Surg Endosc
DOI 10.1007/s00464-010-1329-y

Development, validation and implementation of a monitoring tool for training in laparoscopic colorectal surgery in the English National Training Program

Danilo Miskovic · Susannah M. Wyles · Fiona Carter · Mark G. Coleman · George B. Hanna

Received: 17 May 2010 / Accepted: 9 August 2010

Abstract
Introduction The National Training Program for laparoscopic colorectal surgery (LCS) provides supervised training to colorectal surgeons in England. The purpose of this study was to create, validate, and implement a method for monitoring training progression in laparoscopic colorectal surgery that met the requirements of a good assessment tool.
Methods A generic scale for different tasks in LCS was created under the guidance of a national expert group. The scores were defined by the extent to which the trainees were dependent on support (1 = unable to perform, 5 = unaided (benchmark), 6 = proficient). Trainers were asked to rate their trainees after each supervised case; trainees completed a similar self-assessment form. Construct validity was evaluated comparing scores of trainees at different experience levels (1–5, 6–10, 11–15, 16+) using the Wilcoxon signed-rank test and ANOVA. Internal consistency was determined by Crohnbach's alpha, interrater reliability by comparing peer- and self-assessment (interclass correlation coefficient, ICC). Proficiency gain curves were plotted using CUSUM charts.
Results Analysis included 610 assessments (333 by trainers and 277 by trainees). There was high interrater reliability (ICC = 0.867), internal consistency ($\alpha = 0.920$), and construct validity [$F(3,40) = 6.128$, $p < 0.001$]. Detailed analysis of proficiency gain curves demonstrates that theater setup, exposure, and anastomosis were performed independently after 5 to 15 sessions, and the dissection of the vascular pedicle took 24 cases. Mobilization of the colon and of the splenic/hepatic flexure took more than 25 procedures. Median assessment time was 3.3 (interquartile range (IQR) 1–5) minutes and the tool was accepted as useful [median score 5 of 6 (IQR 4–5)].
Discussion A valid and reliable monitoring tool for surgical training has been implemented successfully into the National Training Program. It provides a description of an individualized proficiency gain curve in terms of both the level of support required and the competency level achieved.

Keywords Education · Laparoscopic surgery · Colorectal surgery · Training · Endoscopy

This study is conducted on behalf of the National Training Program in Laparoscopic Colorectal Surgery.

Presented at the 12th WCES, April 14–17, 2010, National Harbor, MD.

D. Miskovic · S. M. Wyles · G. B. Hanna (✉)
Department of Surgery and Cancer, St Mary's Hospital, Imperial College, Praed St., London W2 1NY, UK
e-mail: g.hanna@imperial.ac.uk

F. Carter
Yeovil District Hospital NHS Foundation Trust, Yeovil, UK

M. G. Coleman
Coordination Office, National Training Program in Laparoscopic Colorectal Surgery, Plymouth, UK

The National Training Program in England provides structured laparoscopic training to established consultant colorectal surgeons, with the goal to train 250 consultants to a level of competence in laparoscopic colorectal surgery (LCS) within 5 years [1]. Monitoring the training progression of trainees is an essential component of such educational programs and requires regular structured assessment [2, 3]. This approach not only provides

Published online: 11 September 2010

information on the effectiveness of the program but also permits the detection of prolonged proficiency gain curves of individual trainees who may require additional support. Formative assessment of surgical performance after each training case also can provide a framework for debriefing and promote the quality and efficiency of training [4].

Previous reports on attempts to implement assessment tools into LCS training programs lacked validity and clear educational impact [5, 6]. Although observational assessment tools have been shown to be reliable and valid for basic surgical skills, their value is restricted for advanced laparoscopic surgery because of a ceiling effect [7–9]. There are numerous studies on the description of the learning curve using clinical outcome parameters and conversion rates [10–12]. These learning curves are often used to define the "proficiency gain process" of a surgeon, thus coining the term "proficiency gain curve" [13]. The use of morbidity and mortality data as the only monitoring parameters, however, raises several problems: aside from the apparent ethical issues, clinical outcomes tend to reflect the proficiency level of the trainer present in the operating room rather than the trainee's actual operative skills [14–18]. Furthermore, clinical outcome data do not provide a descriptive analysis on the operative areas that require improvement.

