In the operating theatre, there was no high-level evidence as all seven studies provided only level 3b or 4 evidence. These seven studies showed construct validity and sporadically addressed reliability, which was above 0·8 in only one study[48].

Motion analysis

Motion analysis uses parameters that are extracted from motion of the hands or laparoscopic instruments. Nineteen studies were identified concerning this method of assessment[25,56–59,63–76]. These studies addressed five different instruments: the Imperial College Surgical Assessment Device (ICSAD; Imperial College, London, UK), the Advanced Dundee Psychomotor Tester (ADEPT; University of Dundee, Dundee, UK), the ProMIS™ Augmented Reality Simulator (Haptica, Dublin, Ireland), the Hiroshima University Endoscopic Surgical Assessment Device (HUESAD; Hiroshima University, Hiroshima, Japan) and the TrEndo Tracking System (Delft University of Technology, Delft, The Netherlands) (*Table 4*).

Nine studies addressed the ICSAD[25,56,58,59,63–65,73,75]. Construct validity was established, mostly for the parameters time and number of movements. Only Aggarwal and colleagues[56] used the ICSAD in the operating theatre. Intertest reliability was not found to be high and only moderate correlation existed with OSATS, which was considered the standard for objective assessment. In a study by Datta and co-workers[59], the same correlation with OSATS was found, although it was used in a laboratory and not in the operating room. Levels of evidence of all studies were consistently level 3b.

The ADEPT showed construct validity for one of three parameters in a level 1b study by Francis *et al.*[68]. Two other studies addressed correlation with clinical assessment and reliability, but these had a lower level of evidence and fewer participants[67,69].

ProMIS™ is a hybrid simulator, which combines a live and virtual environment. Tasks on this simulator are done in a box trainer, but a virtual interface is placed over the image of the camera in the box trainer. Two other cameras are used for motion tracking of the instruments. In a level 2b study by Van Sickle and colleagues[72], construct validity was established and internal consistency was 0·95. However, this study used only ten participants. Other studies used more participants, but had lower levels of evidence (*Table 4*).

The HUESAD was developed to analyse movements in vertical and horizontal planes. In a study by Egi and co-workers[66] construct validity was established, comparing novices and experts. However, this was the only study about the HUESAD, and had only level 3b evidence.

The TrEndo Tracking System was designed for motion analysis, to be used in a box trainer. In a study by

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Table 3 Specifications and most important results for included studies addressing Objective Structured Assessment of Technical Skills

						Construct validity		Reliability			
								Internal consistency		Inter-rater reliability	
Reference	Setting	Mode	Task/procedure	Level	*n*	GRS	Checklist	GRS	Checklist	GRS	Checklist
53	Lab	Video	Episiotomy repair	1b/2b*	40	Yes	Yes	NA	NA	0·59	0·80
50	Lab	Live	Episiotomy repair	4	18	Yes	Yes	0·95	0·95	NA	NA
42	Lab	Live	Cystoscopy and colposuspension	1b/2b*	55	Yes	Yes	0·93†	0·72†	0·69†	0·68†
								0·92‡	0·85‡	0·84‡	0·92‡
54	Lab	Live	Hysteroscopy	1b/2b*	48	Yes	Yes	0·98	0·79	0·42–0·93	0·75
47	VRS	Live	Myoma resection	2b	13	Yes	No	0·98	0·70	0·88§	0·78§
45	Lab	Live	Basic skills for gynaecology trainees	4	24	Yes, $r > 0·8$¶	Yes, $r > 0·8$¶	0·89	> 0·8	0·87	> 0·8
44	Lab	Live	Basic skills for gynaecology trainees	4	24	Yes, $r > 0·8$¶	Yes, $r > 0·8$¶	0·94	0·77	0·91	0·92
46	Lab	Live	Basic skills for gynaecology trainees	3b	16	Yes	Yes	0·96	0·96	0·95§	> 0·8§
43	Lab	Live	Basic skills for gynaecology trainees	3b	116	Yes	Yes	0·97	0·95	0·95§	0·88§
49	Lab	Live	Basic surgical skills	3b	20	Yes	Yes	< 0·8	< 0·8	< 0·8	< 0·8
51	Lab	Live	Basic surgical skills	3b	48	Yes	Yes	0·84	0·78	NA	NA
38	Lab	Live	Basic surgical skills	4	77	Yes	Yes	0·82	< 0·8	NA	NA
41	Lab	Live	Basic surgical skills	3b	6		$r < 0·8$#	NA	NA	NA	NA
61	Lab	Live	Basic surgical skills	3b	53	0·70**	0·58**	0·85	0·79	NA	NA
52	Lab	Video	Small bowel anastomosis	3b	40	NA	NA	NA	NA	> 0·8	NA
59	Lab	Video	Vascular anastomosis/patch	3b	50	Yes, $r = 0·59$††	No	NA	NA	0·89	0·81
58	Lab	Video	Vascular anastomosis/patch	3b	23	Yes	Yes	NA	NA		
26	Lab	Video	LAR and Nissen	3b	29	No	NA	NA	NA	0·72	NA
57	ARS	Video	Three tasks on ProMIS™	3b	20	$r = 0·77$‡‡	$r = 0·77–0·81$‡‡	NA	NA	0·93	0·88
40	OR	Live	Different gynaecology procedures	3b	119	Yes§§	Yes§§	NA	NA	NA	NA
62	OR	Live	LC	3b	22	Yes	NA	NA	NA	0·28	NA
60	OR	Video	LC	3b	22	No	NA	NA	NA	0·57	NA
56	OR	Video	LC	3b	47	Yes	NA	0·72¶¶	NA	0·72	NA
48	OR	Video	Salpingectomy	3b	21	Yes	Yes	NA	NA	0·83	0·83
55	OR	Live	Three different operations	4	41	Yes $r = 0·89$¶	Yes $r = 0·89$¶	NA	NA	0·73	0·78
39	OR	Live	Carotid endarterectomy	4	28	$r = 0·69–0·82$‡‡ Yes##	$r = 0·21–0·83$‡‡ Yes##	NA	NA	NA	NA

*1b for global rating scale (GRS), 2b for checklist. †For cystoscopy. ‡For colposuspension. §Correlation between blinded and unblinded raters. ¶Concurrent validity: correlation of GRS with checklist. #Concurrent validity: correlation of Objective Structured Assessment of Technical Skills (OSATS) score with faculty ratings. **Regression coefficient for relation with postgraduate year. ††Concurrent validity: correlation of OSATS GRS score with Imperial College Surgical Assessment Device motion analysis. ‡‡Construct validity: correlation with experience. §§Face validity. ¶¶Intertest reliability. ##Discriminant validity. Lab, laboratory; NA, not addressed; VRS, virtual reality simulator; LAR, low anterior resection; ARS, augmented reality simulator; OR, operating room; LC, laparoscopic cholecystectomy.

