GENERAL SURGERY

Royal Australasian College of Surgeons

Validation and reliability testing of a rating scale for objective assessment of performance in laparoscopic appendicectomy surgery

Pramudith Sirimanna,* Praveen Ravindran,† Michelle Smigielski,‡ Marc A Gladman§ and Vasi Naganathan¶

*Department of Surgery, Sydney Medical School—Concord, University of Sydney, Sydney, New South Wales, Australia
†Australian National University and Australian Robotic Colorectal Surgery, Canberra, Australian Capital Territory, Australia
‡Department of Surgery, Liverpool Hospital, Sydney, New South Wales, Australia
§Adelaide Medical School, University of Adelaide, Adelaide, South Australia, Australia and
¶Centre for Education and Research on Ageing, Concord Hospital and University of Sydney, Sydney, New South Wales, Australia

Key words
competency-based training, laparoscopic appendicectomy, rating scales, surgical education, technical skills assessment.

Correspondence
Dr Pramudith Sirimanna, Departments of Surgery, Concord Repatriation General Hospital, Hospital Road, Concord, Sydney, NSW 2139, Australia.
Email: pramsirimanna@gmail.com

P. Sirimanna PhD, FRACS; **P. Ravindran** MBBS, FRACS; **M. Smigielski** MSSB, FRACS; **M. A Gladman** PhD, FRACS; **V. Naganathan** PhD, FRACP.

Accepted for publication 2 June 2022.

doi: 10.1111/ans.17862

Abstract

Backgrounds: To achieve a competency-based training paradigm, the ability to obtain reliable and valid quantitative assessments of intraoperative performance is required. Through this, weaknesses can be identified and practiced, and competency assessed. This study aimed to determine the validity and reliability an objective evaluation tool for assessment of performance in laparoscopic appendicectomy (LA).

Methods: A prospective single-blinded observational study design was used. Videos of inexperienced (performed <10 LAs) and experienced (performed >100 LAs) surgeons performing LA surgery were collected. Surgical performance during each recording was rated by two independent, blinded expert surgeons using the LA Rating Scale (LARS) and the modified Objective Structured Assessment of Technical Skill (OSATS) scale.

Results: The intraclass correlation coefficient (ICC) for LARS was 0.95 (95%CI 0.83–0.98). The ICC for each step ranged from 0.48 to 0.90, and the test–retest ICC for LARS was 0.91 (95%CI 0.69–0.98). Significant differences ($P < 0.001$) between median performance scores as rated by LARS were observed between the inexperienced and experienced surgeons. A Spearman's correlation coefficient of 0.87 ($P < 0.001$) was observed between LARS performance scores and modified OSATS scores.

Conclusion: LARS demonstrated excellent inter-rater and test–retest reliability, and construct and concurrent validity and can be used to quantitatively evaluate performance during LA. This can potentially allow specific weaknesses to be identified and improved upon through deliberate practice. Progress can be tracked through re-evaluation and scores of expert surgeons can be used as performance goals for credentialing in LA.

Introduction

The traditional philosophy of surgical training has been that of a 'time-based' apprenticeship model with no objective method of assessing competence at specific operations. However, given that learners acquire skills at different rates there may be a variation in the time it takes trainees to achieve competence. The revised General Surgery Education and Training (GSET) program that is due to commence in 2022 provides a step towards a competency-based training paradigm with the introduction of entrustable professional activities and procedure-based activities certifying independent practice. To effectively track a trainee's progress and identify those that require further training to reach competence, quantitative and objective measures of competence are needed. These could be used by training boards and trainees to ascertain when individuals are competent to independently perform specific operations safely.

A number of evaluation tools for surgical procedures have been developed over the last two decades starting with the Objective Structured Assessment of Technical Skills (OSATS) tool.[1] Being the most extensively investigated tool, its reliability, construct, concurrent, and predictive validity has been demonstrated.[2] An OSATS scale modified for laparoscopic surgery has also been developed

ANZ Journal of Surgery published by John Wiley & Sons Australia, Ltd on behalf of Royal Australasian College of Surgeons.