The goal of this study was to create, validate, and implement a method for monitoring training progression in laparoscopic colorectal surgery that fulfilled the requirements of a good assessment tool [19, 20].

Methods

Scale development

For *content validity*, a selective literature search on operating techniques in LCS was performed using internet sources, educational videos, and books [21–23]. Based on these sources, a generic task analysis for laparoscopic colorectal resections and a scoring system was created. The proposal was sent to members of the educational committee of the National Training Program, represented by 12 expert laparoscopic surgeons, and an educationalist, from 10 different centers in the United Kingdom. The draft was discussed and amended via two committee telephone conferences, and the final version was approved by all members.

Validity, reliability, and implementation

Ethical approval was obtained from the National Ethics Committee (REC 04/Q040356). Trainees were consented by the local collaborators for assessment. The validation period was defined as the first year after implementation. During the first stage of the pilot phase (October 2008–May 2009), the form was available only in paper format. A web-based electronic form was created for the second phase (June 2009–November 2009). The trainers were all consultant surgeons with a special interest in colorectal cancer surgery with several years of experience in laparoscopic surgery and usually many more than 100 resections. They were all approved as trainers for the National Training Program that entailed a peer-review selection process for trainers.

Construct validity was calculated by comparing four different experience groups. Trainees who performed more than 15 consecutive cases were extracted, and their individual data were split into four experience groups (cases 1–5, 6–10, 11–15, and >15). Due to a lack of a "gold standard" method, it was not possible to establish *criterion-related validity*.

Both trainees and trainers had to fill in the same scoring sheet independently, which facilitated the assessment of *interrater reliability*. Analysis of *test-retest* and *intrarater reliability* was not possible due to the setting of the study. Each form also contained a short survey on its perceived usefulness and the time spent to complete it (*acceptability* and *feasibility*).

Statistical analysis

The Statistical Package for the Social Sciences software (Version 17.0.0, SPSS Chicago, IL, USA) and Excel (Microsoft® Excel® for Mac 2008, Microsoft Corporation, Redmond, WA) was used. For construct validity a comparative test for nonparametric data (Wilcoxon rank test) was used. Overall effects were tested using one-way Analysis of Variances (ANOVA). Interrater reliability between the two was measured using the intraclass correlation coefficient (ICC) and demonstrated on a Bland-Altman plot [24]. Inter-item reliability (IIR) was calculated by using Cronbach's α [25]. Cumulative sum (CUSUM) charts were plotted using the equation $S_i = S_{i-1} + (x_i - x_R)$; $S_0 = 0$; $x_R = 5$: S_i is the cumulative sum, x_i the average score per procedure number and x_R the target score [26]. For feasibility, the mean time (standard deviations) to complete a form was calculated. Analysis of Likert scales was performed using median and interquartile range (IQR).

Results

Scale development

Right hemicolectomy, sigmoid resection, anterior resection, low anterior resection, total and subtotal colectomy, and laparoscopically assisted abdominoperineal resections

Table 1 List of generic task zones (A–D) and task steps (1–12)

A. Exposure
1. Operating room setup (position of surgeons, scrub nurse, drapes etc.)
2. Patient positioning
3. Laparoscopic access (open, Veress needle or other techniques, and insertion of ports)
4. Exposure of operating field (moving of omentum, small bowel etc.)
B. Dissection of vascular pedicle
5. Dissection of vascular pedicle (incision of peritoneum, creation of window below and above, and dissection with stapler, clips, ultrasound dissection tool or other techniques)
6. Retrocolic dissection of mesentry (right side toward hepatic flexure, left side toward splenic flexure)
7. Identification of landmark (right side: duodenum, left side: left ureter)
C. Mobilization
8. Dissection of flexure (right side: hepatic, left side: splenic)
9. Mesorectal dissection (including total mesorectal excision (TME), only for rectal resections)
10. Dissection of bowel (transection, using stapler other similar device)
D. Anastomosis
11. Extraction of specimen (creation of incision, bringing out specimen, completion of resection)
12. Anastomosis (intra- or extra-corporeal)