Chmarra and colleagues[76], participants were classified as novice, intermediate or expert by analysis of six motion analysis parameters (time, depth perception, path length, motion smoothness, angular area and volume). Data from these six parameters were first compressed using principal component analysis and subsequently classified using linear discriminant analysis. In this way, 23 of 31 participants were classified correctly.

Virtual reality simulators

VR simulators are especially valuable for learning endoscopic motor skills. As several parameters of performance are measured, VR simulators may also be used to assess skills. Twenty-six studies addressed this aspect of VR simulators[77–102]. Levels of evidence ranged from 1b to 4 (*Table 5*).

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Table 4 Specifications and most important results for included studies addressing motion analysis

Reference	Setting	Level	*n*	Construct validity			Other validity	Reliability
ICSAD				Time	No. of movements	Path length		
56	OR	3b	47	Yes*	Yes*	Yes*	< 0·63†	0·63‡
65	Lab	3b	51	0·66§	0·76§	No		
59	Lab	3b	50	Yes	Yes	NA	< 0·60†	
75	Lab	3b	30	NA	NA	NA	0·49–0·65¶	
64	Lab	3b	30	Yes	Yes	NA		
58	Lab	3b	23	Yes	Yes	NA		
63	Lab	3b	30	Yes	Yes	NA		
73	Lab	3b	15	Yes	Yes	Yes	0·78#	
25	Lab	3b	26	Yes	NA	Yes		
ADEPT				Time	Error score	Task score		
69	Device	3b	10	NA	NA	NA	0·789**	
67	Device	3b	20	NA	NA	NA		0·6‡; 0.98††
68	Device	1b	40	No	Yes	No		
ProMIS™				Time	Smoothness	Path length		
72	Device	2b	10	Yes	Yes	Yes		0·95††
70	Device	3b	30	0·78§	0·75§	0·67§	0·18–0·88‡‡	
71	Device	3b	160	Yes	Yes	Yes		
57	Device	3b	20	0·61–0·81§	0·36–0·98§	NA		
74	Device	3b	46	0·07–0·60§	0·11–0·59§	0·00–0·39§		
HUESAD								
66	Lab	3b	37	Yes§§			NA	NA
TrEndo								
76	Lab	3b	31	Yes¶¶			NA	NA

*Only for dissection part of laparoscopic cholecystectomy. †Concurrent validity: correlation with Objective Structured Assessment of Technical Skills (OSATS). ‡Intertest reliability. §Correlation with experience. ¶Correlation of 'surgical efficiency score', based on Imperial College Surgical Assessment Device (ICSAD) motion analysis, with OSATS. #Correlation of path length with procedure-specific checklist. **Concurrent validity: correlation of overall performance on Advanced Dundee Pschomotor Tester (ADEPT) with clinical assessment. ††Internal consistency. ‡‡Correlation with global rating scale. §§Construct validity determined for 'time' and deviation from horizontal and vertical planes. ¶¶Classified 74 per cent of participants correctly using linear discriminant analysis of motion analysis parameters. OR, operating room; Lab, laboratory; NA, not addressed; HUESAD, Hiroshima University Endoscopic Surgical Assessment Device.

Studies on five different simulators were identified: Minimally Invasive Surgical Trainer Virtual Reality (MIST™ VR; Mentice, Göthenburg, Sweden), LapSim (Surgical Science, Göthenburg, Sweden), LAP Mentor™ (Simbionix Corporation, Cleveland, Ohio, USA), Xitact® LS 500 (Mentice, Göthenburg, Sweden) and Simulator for Endoscopy SIMENDO® (DelltaTech, Rotterdam, The Netherlands). These simulators all provided tasks to train basic surgical skills for general surgery, gynaecology or laparoscopy in general. For assessment, most simulators used simple motion analysis parameters, such as path length or economy of motion, and all used time to task completion. Some used a composite score, either a simple sum or predetermined by the manufacturer, and different for every task, whereas others used error scores.

Most studies showed good results for all five simulators. However, they mainly offered only level 3b evidence and so the results should be interpreted with caution. Studies with higher levels of evidence exist for MIST™ VR, LAP Mentor™ and LapSim.

Two studies by Gallagher *et al.*[81,82] with level 1b evidence and one level 2b study by Taffinder and colleagues[94] established construct validity for MIST™ VR parameters. Furthermore, a study by Aggarwal and co-workers[103] used proficiency scores. Another study found comparable results, but had a lower level of evidence[89]. In a study by Cope and Fenton-Lee[78], on the other hand, construct validity could not be established for any parameter, and in studies by Grantcharov and colleagues[83] and Madan and co-workers[87] poor correlation with performance on a pig was found (concurrent validity). These studies also provided level 3b evidence.

For LAP Mentor™, a study by Zhang *et al.*[99] comprised level 1b evidence, and showed construct validity for time and composite score when comparing novices and residents. In a level 3b study by Aggarwal and colleagues[101] construct validity was established for most tasks, by different parameters. Moreover, cut-off values were defined in this study. In other studies results for LAP Mentor™ were less consistent (*Table 5*).

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Table 5 Specifications and most important results for included studies addressing virtual reality simulators

Reference	Level	*n*	Construct validity				
MIST™ VR			Time	Economy of movement	Economy of diathermy	Errors	Composite score
81	1b	36	Yes	Yes	No	Yes*	NA
82	1b	36	Yes	Yes	Yes	Yes	NA
94	2b	30	Yes	Yes	NA	Yes	NA
78	3b	22	No	No	NA	No	NA
83	3b	14	NA	NA	NA	0·5–0·7†	0·05–0·80†
89	3b	8	Yes	Yes	NA	Yes	NA
87	3b	32	< 0·56†	< 0·56†	NA	NA	0·21–0·56†
LapSim			Time	Path length	Angular path	Errors	Composite score
80	3b	24	Yes	Yes	Yes	Yes	NA
85	3b	115	Yes	Yes	Yes	Yes	NA
86	3b	32	Yes	Yes	Yes	Yes‡	Yes
79	3b	54	Yes	Yes§	NA	Yes‡	Yes
93	3b	24	Yes¶	Yes#	NA	NA	NA
84	3b	10	0·74**	0·69–0·98**	NA	0·67–0·89**	NA
97	3b	34	Yes	Yes	Yes	Yes	NA
100	4	47	< 0·51††	NA	NA	0·01–0·42††	NA
102	1b	40	Yes	Yes	NA	No	NA
LAP Mentor™			Time	No. of movements	Economy of movement	Speed	Composite score
99	1b	27	Yes	NA	NA	NA	Yes
98	3b	31	Yes	Yes‡‡	Yes‡‡	Yes‡‡	NA
88	3b	103	NA	NA	NA	NA	Yes§§
77	3b	27	Yes¶¶	NA	NA	NA	Yes¶¶
101	3b	57	Yes##	Yes‡	Yes‡	Yes***	NA
Xitact® LS 500			Time	Path length	Economy of movement	Speed	Composite score
91	3b	20	No	Yes	NA	No	NA
90	3b	307	Yes	Yes†††	NA	NA	NA
92	3b	74	Yes‡‡‡	NA	NA	NA	Yes
SIMENDO®			Time	Path length	Errors		
96	3b	25	Yes	Yes†††	Yes		
95	3b	61	Yes	Yes	No		

*Only significant difference between experts and intermediates. †Concurrent validity: correlation with performance on pig. ‡Only for one task. §Only for two of five tasks. ¶Construct validity for 'summary measure' time error. #Construct validity for 'summary measure' motion economy, with two of three tasks. **Predictive validity: correlation with performance in the operating room. ††Concurrent validity: correlation with performance on box trainer. ‡‡Only for non-dominant hand. §§Only for one of eight tasks. ¶¶Only for two of six tasks. ##For six of nine skills. ***For four of nine skills. †††Only for right instrument. ‡‡‡Also concurrent validity: subject with maximum score on pelvitrainer had significantly shorter task time on Xitact®. NA, not addressed.