and validated.[3] While these tools are reliable and valid methods of assessing generic technical skills, they do not provide an assessment or feedback relating to which specific steps or techniques within an operation a trainee requires further practice. Consequently, procedure-specific evaluation tools have been developed.[4] Most of these instruments, however, are institution specific so their applicability to a national training program is limited. Palter *et al.* used a multi-institutional approach to develop procedure-specific evaluations tools for laparoscopic right hemicolectomy and sigmoid colectomy surgery.[5] These tools were subsequently demonstrated to have strong reliability and construct validity.[6]

The laparoscopic appendicectomy (LA) is the most common general surgery emergency procedure, and often performed by the most junior trainees. Indeed, the ability to perform a LA is a prerequisite for selection into the GSET program and deemed as a core procedure within the GSET program that must be completed independently with minimal supervision and guidance by the end of the third year of training. However, in its current format, certification of this from a supervising surgeon is required, rather than an objective quantitative demonstration of competence. We therefore developed the laparoscopic appendicectomy rating scale (LARS), which is a multi-institutional procedure-specific evaluation tool for LA surgery. The aims of this study were to determine the reliability and validity of LARS in objectively evaluating performance during actual LAs.

Methods

Study design

This study utilized a prospective single-blinded observational study design to evaluate the reliability and validity of LARS.

Participants

At a tertiary hospital in Sydney, New South Wales, patients with suspected appendicitis scheduled to undergo a LA were identified through daily review of the emergency operating list. Informed consent was initially obtained from patients to video record the LA they were scheduled to have. If the patient agreed to having the LA video recorded consent was obtained from the primary surgeon performing the LA. Ethical approval for this study was obtained from the Human Research Ethics Committee – Concord Hospital of the Sydney Local Health District.

The surgeons were categorized based on their operative experience at LA. Experienced surgeons were defined as those who were post-fellowship surgeons and had performed >100 LAs as primary operator to ensure they were well beyond the learning curve. Inexperienced surgeons were those who had performed <10 LAs as primary operator.

Assessment tools

Two assessment tools were used in this study. The first was LARS, which is a procedure-specific evaluation tool to assess performance during LA. It comprises of descriptors of 'poor', 'average' and 'excellent' performance for each step of a LA, scored on a Likert scale between 1 and 5. This tool was developed by obtaining multi-institutional expert consensus regarding the suitability of the steps and descriptors of performance for inclusion into the LARS. The second assessment tool used in this study was the modified OSATS global rating scale.

Video analysis

Videos of the LA were only commenced following intra-abdominal placement of the laparoscope and ceased at the end of the case. There was no audio recorded. Inexperienced surgeons were assisted by experienced surgeons and both surgeons were told to make no changes to their usual practice including the verbal guidance that the experienced surgeon would provide to the inexperience surgeon.

The recordings were reviewed by two trained independent surgeons to assess operative performance during each case using LARS and modified OSATS. These raters were post-fellowship surgeons, with an interest in surgical education, who had performed >100 LAs. The raters were blinded to the experience level of the surgeon. Both raters participated in a 30-minute orientation session with a study team member to familiarize themselves with the evaluation tools. Each rater was walked through the procedural steps included in the tool, as well as the descriptors of performance. They were asked to use the descriptors to evaluate performance and encouraged to use the full range of the 5-point Likert scale as appropriate. The raters were given further time to independently review the tool followed by an opportunity to ask questions. Viewing the videos at a faster playback speed was permitted if it did not interfere with evaluation of performance. For any steps within the rating scales that could not be visualized on the videos the raters were asked to mark the step as "not applicable".

Data analysis

The minimum number of videos needed in each group was determined based on a previous study that evaluated the Global Operative Assessment for Laparoscopic Skills (GOALS).[7] In this study, the minimum relevant difference in mean scores between novice and experience surgeons was 6.4.[7] Based on a standard deviation of 4.5, a power of 80% and an alpha of 0.05, the minimum number of videos required in each group was 8. All statistical analyses were performed using SPSS (statistical package for social sciences version 20.0, Chicago, IL, USA). Descriptive statistics were calculated for the LARS scores using medians and interquartile ranges.