were all classified as laparoscopic colorectal resections. Four generic task zones and 12 generic task steps were identified (Table 1). For some of the operations and techniques some steps were not applicable (e.g., step 9 for right hemicolectomy or step 5 for medial to lateral dissection without laparoscopic dissection of vascular pedicle). Some non-laparoscopic steps (e.g., patient positioning) were included in the list because they may substantially facilitate the procedure. Others (e.g., wound closure) were not included, because they were not deemed to be related to the quality of advanced laparoscopic surgical skills.

The rationale for the scoring system was to avoid judgmental expressions, such as "good" or "poor." Instead the degree of support required to complete the task was used as a more objective and reliable parameter. The amount of physical and verbal support required from the trainer was divided into six categories using a "Juster" scale (Table 2) [27]. For all other situations (e.g., another trainee performed the step, step was not necessary for the operation, or the trainer performed the step when restricted by time constraints etc.), a "not applicable" option was added for each step. A full description of the assessment tool can be downloaded from the internet [28].

Table 2 Rating score for the amount of support required to perform the task

0: Not applicable
1: Not performed, step had to be done by trainer
2: Partly performed, step had to be partly done by trainer
3: Performed, with substantial verbal support
4: Performed with minor verbal support
5: Competent performance, safe (without guidance)
6: Proficient performance, couldn't be better

Construct validity and reliability

Thirty trainers submitted 333 forms, and 52 trainees 277 forms, from a total of 10 national training centers. Twelve trainees underwent more than 15 training sessions during the pilot phase each. The results are summarized in Table 3.

Construct validity was analyzed on 193 cases from 12 trainees who performed more than 15 cases. One-way analysis of variance of four experience groups (groups 1–4) confirmed the difference between the experience levels ($F(3,40) = 6.128$, $p < 0.001$; Fig. 1). There were significant differences between group 1 and 2 (4 vs. 4.6, $p = 0.030$) and between 3 and 4 (4.7 vs. 5, $p = 0.008$), but not for groups 2 and 3 ($p = 0.850$).

Interrater reliability analysis was conducted on 94 cases, for which both the trainer and the paired trainee returned a completed form. For the remaining cases, one of the two failed to complete a form or did not provide the patient identification number. These issues were resolved at a later stage after improving participant instructions and by introducing an online submission using an electronic form. Comparison of complete datasets showed a high level of interrater reliability (ICC = 0.867; $F(94, 94) = 13.989$,

Table 3 Summary of the validation and implementation process

Parameter	Test	*p* Value
Construct validity	ANOVA	<0.001
Reliability		
Interrater	ICC	0.867
Inter-item (trainer)	Cronbach's alpha	0.937
Inter-item (trainee)	Cronbach's alpha	0.920
Feasibility	Score	5/6

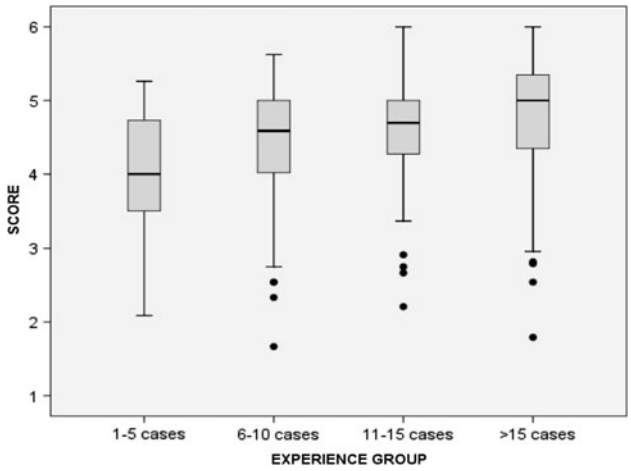

Fig. 1 Box plots for four different experience groups (12 trainees)