LapSim has been studied extensively and most studies showed construct validity. One study, by Aggarwal *et al.*[102], provided level 1b evidence and showed construct validity for time and path length for all exercises. Cut-off values were defined in this study, and also in an investigation by Sherman and co-workers[93].

Video assessment

A separate category was defined for video assessment. With video assessment, a task, performance or operation is videotaped and rated later, which adds to its flexibility. The methods for assessment are the same as in live settings, but the fact that the performance is videotaped may have a considerable impact on the outcome of the assessment. For example, often only the laparoscopic camera shot is taped and not the whole operating theatre, which may blind the observer to certain aspects of the operation. Five studies were identified that explicitly addressed use of video registration on the outcome of assessment (*Table 6*)[60,75,104–106].

Studies by Beard[106] and Driscoll and co-workers[105] established construct validity for video assessment, with level 1b evidence in one study[106]. However, in this study only two videos of two subjects with a large difference in performance level (inexperienced *versus* experienced) were scored by different groups of raters. In another study by Beard and co-workers[104], a good correlation was found between video and live assessment, although that study offered level 3b evidence.

Editing of videotapes alters the assessment. In the Beard and Driscoll studies, raters were permitted to fast forward

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Table 6 Specifications and most important results for included studies addressing effect of videotaping

Reference	Setting	Procedure	Level	*n*	Construct validity		Correlation between video and live	Reliability			
								Internal consistency		Inter-rater reliability	
					Video	Live		Video	Live	Video	Live
60	OR	Laparoscopic cholecystectomy*	3b	22	No	Yes	< 0·33	NA	NA	0·28	0·57
104	OR	Saphenofemoral disconnection†	3b	33	NA	NA	083–0·92	NA	NA	NA	0·91‡
105	OR	Inguinal hernia repair†§	3b/4¶	9	Yes	No	NA	> 0·76	> 0·85	> 0·69	NA
75	Lab	Vascular and bowel anastomosis#	3b	30	< 0·37**	NA	NA	NA	NA	0·59–0·80	NA
106	OR	Saphenofemoral disconnection†	1b	2††	Yes	NA	NA	NA	NA	NA	NA

*Edited videotapes: length 10 min. †Fast forwarding of videotape was permitted. ‡Intertest reliability. §Edited videotapes: only essential steps shown. ¶Videotapes were blinded (level 3b), real-time assessment was not (level 4). #Edited videotapes: length 2 min. **Concurrent validity: correlation of full-length video score with snapshot video score. ††Two videotapes were shown to 14 surgeons, 14 trainees and 13 operating room (OR) nurses. NA, not addressed; Lab, laboratory.

the tape at their own discretion. Scott and colleagues[60] and Datta *et al.*[75] assessed the effect of editing videotapes before rating them. In the former study the videotapes were shortened to 10 min, showing only the essential parts. A poor correlation with live assessment was found[60]. In the investigation by Datta and colleagues[75], a 2-min snapshot tape of a task was recorded and the rating was compared with that for a full-length videotape. The results were similarly poor.

Miscellaneous

Nine studies did not fully fit any other category[75,107–114]. Levels of evidence of these studies ranged from 2b to 4 and concerned six different methods of assessment (*Table 7*).

One method of specific interest is outcome measurement, as it is often applied in clinical practice. With this method, numbers of complications and deaths are kept in logbooks or portfolios. Haddad and colleagues[107] compared complications between junior and senior surgeons in a study of 691 procedures, which were classified with respect to extent of operation. A difference was found for moderately extensive operations, for which more complications were attributed to senior surgeons. However, this difference was considered to result from allocation of more difficult cases to more senior surgeons, leading to bias. Therefore, patient outcome was not considered to be a useful method of assessment.

Another interesting method is the use of (hidden) Markov modelling. This is a mathematical way of compressing large amounts of data and producing one single measure to indicate a subject's distance from an ideal learning curve. With the studies shown in *Table 8*, this method was used to compress motion data

Table 7 Specifications and most important results for all included studies addressing miscellaneous methods of assessment

Reference	Setting	Mode	Method	Level	*n*	Validity	Reliability	
							Internal consistency	Inter-rater reliability
110	Lab	Live	Force/torque metrics	2b	4	Yes*	NA	NA
111	Lab	Live	(Hidden) Markov modelling	2b	8	Yes*	NA	NA
109	Lab	Live	(Hidden) Markov modelling	2b	10	Yes*	NA	NA
108	Lab	Live	(Hidden) Markov modelling	3b	11	0·93†	NA	NA
75	Lab	Video	Quality of final product	3b	30	0·34–0·55†	NA	0·80–0·84
113	Lab	Video	Time, errors and needle manipulations for suturing	3b	32	Yes*‡	NA	0·86–0·91§
107	OR	Data¶	Patient outcome	4	691	No	NA	NA
114	OR	Live	Patient outcome	4	29	Yes#	NA	NA
112	Lab	Video	Error scoring	4	60	> 0·8**	NA	NA

*Construct validity. †Concurrent validity: correlation with Objective Structured Assessment of Technical Skills. ‡Good for time, limited for errors and needle manipulations. §Tapes were rescored until reliability was above 0·80. ¶Prospective data collection. #Leakage in laboratory (Lab) task for vascular anastomosis predicted leakage in operating room (OR) and time in OR. **Concurrent validity: correlation with Objective Structured Clinical Examination for performance in simulated laparoscopic cholecystectomy. NA, not addressed.