Reliability assessment was deconstructed into inter-rater reliability and test–retest reliability. Inter-rater reliability was determined using the intraclass correlation coefficient (ICC, 2-way mixed-effects model, absolute agreement), which measures the internal consistency of the total scores and scores for each step as rated by the two blinded surgeons. A cut-off value of an ICC > 0.8 has been previously suggested to be a benchmark for demonstrating good reliability.[8] The scores of the two raters were also assessed for correlation using Spearman's correlation coefficient. To determine the test–retest reliability, a single recorded case was randomly selected for re-rating by both raters using LARS 6 months following the

Table 1 Inter-rater reliability coefficients between the two independent blinded raters for the Laparoscopic Appendicectomy Rating Scale (LARS) and its individual components

Component of LARS	N	ICC	95% CI	*P*-value
Total score	18	0.95	0.83–0.98	<0.001
Suction/Lavage any free fluid/pus	18	0.90	0.73–0.96	<0.001
Appropriately position patient to aid exposure	18	0.56	−0.17 to 0.83	0.05
Retract/Sweep the small bowel/omentum to aid exposure	18	0.84	0.56–0.94	<0.001
Identifies and exposes appendix using anatomical landmarks	18	0.69	0.21–0.89	0.007
Divide peritoneal/inflammatory adhesions of appendix ± caecum and/or proximal right colon as required to mobilize and locate the base of the appendix	18	0.83	0.53–0.94	<0.001
Retract appendix/mesoappendix to orientate and expose area for dissection	18	0.85	0.61–0.95	<0.001
Dissection of the mesoappendix	18	0.90	0.72–0.96	<0.001
Divide mesoappendix/appendicular artery if relevant (i.e., clips appendicular artery)	18	0.81	0.50–0.93	0.001
Assessment of the condition, and appropriate ligation and division of appendix at the base	17	0.83	0.52–0.94	0.001
Place the appendix into a bag	17	0.72	0.21–0.90	0.009
Deliver the appendix through the umbilical port	17	0.48	−0.53 to 0.81	0.114
Inspect operative bed for bleeding	17	0.87	0.63–0.95	<0.001

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient; *N*, number of cases analyses.

original assessment of the selected case. Utilizing the ICC, the total scores rated by both raters during the original assessment of the case was correlated to the total scores rated during the re-assessment.

The ability of LARS to detect differences in experience level between the groups (construct validity) was measured using Mann–Whitney U test before and after pooling the scores of the two raters for each step. Concurrent validity was assessed by correlating the LARS scores with the corresponding modified OSATS scores using Spearman's correlation coefficient. For the analyses of validity, each rater's scores as well as the combined LARS scores were expressed as a percentage of the total possible score. Similarly, the combined modified OSATS scores were expressed as a percentage of the total score. This was conducted in case of a failure to capture or an inability to analyse part of a procedure.

Results

Eighteen videos of LAs comprising of nine by inexperienced surgeons and nine by experienced surgeons were obtained. There were nine different experienced surgeons and eight different inexperienced surgeons (two separate procedures by one inexperienced surgeon). One of the videos of the inexperienced surgeons was unable to be analysed following the dissection of the mesoappendix and division of the appendicular artery as the video recording system failed after this point in the procedure. For all surgeons, the performance rating for the one of the steps ('Inspection of all four quadrants and inspect for other differentials [including small bowel run] if appendix macroscopically normal was either performed or not') was excluded from the analysis as the binary nature of this data had the potential to skew the results of the reliability results.

Reliability assessment

The overall ICC for LARS was 0.95 (95% CI 0.83–0.98; $P < 0.001$) and the scores as rated by the two independent blinded surgeons correlated with a coefficient of 0.91. The ICC for each individual step ranged from 0.48 and 0.90 (Table 1). For the inexperienced surgeons, the inter-rater agreement between the raters was 0.86 (95% CI 0.37–0.98, $P = 0.008$) and for experienced surgeons was 0.57 (95% CI 0.27–0.90, $P = 0.013$). The test–retest ICC for LARS was 0.81 (95% CI 0.39–0.95, $P < 0.004$) for rater 1, and 0.92 (95% CI 0.72–0.98, $P < 0.001$) for rater 2. When the two scores of both raters were combined the test–retest ICC was 0.91 (95% CI 0.69–0.98, $P < 0.001$). The overall ICC for the modified OSATS was 0.90 (95% CI 0.87–0.98; $P < 0.001$).