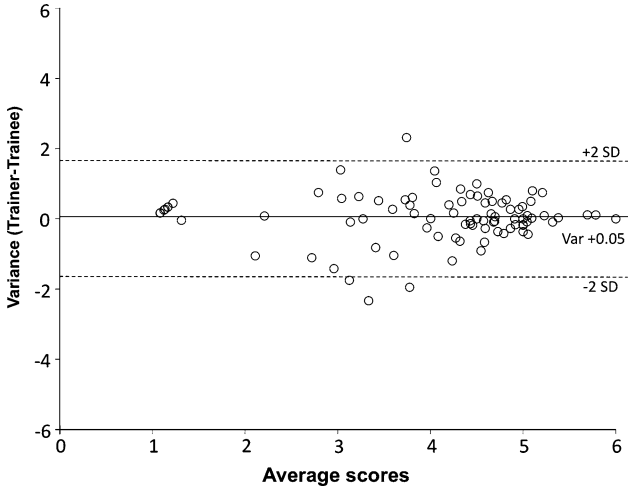

Fig. 2 Bland-Altman plot shows minimal interrater difference

$p < 0.001$; Fig. 2). On the analysis of the complete data set (610 forms), there was excellent IIR with $\alpha = 0.937$ for the trainers' assessments and $\alpha = 0.920$ for the trainees' assessments.

Implementation

The median assessment time was 3.3 (interquartile range (IQR) = 1–5) minutes. The median score for the usefulness of the form was 5 (IQR = 4–5: 1 = not useful at all, 6 = very useful).

For the construction of proficiency gain curves, the trainers' scores of all 333 procedures were used. The target score was set at 5 ["competent performance, safe (without guidance)"]. The average score indicates that the target score was consistently achieved after the 21st procedure. However, this value is different for different components of the operation. For a detailed analysis, the tasks steps were regrouped and CUSUM charts of six different skill areas were computed. The areas "Set-up" and "Anastomosis" reached target scores from a very early stage. The curve for "Exposure" shows an inflection point around the 17th procedure, and for "Vascular pedicle," this occurred around the 22nd procedure. For "Mobilization" and "Hepatic/ splenic flexure," no inflection point was observed until procedure 25 (Fig. 3a). Furthermore, proficiency gain curves can differ substantially between individual trainees, as shown in Fig. 3b.

Discussion

In this study, reliability, validity, and feasibility of an observational assessment tool to monitor the proficiency gain process of laparoscopic colorectal surgery was demonstrated. To our knowledge, there is no comparable data in a similar setting of a National Educational Program for advanced laparoscopic surgery.

The main strength of this tool is the combination of a task-specific checklist with a generic scale, which is highly practical and yet remains reliable. The use of generic task steps for laparoscopic colorectal resections allows the identification of areas within the operation that may be more difficult to master and can thus help the trainers to focus their teaching on these areas. The descriptive scale ("Juster" scale) details the amount of support needed to complete a task, which is a more objective parameter than judgmental scales [20, 27]. This may have added to the high level of interrater reliability. Other tools, such as the Objective Structured Assessment of Technical Skill (OSATS), also use descriptive scales, but each category has its specific descriptors, which makes the assessment more difficult and creates a learning curve for the assessor [8]. In a recent study, the authors constructed an assessment tool for laparoscopic colorectal surgery following a task-specific approach based on OSATS [29]. Using OSATS requires assessor training, which is logistically difficult to coordinate for a National Training Program with a high number of trainers in different locations. One previous study used a computer-based database to enter performance-related parameters of trainees in a single-center study, showing construct validity when comparing trainees at different levels [6]. However, the conclusions and the educational value remain unclear. Other generic assessment tools for laparoscopic surgery have shown construct validity, but none of them was used for the continuous description of the learning curve [8, 9]. The monitoring tool presented in the current study has been widely accepted within the National Training Program in England, as shown by the reported high trainer satisfaction levels. Another indicator for its popularity is the fact that it is