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Table 8 Specifications and most important results for all included studies addressing Fundamentals of Laparoscopic Skills (FLS) manual skills test

			Validity				
Reference	Level	*n*	Construct	Concurrent	Reliability	Sensitivity (%)	Specificity (%)
115	1b	50	NA	0·51*	NA	NA	NA
116	3b	12	NA	NA	0·77–0·86† 0·98–1·00‡ 0·37–0·89§	NA	NA
117	4	42	0·26–0·69¶	NA	NA	NA	NA
118	4	10	Yes#, 0·82¶	NA	NA	NA	NA
119	4	165	Yes	NA	NA	82**	82**
120	4	12	NA	0·15–0·76††	NA	NA	NA
121	4	215	Yes	0·81††	NA	NA	NA
122	4	58	Yes	NA	NA	NA	NA
36	4	40	Yes	0·77‡‡	NA	91§§	86§§

*Correlation with In-Training Evaluation Reports. †Internal consistency. ‡Inter-rater reliability. §Intertest reliability. ¶Correlation of total score with postgraduate year. #For two of three tasks. **For a total score cut-off of 270. ††Correlation with performance *in vivo*. ‡‡Correlation with Global Operative Assessment of Laparoscopic Skills score. §§For a mean score cut-off of 70. NA, not addressed.

and force/torque data. Three studies, two of which provided level 2b evidence, established construct and concurrent validity[108,109,111]. However, limited numbers of participants were used; larger studies are needed to provide more solid evidence to show whether this method can truly distinguish between individuals with different performance.

Fundamentals of Laparoscopic Skills

Assessment of the FLS manual skills test is based on the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS). It consists of five tasks that are rated by two metrics: 'time to complete the task' and 'accuracy', calculated by predetermined penalties. As FLS consists of this fixed set of box trainer tasks, which is used for assessment and is not a specific rating method, it does not actually meet the inclusion criteria for this review (see Methods). However, FLS is the official manual skills test for surgical residents in the USA and is also used in other countries. Therefore, it is considered important to include.

Nine studies were identified concerning MISTELS (*Table 8*)[36,115–122]. Construct validity was established in six[36,117–119,121,122], of which four[36,119,121,122] found highly significant differences between subjects with different training levels, and two[117,118] found correlations with training levels varying from poor to good. Two large studies were performed by Fraser and colleagues[119] and Fried and co-workers[121], with 165 and 215 subjects respectively. Unfortunately, these studies provided only level 4 evidence because raters were not blinded to the training level of the participants.

Four studies addressed concurrent validity[36,115,120,121], comparing MISTELS with other assessment methods. One study had level 1b evidence and found a moderate correlation with In-Training Evaluation Reports[115].

Vassiliou and colleagues[116] studied the reliability of MISTELS. Good internal consistency and excellent inter-rater reliability were found, tested by comparing blinded *versus* unblinded raters. This was the only study that clearly stated the use of blinded raters; it had level 3b evidence as 12 non-consecutive participants were studied.

Finally, in two studies a cut-off value was calculated for the FLS score to use for certification[36,119]. Such a cut-off value is essential for summative assessment. Both studies offered level 4 evidence because the raters were not blinded.

Discussion

This study provided a critical overview and rating of current methods for objective assessment of technical surgical skills. Methods for objective assessment are needed for assessing trainees' performance, and that of practising surgeons. In an era with intense focus on training, and on quality and safety of surgery, these are important issues. Solid proof of validity and reliability of assessment methods is essential.

From all studies included in this review it can be concluded that OSATS is presently most accepted as the 'gold standard' for objective skills assessment. However, a high level of evidence for OSATS has been reached only for use with gynaecological bench tasks in a laboratory setting. Evidence for use in the operating theatre is of lower

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

grade and less abundant. Therefore, it remains unknown whether OSATS can distinguish between different levels of performance in the operating theatre. Studies by Martin and colleagues[49] and Beard and co-workers[104] were the only ones to correlate performance in bench tasks with performance in live animals (pigs) and the operating room, finding moderate correlations, not exceeding 0·8. Furthermore, cut-off values have not been defined for OSATS. These shortcomings should not prevent its use for feedback and discussion (formative assessment), but arguably prohibit the making of more important decisions based on OSATS. Presently OSATS should not be used for summative assessment in the operating theatre if these results are to be used for important examination decisions.

The same shortcomings affect other methods of assessment. Procedure-specific checklists and global rating scales have mostly been assessed in single studies. Only one procedure-specific checklist with a high level of evidence is available, and it is for laparoscopic cholecystectomy[21]. GOALS is the only global rating scale that has been assessed in multiple studies, but only one of these provided a high level of evidence[37]. Therefore, the evidence is limited for both the checklist and GOALS. Their use for formative assessment is an option, but using them for summative assessment is not recommended.

The use of motion analysis devices and VR simulators for skills assessment has been studied more extensively. For motion analysis, construct validity was established for the ICSAD and the ProMIS™ in multiple studies. These devices can differentiate between levels of performance and can be used for formative assessment. However, all these studies had a low level of evidence and the only study for the ProMIS™ that had level 2b evidence used ten participants[72]. Moreover, cut-off values for the scores have not been defined. Consequently, these devices should be used only for formative assessment.

VR simulators have all been tested in multiple studies and all seem able to distinguish in performance. However, high-level evidence is scarce and available only for MIST™ VR, LAP Mentor™ and LapSim. The results for MIST™ VR are consistent and, together with proficiency scores, form the evidence for its use for summative assessment. The results for LAP Mentor™ are not. The only study with a high level of evidence showed construct validity for two parameters. Cut-off values have been defined for this simulator and so summative assessment is possible, but not recommended. For LapSim, all studies, including one with level 1b evidence, showed construct validity and cut-off values have been defined. Therefore, the evidence is solid enough to use LapSim for summative assessment.

The results for a video assessment suggest that it is possible to use a videotape of a procedure to distinguish between individuals with a large difference in experience. There is no evidence that videotaping can distinguish between more subtle differences in performance level. Consequently, using videotapes for summative assessment is not yet possible. Moreover, editing, especially shortening videos, before assessment has a clear adverse impact on the outcome of assessment.

The FLS manual skills test was studied broadly and results were mostly good. However, only one study used blinded raters[116]. Therefore, all other studies had lower levels of evidence. The impact of blinding on scoring of the MISTELS tasks is questionable, because predefined errors cannot easily be scored in different ways and are therefore objective measures[117–119]. Likewise, the time to complete a task is an objective measure. In a study by Vassiliou and colleagues[116] a very high correlation was found between blinded and unblinded raters. It could be argued that the results of studies without blinding are unlikely to change if raters are blinded; these studies could be considered as having a higher level of evidence. Therefore, FLS seems well suited for formative assessment and, as cut-off values have also been defined, it could even be used for summative assessment.

Certain factors made it difficult to compare different methods of assessment. Much heterogeneity exists among the included studies, because existing methods have been adapted for use in other studies, and methodology and statistical evaluation differ. Different settings or tasks were used. Variation in the results of the studies could therefore be explained by these differences instead of differences in the assessment methods.

To determine the quality of the studies included in this review, the Oxford Centre for Evidence-based Medicine levels of evidence for diagnostic studies were used. This category provided the best fit with the design of validating studies. For this category, high levels of evidence are reached only when blinded reference standards (observer blinded to training level of the subject) are used and when cohorts of study objects comprise consecutive participants. Unblinded rating is less objective, which affects not only reliability but also validity, as the validity relies on the outcome of a test which, in turn, is influenced by reliability. Non-consecutive cohorts carry the risk of selection bias because certain eligible subjects might have been excluded deliberately. In many articles it was not clearly stated whether the cohorts were consecutive. These studies were assigned level of evidence 3b. This explains the large number of studies with this level of evidence. It could be argued that this approach was not fair. It is common for

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

all eligible subjects to be asked to participate voluntarily in this kind of study, which might result in a consecutive cohort. This fact was certainly considered but, as it was explicitly mentioned in a number of studies that all eligible subjects were included, it was considered appropriate to keep to the strict definition.