Table 2 Individual and combined LARS and OSATS rating scores for inexperienced and experienced surgeons

	Inexperienced surgeons, $N = 9$ (median, range)	Experienced surgeons, $N = 9$ (median, range)	*P*-value
Rater 1 LARS score (% of total possible score)	54.9% (54.7–64.6%)	83.1% (78.5–86.2%)	<0.001
Rater 2 LARS score (% of total possible score)	63.1% (53.6–63.8%)	89.2% (82.3–93.1%)	<0.001
Combined LARS score (% of total possible score)	58.5% (54.9–63.8%)	86.9% (80.4–88.8%)	<0.001
Rater 1 OSATS score (% of total possible score)	50.0% (45.0–55.0%)	85.0% (75.0–90.0%)	
Rater 2 OSATS score (% of total possible score)	50.0% (45.0–55.0%)	80.0% (77.5–85.0%)	
Combined OSATS score (% of total possible score)	47.5% (47.5–55.0%)	82.5% (78.75–86.25%)	

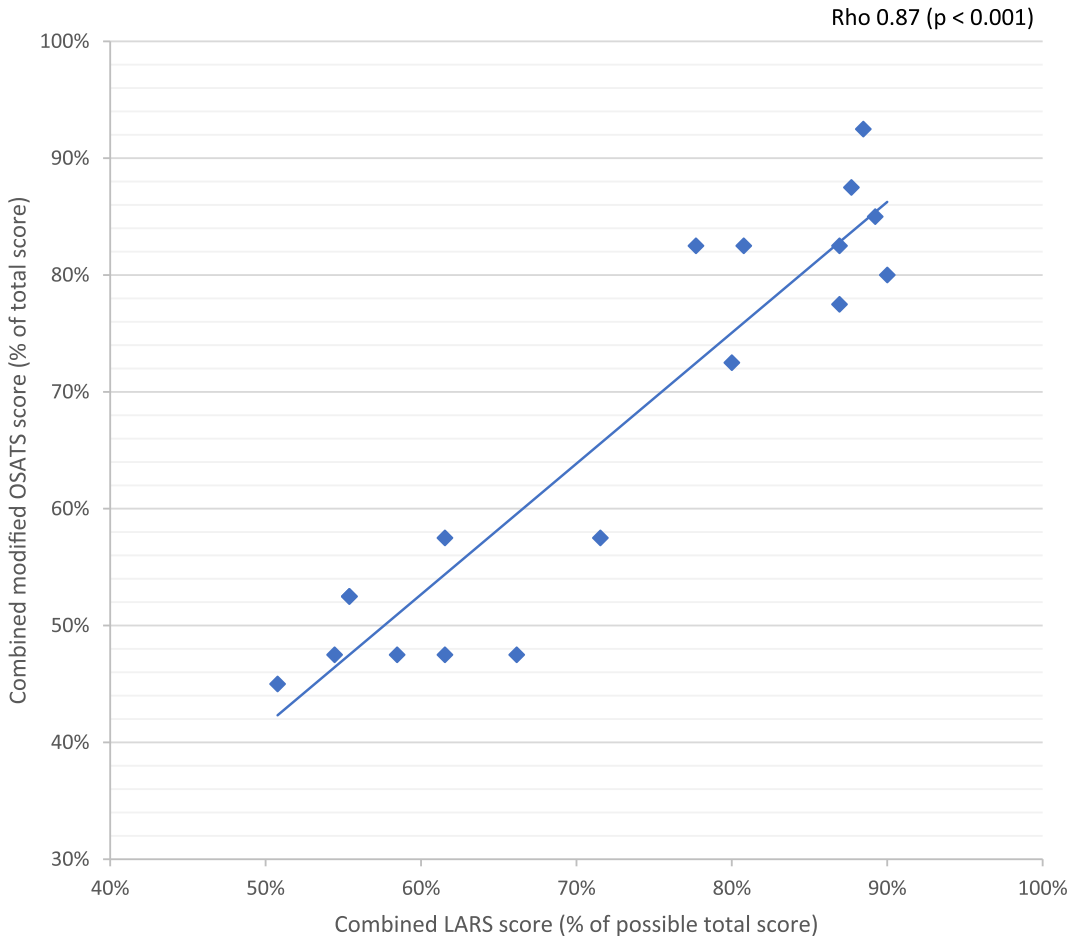

Fig. 1. Correlation between combined LARS scores combined modified OSATS scores.