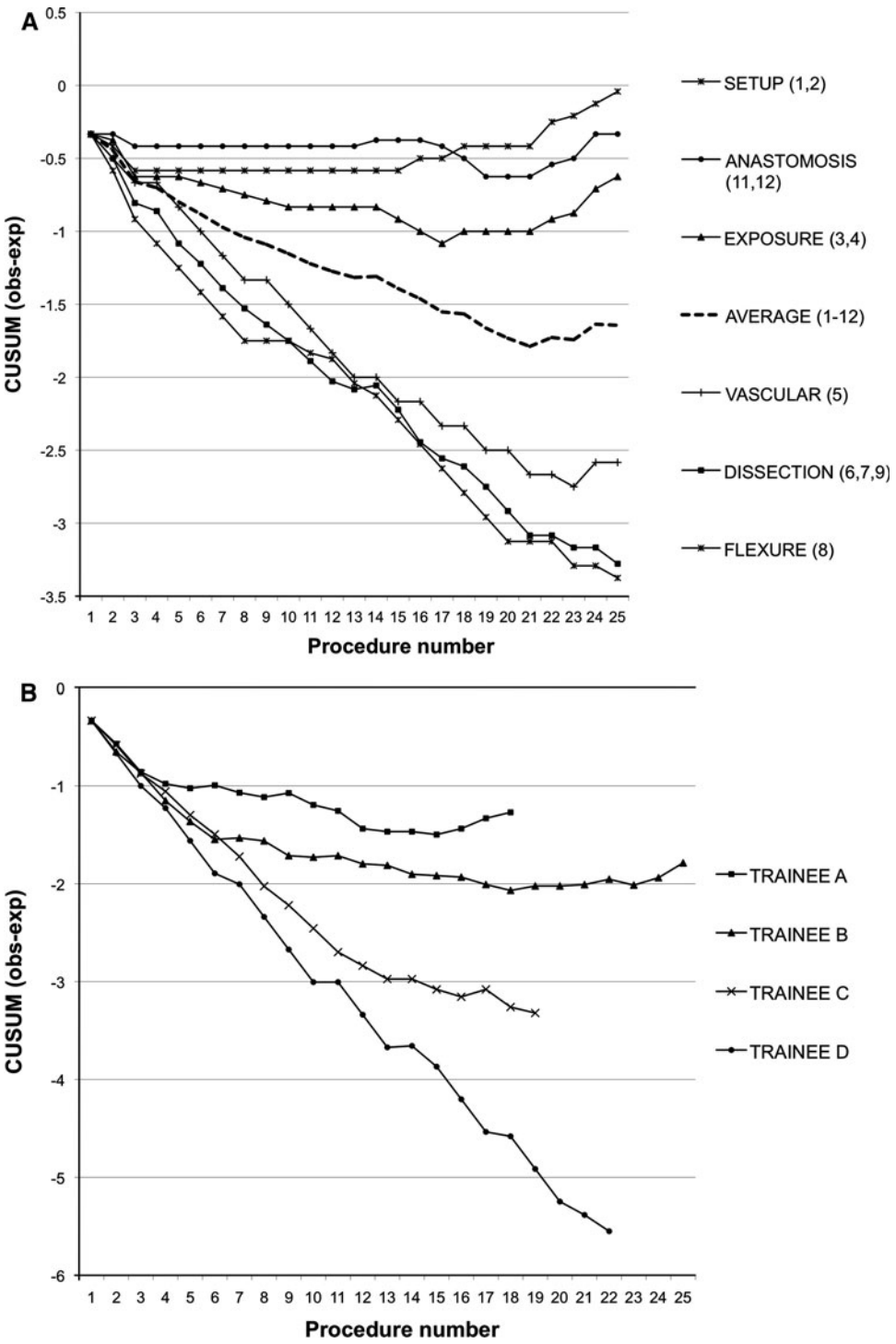

Fig. 3 **a** CUSUM chart for average and detailed scores (in brackets the task step numbers used to regroup the skill areas (see text and Table 1 for explanation). **b** Examples of four different individual trainee-curves. Trainees A and B develop as expected, trainee C may be just about to reach the target, whereas trainee D shows little improvement over time

being used increasingly outside the scope of the National Training Program by general surgical trainees (registrars) and their trainers in the United Kingdom.