A threshold of 0·8 was chosen for good reliability. This choice was based on previous literature. However, this threshold is rigid and some assessments may not need to be that strict. While that may be true, this threshold was adhered to in order to distinguish methods that are sufficiently reliable for high-state assessment. Yet many studies showed reliability of assessment above this threshold, indicating that it was possible to meet the requirement.

Studies of assessment methods are hampered by the fact that it is difficult to set up a good study with participants who work in clinical practice. Time is always short and schedules not very flexible, which makes it hard to find large and consecutive groups of participants. Using blinded raters is an even bigger challenge. As they must be blinded to the assessed subject's training and clinical level, they have to originate from another hospital at least. Future studies should adhere to criteria that ensure true measures of validity and reliability, as mentioned above, and address the levels of evidence. One solution might be to embed future studies within residency programmes or training curricula. This will facilitate participation of all eligible subjects and use of blinded raters. Consequently, it will require cooperation of residency program directors.

Different methods of skills assessment are appropriate in different situations. Which method is the most appropriate depends on several aspects. The first issue to consider is the type of assessment and its consequences. To enable formative assessment, a method should give a reasonable impression of a subject's performance, as this form of assessment is used only as material for feedback and tracking progress over time. A negative outcome will not have direct consequences. On the contrary, significant consequences might result from summative assessment. Therefore, methods used for this form of assessment must be highly accurate and reliable, and should be used to test a subject's performance against predefined criteria. A crucial element for this is defining cut-off values for the scores that serve as these criteria. Furthermore, it should be clear what kind of performance is to be assessed, whether general skills or precise and systematic completion of a specific procedure. The latter requires a procedure-specific checklist, although this might not always give a proper representation of skills in general. In addition, it is important to realize that the value of a good assessment method can diminish when it is used in an inappropriate setting. An assessment method suitable for standard tasks in a laboratory setting may not be suitable for a real operation, where performance and outcome are influenced by many factors. Formulating clear answers to the issues discussed above is of great importance and should be adhered to when selecting a proper method for assessment of skills.

Acknowledgements

This study was funded by the national healthcare insurance company DSW (Schiedam, The Netherlands). The authors declare no conflict of interest.

References

1 Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993; **165**: 358–361.
2 Halsted WS. The training of the surgeon. *Bulletin of the Johns Hopkins Hospital* 1904; **15**: 267–275.
3 Darzi A, Smith S, Taffinder N. Assessing operative skill. Needs to become more objective. *BMJ* 1999; **318**: 887–888.
4 *Risico's minimaal invasieve chirurgie onderschat*. http://www.igz.nl/zoeken/document.aspx?doc=Bijlage_rapport_risico's_minimaalinvasieve_chirurgie_onderschat&URL [accessed 14 April 2010].
5 Darzi A, Datta V, Mackay S. The challenge of objective assessment of surgical skill. *Am J Surg* 2001; **181**: 484–486.
6 Kohn LT, Coorigan JM, Donaldson MS (eds). *To Err is Human: Building a Safer Health System*. National Academy Press: Washington, DC, 1999.
7 Kopta JA. Approach to evaluation of operative skills. *Surgery* 1971; **70**: 297–303.
8 Figert PL, Park AE, Witzke DB, Schwartz RW. Transfer of training in acquiring laparoscopic skills. *J Am Coll Surg* 2001; **193**: 533–537.
9 Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *BMJ* 2003; **327**: 1032–1037.
10 Stassen LP, Bemelman WA, Meijerink J. Risks of minimally invasive surgery underestimated: a report of the Dutch Health Care Inspectorate. *Surg Endosc* 2010; **24**: 495–498.
11 Aggarwal R, Moorthy K, Darzi A. Laparoscopic skills training and assessment. *Br J Surg* 2004; **91**: 1549–1558.
12 Fried GM, Feldman LS. Objective assessment of technical performance. *World J Surg* 2008; **32**: 156–160.
13 Grantcharov TP, Bardram L, Funch-Jensen P, Rosenberg J. Assessment of technical surgical skills. *Eur J Surg* 2002; **168**: 139–144.
14 Jaffer A, Bednarz B, Challacombe B, Sriprasad S. The assessment of surgical competency in the UK. *Int J Surg* 2009; **7**: 12–15.
15 Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, *et al*. *Levels of Evidence and Grades of*