Validity assessment

Significant differences ($P < 0.001$) were observed between the median LARS scores of the inexperienced surgeons compared to the experienced surgeons for both raters and when scores of both raters were combined (58.5% for inexperienced surgeons versus 86.9% for experienced surgeons, $P < 0.001$) (Table 2). With regards to concurrent validity, a Spearman's correlation coefficient of 0.87 ($P < 0.001$) was observed between the combined LARS scores of both raters and the combined modified OSATS scores of both raters (Fig. 1).

Discussion

This study showed that a systematically developed procedure-specific evaluation tool for LA had excellent inter-rater reliability. LARS was also shown to have construct validity as there were significant differences in the performance scores between inexperienced surgeons and experienced surgeons. LARS scores correlated strongly with the modified OSATS scores, indicating concurrent validity. This means that LARS can be used to accurately assess performance and delineate differences between varying skill levels reliably, regardless of the rater, and to the same degree an established method of skills assessment. It can be used to track the progress of trainees' performances and identify weaknesses, facilitating deliberate practice. LARS has potential value for training programs such as GSET. It could be used to create competence benchmarks for use as performance goals or credentialing trainees for independent practice.

The reliability results of the LARS in this study is higher than achieved in studies of existing rating scales such as GOALS (ICC–0.89 for trained observers),[7] OSATS (Cronbach's alpha = 0.72)[3] and modified OSATS (Cronbach's alpha = 0.76).[3] However, the reliability for the modified OSATS in our study was similar to LARS. The higher reliability for the modified OSATS demonstrated in this study may be a function of the greater difference in experience level between the two groups in our study compared to the previous study. This wide difference in experience level may have made it easier to differentiate performance using the modified OSATS.

The advantage of LARS over global rating scales such as the modified OSTAS is that it can deliver objective feedback on performance at specific steps of a LA. This can allow trainees to identify weaknesses specific to LA that need to be built upon and improved. Procedure-specific rating scales confer a greater ability to differentiate between the performance of surgeons with different experience

levels.[4,9] In an overview of surgical performance measurements, Aggarwal[10] suggested that global scales evaluate overall surgical skill providing a holistic evaluation of technical expertise, which may have a role in overall board/fellowship certification. Comparatively, procedure-specific rating scales evaluate surgical technique and surgical performance, particularly if different techniques are employed, and may be better suited for credentialing trainees for independent practice at specific procedures. Birkmeyer *et al*.[11] demonstrated that greater scores on a modified global rating scale for bariatric surgery correlated with fewer post-operative complication, lower rates of re-operation and readmission. Perhaps a combination of these two useful approaches can provide the most comprehensive assessment of surgical skill and performance.

Ten of the twelve components of LARS demonstrated statistically significant high inter-rater reliability. The two components that did not were the sub-steps 'Appropriately position patient to aid exposure' and 'Deliver the appendix through the umbilical port'. A possible explanation for this is that these sub-steps are inherently difficult to visualize during videos. Similarly, determining if the 'appendix had been identified and exposed using anatomical landmarks' is difficult to determine on a video. These issues were reported by both raters. There was greater agreement between raters for the inexperienced surgeons compared to the experienced surgeons. Perhaps despite participating in a calibration session to familiarize themselves with LARS and instructions on how to use the descriptors of performance, the raters may have differed in their interpretation of what performance constitutes a high score (4 or 5) on the Likert scale. Whereas they may have had similar interpretations of scoring average and low performance. Rater 2 had a wider range of scores for experienced surgeons and higher scores overall than Rater 1. The development of more intensive supervised 'train the trainer' sessions and joint calibration sessions attended by all the raters working through video examples together would improve the reliability of LARS further.

There are some important limitations of this study. The use of videos to retrospectively analyse the performance meant that some items in LARS could not be visualized well and therefore could not be accurately assessed creating less reliable scores for these components as discussed above. The descriptors of performance within LARS required an assessment of 'the need for guidance' as an indication of how well a particular step was performed. This was difficult to fully assess in the videos, especially without audio recordings. A prospective intra-operative assessment would have helped circumvent these problems. However, this would be unblinded, which may introduce bias. The blinded nature of the raters is a strength of this study. Additionally, other descriptors were deliberately included when designing LARS to allow video analysis. Both raters stated they were able to still accurately assess performance despite not being able to tell if verbal guidance was given. As supervising surgeons were asked not to make any changes to their usual practice, any verbal guidance given may have impacted inexperienced surgeon performance. Nevertheless, significant differences between inexperienced and experienced surgeon performance were observed by the two blinded raters suggesting that any verbal guidance did not impact demonstration of the construct validity of LARS.