CUSUM charts were used to identify the point at which the trainees reach a satisfactory score. Although the proficiency gain curve for the overall score indicates that the target score is generally achieved after the 21st procedure, this value is different for individual trainees and for various components of the operation. Certain steps of the procedure (vascular pedicle, mobilization, hepatic and splenic flexure) may take longer to learn than the overall score suggests. This information may be useful for task-specific training models, where trainees perform certain parts of the procedure according to their proficiency level rather than a whole case. Previously reported learning curves by self-taught laparoscopic surgeons using clinical outcome data

suggested that 50–70 cases were required for competent performance [10, 12, 30]. These proficiency gain curves, however, are hardly comparable to the present data because the current trainees are established consultants who have been trained actively by expert laparoscopic surgeons, hence the learning curve may be shortened substantially [17, 31–33]. The data also demonstrated variations between the different candidates in achieving their path to competency. This highlights the need for an individualized competency-based assessment rather than a set number of cases to be signed off for independent practice.

At the beginning of the implementation of the monitoring tool, the submission rate was lower than expected. Two events account for a significant improvement in the submission rate. The first was the introduction of a web-based assessment form by the coordinating center, which facilitated the process substantially. The second was that the governing body of the National Training Program used the number of completed forms as the evidence of training activity and consequently for funding allocation.

There are some limitations to this study. Content validity was not evaluated in a systematic way. Although multiple sources (literature, expert opinion, internet search) were used to establish content validity, certain procedural skills may have been missed. Also, we did not apply the generalizability theory to further explore reliability, because there were too few facets to implement into the equation [20, 34]. Furthermore, although trainees and trainers were instructed to fill in the form independently, there was no control and there is a potential risk for false-positive interrater reliability.

Future research may expand to other disciplines. The generic scale can be easily used for other procedures or interventions by changing the task list accordingly. Although the application of this tool enables the description of proficiency gain curves, it does not necessarily provide reliable information on the competency level at the end of the training. Further assessment methods need to be developed for this stage, such as human reliability methods, to identify errors and near misses, and use these as a surrogate marker for the maturity of technical performance [35, 36]. Furthermore, it would be interesting to measure the impact of structured feedback on the training progression, which is assumed but has yet to be shown in surgery.

In conclusion, a valid and reliable monitoring tool for surgical training has been implemented successfully into a National Training Program. It provides a description of individualized proficiency gain curve in terms of the level of support required and the competency level achieved.

Acknowledgments This work is funded by the National Cancer Action Team as part of the National Training Program in Laparoscopic Colorectal Surgery. We would like to thank all members of the educational committee and the steering group representing the centers participating in the National Training Program for their boundless effort to support the educational activities: Austin Acheson and Charles Maxwell-Armstrong (Nottingham), Tom Cecil (Basingstoke), Chris Cunningham (Oxford), Vivek Datta and Sav Papagrigoriadis (London), Nader Francis (Yeovil), John Griffith (Bradford), James Gunn (Hull), Alan Horgan (Newcastle-upon-Tyne), Robin Kennedy (Harrow, London), Roger Motson (Colchester), Amjad Parvaiz (Portsmouth), and Timothy Rockall (Guildford).

Disclosures Miskovic, S.M. Wyles, F. Carter, M.G. Coleman, and G.B. Hanna have no conflicts of interest or financial ties to disclose.