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Recommendation. Oxford Centre for Evidence-based Medicine: Oxford, 2001.
16 Gallagher AG, Ritter EM, Satava RM. Fundamental principles of validation, and reliability: rigorous science for the assessment of surgical education and training. *Surg Endosc* 2003; **17**: 1525–1529.
17 Carter FJ, Schijven MP, Aggarwal R, Grantcharov T, Francis NK, Hanna GB *et al*.; Work Group for Evaluation and Implementation of Simulators and Skills Training Programmes. Consensus guidelines for validation of virtual reality surgical simulators. *Surg Endosc* 2005; **19**: 1523–1532.
18 Cooper C. *Individual Differences*. Arnold: London, 1998.
19 Eubanks TR, Clements RH, Pohl D, Williams N, Schaad DC, Horgan S *et al*. An objective scoring system for laparoscopic cholecystectomy. *J Am Coll Surg* 1999; **189**: 566–574.
20 Larson JL, Williams RG, Ketchum J, Boehler ML, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating surgery residents. *Surgery* 2005; **138**: 640–647.
21 Sarker SK, Chang A, Vincent C. Technical and technological skills assessment in laparoscopic surgery. *JSLS* 2006; **10**: 284–292.
22 Sarker SK, Chang A, Vincent C, Darzi SA. Development of assessing generic and specific technical skills in laparoscopic surgery. *Am J Surg* 2006; **191**: 238–244.
23 Swift SE, Carter JF. Institution and validation of an observed structured assessment of technical skills (OSATS) for obstetrics and gynecology residents and faculty. *Am J Obstet Gynecol* 2006; **195**: 617–621.
24 Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann Surg* 2008; **247**: 372–379.
25 Moorthy K, Munz Y, Dosis A, Bello F, Chang A, Darzi A. Bimodal assessment of laparoscopic suturing skills: construct and concurrent validity. *Surg Endosc* 2004; **18**: 1608–1612.
26 Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J *et al*. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc* 2004; **18**: 1800–1804.
27 Beckmann CR, Lipscomb GH, Ling FW, Beckmann CA, Johnson H, Barton L. Computer-assisted video evaluation of surgical skills. *Obstet Gynecol* 1995; **85**: 1039–1041.
28 Bramson R, Sadoski M, Sanders CW, van Walsum K, Wiprud R. A reliable and valid instrument to assess competency in basic surgical skills in second-year medical students. *South Med J* 2007; **100**: 985–990.
29 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.
30 Delaney PV, Hennessy TP, Quill RD, Kalizser M. Assessment of operative surgical skills. *Ir Med J* 1978; **71**: 438–441.
31 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.
32 Fung KFK, Fung MFK, Bordage G, Norman G. Interactive voice response to assess residents' laparoscopic skills: an instrument validation study. *Am J Obstet Gynecol* 2003; **189**: 674–678.
33 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.
34 Sidhu RS, Vikis E, Cheifetz R, Phang T. Self-assessment during a 2-day laparoscopic colectomy course: can surgeons judge how well they are learning new skills? *Am J Surg* 2006; **191**: 677–681.
35 Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondré K, Stanbridge D *et al*. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg* 2005; **190**: 107–113.
36 McCluney AL, Vassiliou MC, Kaneva PA, Cao J, Stanbridge DD, Feldman LS *et al*. FLS simulator performance predicts intraoperative laparoscopic skill. *Surg Endosc* 2007; **21**: 1991–1995.
37 Chang L, Hogle NJ, Moore BB, Graham MJ, Sinanan MN, Bailey R *et al*. Reliable assessment of laparoscopic performance in the operating room using videotape analysis. *Surg Innov* 2007; **14**: 122–126.
38 Ault G, Reznick R, MacRae H, Leadbetter W, DaRosa D, Joehl R *et al*. Exporting a technical skills evaluation technology to other sites. *Am J Surg* 2001; **182**: 254–256.
39 Beard JD, Choksy S, Khan S; Vascular Society of Great Britain and Ireland. Assessment of operative competence during carotid endarterectomy. *Br J Surg* 2007; **94**: 726–730.
40 Bodle JF, Kaufmann SJ, Bisson D, Nathanson B, Binney DM. Value and face validity of objective structured assessment of technical skills (OSATS) for work based assessment of surgical skills in obstetrics and gynaecology. *Med Teach* 2008; **30**: 212–216.
41 Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med* 1996; **71**: 1363–1365.
42 Fialkow M, Mandel L, VanBlaricom A, Chinn M, Lentz G, Goff B. A curriculum for Burch colposuspension and diagnostic cystoscopy evaluated by an objective structured assessment of technical skills. *Am J Obstet Gynecol* 2007; **197**: 544.e1–544.e6.
43 Goff B, Mandel L, Lentz G, Vanblaricom A, Oelschlager AM, Lee D *et al*. Assessment of resident surgical skills: is testing feasible? *Am J Obstet Gynecol* 2005; **192**: 1331–1338.
44 Goff BA, Lentz GM, Lee D, Fenner D, Morris J, Mandel LS. Development of a bench station objective

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

structured assessment of technical skills. *Obstet Gynecol* 2001; **98**: 412–416.
45 Goff BA, Lentz GM, Lee D, Houmard B, Mandel LS. Development of an objective structured assessment of technical skills for obstetric and gynecology residents. *Obstet Gynecol* 2000; **96**: 146–150.
46 Goff BA, Nielsen PE, Lentz GM, Chow GE, Chalmers RW, Fenner D *et al*. Surgical skills assessment: a blinded examination of obstetrics and gynecology residents. *Am J Obstet Gynecol* 2002; **186**: 613–617.
47 Goff BA, VanBlaricom A, Mandel L, Chinn M, Nielsen P. Comparison of objective, structured assessment of technical skills with a virtual reality hysteroscopy trainer and standard latex hysteroscopy model. *J Reprod Med* 2007; **52**: 407–412.
48 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.
49 Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C *et al*. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997; **84**: 273–278.
50 Nielsen PE, Foglia LM, Mandel LS, Chow GE. Objective structured assessment of technical skills for episiotomy repair. *Am J Obstet Gynecol* 2003; **189**: 1257–1260.
51 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.
52 Shah J, Munz Y, Manson J, Moorthy K, Darzi A. Objective assessment of small bowel anastomosis skill in trainee general surgeons and urologists. *World J Surg* 2006; **30**: 248–251.
53 Siddiqui NY, Stepp KJ, Lasch SJ, Mangel JM, Wu JM. Objective structured assessment of technical skills for repair of fourth-degree perineal lacerations. *Am J Obstet Gynecol* 2008; **199**: 676.e1–676.e6.
54 VanBlaricom AL, Goff BA, Chinn M, Icasiano MM, Nielsen P, Mandel L. A new curriculum for hysteroscopy training as demonstrated by an objective structured assessment of technical skills (OSATS). *Am J Obstet Gynecol* 2005; **193**: 1856–1865.
55 Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg* 1994; **167**: 423–427.
56 Aggarwal R, Grantcharov T, Moorthy K, Milland T, Papasavas P, Dosis A *et al*. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg* 2007; **245**: 992–999.
57 Broe D, Ridgway PF, Johnson S, Tierney S, Conlon KC. Construct validation of a novel hybrid surgical simulator. *Surg Endosc* 2006; **20**: 900–904.
58 Brydges R, Sidhu R, Park J, Dubrowski A. Construct validity of computer-assisted assessment: quantification of movement processes during a vascular anastomosis on a live porcine model. *Am J Surg* 2007; **193**: 523–529.
59 Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. *Am J Surg* 2002; **184**: 70–73.
60 Scott DJ, Rege RV, Bergen PC, Guo WA, Laycock R, Tesfay ST *et al*. Measuring operative performance after laparoscopic skills training: edited videotape *versus* direct observation. *J Laparoendosc Adv Surg Tech A* 2000; **10**: 183–190.
61 Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998; **73**: 993–997.
62 Scott DJ, Bergen PC, Rege RV, Laycock R, Tesfay ST, Valentine RJ *et al*. Laparoscopic training on bench models: better and more cost effective than operating room experience? *J Am Coll Surg* 2000; **191**: 272–283.
63 Bann SD, Khan MS, Darzi AW. Measurement of surgical dexterity using motion analysis of simple bench tasks. *World J Surg* 2003; **27**: 390–394.
64 Brydges R, Classen R, Larmer J, Xeroulis G, Dubrowski A. Computer-assisted assessment of one-handed knot tying skills performed within various contexts: a construct validity study. *Am J Surg* 2006; **192**: 109–113.
65 Datta V, Mackay S, Mandalia M, Darzi A. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *J Am Coll Surg* 2001; **193**: 479–485.
66 Egi H, Okajima M, Yoshimitsu M, Ikeda S, Miyata Y, Masugami H *et al*. Objective assessment of endoscopic surgical skills by analyzing direction-dependent dexterity using the Hiroshima University Endoscopic Surgical Assessment Device (HUESAD). *Surg Today* 2008; **38**: 705–710.
67 Francis NK, Hanna GB, Cuschieri A. Reliability of the Advanced Dundee Endoscopic Psychomotor Tester for bimanual tasks. *Arch Surg* 2001; **136**: 40–43.
68 Francis NK, Hanna GB, Cuschieri A. The performance of master surgeons on the Advanced Dundee Endoscopic Psychomotor Tester: contrast validity study. *Arch Surg* 2002; **137**: 841–844.
69 Macmillan AI, Cuschieri A. Assessment of innate ability and skills for endoscopic manipulations by the Advanced Dundee Endoscopic Psychomotor Tester: predictive and concurrent validity. *Am J Surg* 1999; **177**: 274–277.
70 Pellen M, Horgan L, Roger Barton J, Attwood S. Laparoscopic surgical skills assessment: can simulators replace experts? *World J Surg* 2009; **33**: 440–447.
71 Pellen MG, Horgan LF, Barton JR, Attwood SE. Construct validity of the ProMIS laparoscopic simulator. *Surg Endosc* 2009; **23**: 130–139.
72 Van Sickle KR, McClusky DA III, Gallagher AG, Smith CD. Construct validation of the ProMIS simulator using a novel laparoscopic suturing task. *Surg Endosc* 2005; **19**: 1227–1231.
73 Smith SG, Torkington J, Brown TJ, Taffinder NJ, Darzi A. Motion analysis. *Surg Endosc* 2002; **16**: 640–645.