This study was conducted at one institution which raises questions about the generalisability of the results. However, LARS was developed using the input from surgeons from multiple institutions and designed with the purpose of being widely applicable by accounting for the range of operative techniques that occur in 'everyday' practice. This is one of the reasons we think it has potential to be utilized in surgical training and assessment widely. Nevertheless, there would be value in further studies looking at the reliability and validity of LARS in other institutions, as well as correlating intra-operative prospective assessment and post-operative video-based assessment, and comparing LARS scores to patient outcomes, for example, complications. Whilst collecting a single video from participants allowed a broader cohort of surgeons to be sampled, there is potential for an erroneously poor/good performance to impact the results. However, construct validity was still demonstrated. Investigating trainees' performance over time and multiple procedures following implementation of LARS within a training program would help improve its validity, reliability and generalisability.

This study demonstrated that a multi-institutionally developed procedure-specific rating scale for LA can reliably evaluate performance in an objective manner *in vivo*. The tool was able to accurately differentiate between varying levels of surgical experience accounting for variations in operative techniques. LARS can be used to provide surgical trainees with meaningful objective feedback on their performance. Feedback on the specific steps of a LA gives trainees the opportunity to deliberately practice these steps. Reassessment using LARS following this creates a cycle within which a trainee's progress can be tracked over time. Benchmark LARS scores of experienced surgeon performance, like those demonstrated in this study, could be used for high-stakes assessment and certification for independent practice. Indeed, LARS has the potential to be used in determining eligibility for entrance into the revised GSET program as well as providing a method of assessing and credentialing competence within this new competency-based program.

Acknowledgement

Open access publishing facilitated by The University of Sydney, as part of the Wiley - The University of Sydney agreement via the Council of Australian University Librarians.

Conflict of interest

None declared.

Funding information

Pramudith Sirimanna was funded by a National Health and Medical Research Council (NHMRC) postgraduate scholarship during the conduct of the study (Scholarship number GNT1093784). Marc Gladman has received support from Karl Storz and Applied Medical, during the conduct of the study and personal fees from Health Ed, Elsevier, and Abbvie LTD, outside the submitted work.

Author contributions

Pramudith Sirimanna: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; writing – original draft; writing – review and editing. **Praveen Ravindran:** Data curation; formal analysis; investigation. **Michelle Smigielski:** Data curation; investigation. **Marc A Gladman:** Conceptualization; funding acquisition; methodology; resources; supervision. **Vasi Naganathan:** Formal analysis; project administration; supervision; writing – review and editing.

References

1. Martin JA, Regehr G, Reznick R *et al*. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br. J. Surg.* 1997; **84**: 273–8.
2. Datta V, Bann S, Aggarwal R, Mandalia M, Hance J, Darzi A. Technical skills examination for general surgical trainees. *Br. J. Surg.* 2006; **93**: 1139–46.
3. 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**: 372–9.
4. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: a systematic review. *Am. J. Surg.* 2011; **202**: 469–80.e6.
5. Palter VN, Graafland M, Schijven MP, Grantcharov TP. Designing a proficiency-based, content validated virtual reality curriculum for laparoscopic colorectal surgery: a Delphi approach. *Surgery* 2012; **151**: 391–7.
6. Palter VN, Grantcharov TP. A prospective study demonstrating the reliability and validity of two procedure-specific evaluation tools to assess operative competence in laparoscopic colorectal surgery. *Surg. Endosc.* 2012; **26**: 2489–503.
7. Vassiliou MC, Feldman LS, Andrew CG *et al*. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am. J. Surg.* 2005; **190**: 107–13.
8. 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**: 1525–9.
9. Kramp KH, van Det MJ, Veeger NJ, Pierie JP. Validity, reliability and support for implementation of independence-scaled procedural assessment in laparoscopic surgery. *Surg. Endosc.* 2016; **30**: 2288–300.
10. Aggarwal R. Intraoperative surgical performance measurement and outcomes: choose your tools carefully. *JAMA Surg.* 2017; **152**: 995–6.
11. Birkmeyer JD, Finks JF, O'Reilly A *et al*. Surgical skill and complication rates after bariatric surgery. *N. Engl. J. Med.* 2013; **369**: 1434–42.