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

74 Oostema JA, Abdel MP, Gould JC. Time-efficient laparoscopic skills assessment using an augmented-reality simulator. *Surg Endosc* 2008; **22**: 2621–2624.
75 Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. *Am J Surg* 2006; **192**: 372–378.
76 Chmarra MK, Klein S, de Winter JC, Jansen FW, Dankelman J. Objective classification of residents based on their psychomotor laparoscopic skills. *Surg Endosc* 2009; [Epub ahead of print].
77 Andreatta PB, Woodrum DT, Gauger PG, Minter RM. LapMentor metrics possess limited construct validity. *Simul Healthc* 2008; **3**: 16–25.
78 Cope DH, Fenton-Lee D. Assessment of laparoscopic psychomotor skills in interns using the MIST Virtual Reality Simulator: a prerequisite for those considering surgical training? *ANZ J Surg* 2008; **78**: 291–296.
79 Duffy AJ, Hogle NJ, McCarthy H, Lew JI, Egan A, Christos P *et al*. Construct validity for the LAPSIM laparoscopic surgical simulator. *Surg Endosc* 2005; **19**: 401–405.
80 Eriksen JR, Grantcharov T. Objective assessment of laparoscopic skills using a virtual reality stimulator. *Surg Endosc* 2005; **19**: 1216–1219.
81 Gallagher AG, Richie K, McClure N, McGuigan J. Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. *World J Surg* 2001; **25**: 1478–1483.
82 Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. Learning curves and reliability measures. *Surg Endosc* 2002; **16**: 1746–1752.
83 Grantcharov TP, Rosenberg J, Pahle E, Funch-Jensen P. Virtual reality computer simulation. *Surg Endosc* 2001; **15**: 242–244.
84 Kundhal PS, Grantcharov TP. Psychomotor performance measured in a virtual environment correlates with technical skills in the operating room. *Surg Endosc* 2009; **23**: 645–649.
85 Langelotz C, Kilian M, Paul C, Schwenk W. LapSim virtual reality laparoscopic simulator reflects clinical experience in German surgeons. *Langenbecks Arch Surg* 2005; **390**: 534–537.
86 Larsen CR, Grantcharov T, Aggarwal R, Tully A, Sørensen JL, Dalsgaard T *et al*. Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. *Surg Endosc* 2006; **20**: 1460–1466.
87 Madan AK, Frantzides CT, Sasso LM. Laparoscopic baseline ability assessment by virtual reality. *J Laparoendosc Adv Surg Tech A* 2005; **15**: 13–17.
88 McDougall EM, Corica FA, Boker JR, Sala LG, Stoliar G, Borin JF *et al*. Construct validity testing of a laparoscopic surgical simulator. *J Am Coll Surg* 2006; **202**: 779–787.
89 McNatt SS, Smith CD. A computer-based laparoscopic skills assessment device differentiates experienced from novice laparoscopic surgeons. *Surg Endosc* 2001; **15**: 1085–1089.
90 Rosenthal R, Gantert WA, Hamel C, Hahnloser D, Metzger J, Kocher T *et al*. Assessment of construct validity of a virtual reality laparoscopy simulator. *J Laparoendosc Adv Surg Tech A* 2007; **17**: 407–413.
91 Rosenthal R, Gantert WA, Scheidegger D, Oertli D. Can skills assessment on a virtual reality trainer predict a surgical trainee's talent in laparoscopic surgery? *Surg Endosc* 2006; **20**: 1286–1290.
92 Schijven M, Jakimowicz J. Construct validity: experts and novices performing on the Xitact LS500 laparoscopy simulator. *Surg Endosc* 2003; **17**: 803–810.
93 Sherman V, Feldman LS, Stanbridge D, Kazmi R, Fried GM. Assessing the learning curve for the acquisition of laparoscopic skills on a virtual reality simulator. *Surg Endosc* 2005; **19**: 678–682.
94 Taffinder N, Sutton C, Fishwick RJ, McManus IC, Darzi A. Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: results from randomised controlled studies using the MIST VR laparoscopic simulator. *Stud Health Technol Inform* 1998; **50**: 124–130.
95 Verdaasdonk EG, Stassen LP, Monteny LJ, Dankelman J. Validation of a new basic virtual reality simulator for training of basic endoscopic skills: the SIMENDO. *Surg Endosc* 2006; **20**: 511–518.
96 Verdaasdonk EG, Stassen LP, Schijven MP, Dankelman J. Construct validity and assessment of the learning curve for the SIMENDO endoscopic simulator. *Surg Endosc* 2007; **21**: 1406–1412.
97 Woodrum DT, Andreatta PB, Yellamanchilli RK, Feryus L, Gauger PG, Minter RM. Construct validity of the LapSim laparoscopic surgical simulator. *Am J Surg* 2006; **191**: 28–32.
98 Yamaguchi S, Konishi K, Yasunaga T, Yoshida D, Kinjo N, Kobayashi K *et al*. Construct validity for eye–hand coordination skill on a virtual reality laparoscopic surgical simulator. *Surg Endosc* 2007; **21**: 2253–2257.
99 Zhang A, Hünerbein M, Dai Y, Schlag PM, Beller S. Construct validity testing of a laparoscopic surgery simulator (Lap Mentor): evaluation of surgical skill with a virtual laparoscopic training simulator. *Surg Endosc* 2008; **22**: 1440–1444.
100 Newmark J, Dandolu V, Milner R, Grewal H, Harbison S, Hernandez E. Correlating virtual reality and box trainer tasks in the assessment of laparoscopic surgical skills. *Am J Obstet Gynecol* 2007; **197**: e1–e4.
101 Aggarwal R, Crochet P, Dias A, Misra A, Ziprin P, Darzi A. Development of a virtual reality training curriculum for laparoscopic cholecystectomy. *Br J Surg* 2009; **96**: 1086–1093.
102 Aggarwal R, Grantcharov TP, Eriksen JR, Blirup D, Kristiansen VB, Funch-Jensen P *et al*. An evidence-based virtual reality training program for novice laparoscopic surgeons. *Ann Surg* 2006; **244**: 310–314.
103 Aggarwal R, Grantcharov T, Moorthy K, Hance J, Darzi A. A competency-based virtual reality training curriculum for

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

the acquisition of laparoscopic psychomotor skill. *Am J Surg* 2006; **191**: 128–133.
104 Beard JD, Jolly BC, Newble DI, Thomas WE, Donnelly J, Southgate LJ. Assessing the technical skills of surgical trainees. *Br J Surg* 2005; **92**: 778–782.
105 Driscoll PJ, Paisley AM, Paterson-Brown S. Video assessment of basic surgical trainees' operative skills. *Am J Surg* 2008; **196**: 265–272.
106 Beard JD. Setting standards for the assessment of operative competence. *Eur J Vasc Endovasc Surg* 2005; **30**: 215–218.
107 Haddad M, Zelikovski A, Gutman H, Haddad E, Reiss R. Assessment of surgical residents' competence based on postoperative complications. *Int Surg* 1987; **72**: 230–232.
108 Leong JJ, Nicolaou M, Atallah L, Mylonas GP, Darzi AW, Yang GZ. HMM assessment of quality of movement trajectory in laparoscopic surgery. *Comput Aided Surg* 2007; **12**: 335–346.
109 Rosen J, Hannaford B, Richards CG, Sinanan MN. Markov modeling of minimally invasive surgery based on tool/tissue interaction and force/torque signatures for evaluating surgical skills. *IEEE Trans Biomed Eng* 2001; **48**: 579–591.
110 Rosen J, MacFarlane M, Richards C, Hannaford B, Sinanan M. Surgeon–tool force/torque signatures – evaluation of surgical skills in minimally invasive surgery. *Stud Health Technol Inform* 1999; **62**: 290–296.
111 Rosen J, Solazzo M, Hannaford B, Sinanan M. Objective laparoscopic skills assessments of surgical residents using Hidden Markov Models based on haptic information and tool/tissue interactions. *Stud Health Technol Inform* 2001; **81**: 417–423.
112 Tang B, Hanna GB, Carter F, Adamson GD, Martindale JP, Cuschieri A. Competence assessment of laparoscopic operative and cognitive skills: Objective Structured Clinical Examination (OSCE) or Observational Clinical Human Reliability Assessment (OCHRA). *World J Surg* 2006; **30**: 527–534.
113 Van Sickle KR, Baghai M, Huang IP, Goldenberg A, Smith CD, Ritter EM. Construct validity of an objective assessment method for laparoscopic intracorporeal suturing and knot tying. *Am J Surg* 2008; **196**: 74–80.
114 Wilasrusmee C, Lertsithichai P, Kittur DS. Vascular anastomosis model: relation between competency in a laboratory-based model and surgical competency. *Eur J Vasc Endovasc Surg* 2007; **34**: 405–410.
115 Feldman LS, Hagarty SE, Ghitulescu G, Stanbridge D, Fried GM. Relationship between objective assessment of technical skills and subjective in-training evaluations in surgical residents. *J Am Coll Surg* 2004; **198**: 105–110.
116 Vassiliou MC, Ghitulescu GA, Feldman LS, Stanbridge D, Leffondré K, Sigman HH *et al*. The MISTELS program to measure technical skill in laparoscopic surgery: evidence for reliability. *Surg Endosc* 2006; **20**: 744–747.
117 Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL. Development of a model for training and evaluation of laparoscopic skills. *Am J Surg* 1998; **175**: 482–487.
118 Derossis AM, Antoniuk M, Fried GM. Evaluation of laparoscopic skills: a 2-year follow-up during residency training. *Can J Surg* 1999; **42**: 293–296.
119 Fraser SA, Klassen DR, Feldman LS, Ghitulescu GA, Stanbridge D, Fried GM. Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. *Surg Endosc* 2003; **17**: 964–967.
120 Fried GM, Derossis AM, Bothwell J, Sigman HH. Comparison of laparoscopic performance *in vivo* with performance measured in a laparoscopic simulator. *Surg Endosc* 1999; **13**: 1077–1081.
121 Fried GM, Feldman LS, Vassiliou MC, Fraser SA, Stanbridge D, Ghitulescu G *et al*. Proving the value of simulation in laparoscopic surgery. *Ann Surg* 2004; **240**: 518–525.
122 Swanstrom LL, Fried GM, Hoffman KI, Soper NJ. Beta test results of a new system assessing competence in laparoscopic surgery. *J Am Coll Surg* 2006; **202**: 62–69.

Commentary

Objective assessment of technical surgical skills (*Br J Surg* 2010, 97: 972–987)

This review of the published literature seems thorough, although an entire edition of the *ANZ Journal of Surgery* was dedicated to the proceedings of the first International Conference on Surgical Education last year[1]. A systematic review requires a search for other evidence, as educational research remains poorly funded and often unpublished. A search of the websites of surgical training organizations and colleges, such as the UK Intercollegiate Surgical Curriculum Programme (http://www.iscp.ac.uk/), the American Board of Surgery (http://www.absurgery.org/) and the Australasian College of Surgeons (http://www.surgeons.org/), would have revealed a wealth of information about current workplace-based assessment methods for surgical trainees.

Such a search would have revealed additional methods such as direct observation of procedural skills (DOPS) and procedure-based assessment (PBA). DOPS is used by many specialties, as well as surgery, to assess trainees on simple

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

procedures and PBA is the main method used to assess the technical skills of specialist surgical trainees in the UK. Whereas Objective Structured Assessment of Technical Skills (OSATS) was originally developed for use in the skills laboratory, PBA has been specifically designed for use in the operating room[2].

Focusing on technical skills will not guarantee patient safety. Many surgical errors and poor outcomes are due to poor communication and teamwork[3]. Non-Technical Skills for Surgeons (NOTSS) has been designed to assess situational awareness, communication/teamworking, decision making and leadership in the operating room (http://www.abdn.ac.uk/iprc/notss).

J. D. Beard
Sheffield Vascular Institute, Northern General Hospital, Sheffield S5 7AU, UK
(e-mail: Jonathan.D.Beard@sth.nhs.uk)
DOI: 10.1002/bjs.7118

References

1 Proceedings of the first International Conference on Surgical Education and Training, Melbourne 2008. *ANZ J Surg* 2009; **79**: 95–216.
2 Beard JD. Assessment of surgical competence. *Br J Surg* 2007; **94**: 1315–1316.
3 Gawande AA, Zinner MJ, Studdert DM, Brennan TA. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery* 2003; **133**: 614–621.

Downloaded from https://academic.oup.com/bjs/article/97/7/972/6150566 by Maastricht University user on 29 November 2021

Copyright © 2010 British Journal of Surgery Society Ltd
Published by John Wiley & Sons Ltd