References

1. The National Training Programme in laparoscopic colorectal surgery. Lapco website. www.lapco.nhs.uk
2. van der Vleuten CP, Schuwirth LW (2005) Assessing professional competence: from methods to programmes. Med Educ 39:309–317
3. Joint Advisory Group on GI Endoscopy (JAG) website. www.thejag.org.uk
4. Black P, Wiliam D (2009) Developing the theory of formative assessment. Educ Assess Eval Account 21:5–31
5. Sidhu RS, Vikis E, Cheifetz R, Phang T (2006) Self-assessment during a 2-day laparoscopic colectomy course: can surgeons judge how well they are learning new skills? Am J Surg 191:677–681
6. Wohaibi EM, Earle DB, Ansanitis FE, Wait RB, Fernandez G, Seymour NE (2007) A new web-based operative skills assessment tool effectively tracks progression in surgical resident performance. J Surg Educ 64:333–341
7. Munz Y, Moorthy K, Bann S, Shah J, Ivanova S, Darzi SA (2004) Ceiling effect in technical skills of surgical residents. Am J Surg 188:294–300
8. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84:273–278
9. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondre K, Stanbridge D, Fried GM (2005) A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 190:107–113
10. Tekkis PP, Senagore AJ, Delaney CP, Fazio VW (2005) Evaluation of the learning curve in laparoscopic colorectal surgery: comparison of right-sided and left-sided resections. Ann Surg 242:83–91
11. Schlachta CM, Mamazza J, Seshadri PA, Cadeddu M, Gregoire R, Poulin EC (2001) Defining a learning curve for laparoscopic colorectal resections. Dis Colon Rectum 44:217–222
12. Park IJ, Choi GS, Lim KH, Kang BM, Jun SH (2009) Multidimensional analysis of the learning curve for laparoscopic colorectal surgery: lessons from 1,000 cases of laparoscopic colorectal surgery. Surg Endosc 23:839–846
13. Cuschieri A (2006) Nature of human error: implications for surgical practice. Ann Surg 244:642–648
14. Bull C, Yates R, Sarkar D, Deanfield J, de Leval M (2000) Scientific, ethical, and logistical considerations in introducing a new operation: a retrospective cohort study from paediatric cardiac surgery. BMJ 320:1168–1173
15. Dalton S, Ghosh A, Zafar N, Riyad K, Dixon A (2009) Competency in laparoscopic colorectal surgery is achievable with

appropriate training but takes time: a comparison of 300 elective resections with anastomosis. Colorectal Dis, Jul 6 [Epub ahead of print]
16. Rijbroek A, Wisselink W, Rauwerda JA (2003) The impact of training in unselected patients on mortality and morbidity in carotid endarterectomy in a vascular training center and the recommendations of the European Board of Surgery Qualification in Vascular Surgery. Eur J Vasc Endovasc Surg 26:256–261
17. Li JC, Hon SS, Ng SS, Lee JF, Yiu RY, Leung KL (2009) The learning curve for laparoscopic colectomy: experience of a surgical fellow in an university colorectal unit. Surg Endosc 23:1603–1608
18. Miskovic D, Wyles SM, Ni M, Darzi AW, Hanna GB (2010) Mentoring and simulation in laparoscopic colorectal surgery. A systematic review. Ann Surg (in press)
19. van Mook WN, van Luijk SJ, O'Sullivan H, Wass V, Schuwirth LW, van der Vleuten CP (2009) General considerations regarding assessment of professional behaviour. Eur J Intern Med 20:e90–e95
20. Streiner N, Norman G (2008) Health measurement scales. A practical guide to their development and use, 4th edn. Oxford University Press, Oxford
21. Evans J, Jenkins I, Kennedy RH (2009) Laparoscopic right hemicolectomy. St. Marks Multimedia, UK (20 mins)
22. Champault G, Schimmelpenning H (2007) Laparoscopic sigmoidectomy for cancer, operation primer edn, Springer
23. WebSurg website. http://www.websurg.com
24. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310
25. Bland JM, Altman DG (1997) Cronbach's alpha. BMJ 314:572
26. Grigg OA, Farewell VT, Spiegelhalter DJ (2003) Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. Stat Methods Med Res 12:147–170
27. Hoek J, Gendall P (1993) A new method of predicting voting behaviour. J Market Res Soc 35:361–373
28. Lapco website (link to download assessment form). http://www.lapco.nhs.uk/userfiles/file/Form%20A%20GAS%20assessment%20sheet.pdf
29. Sarker SK, Kumar I, Delaney C (2010) Assessing operative performance in advanced laparoscopic colorectal surgery. World J Surg 34(7):1594–1603
30. Dincler S, Koller MT, Steurer J, Bachmann LM, Christen D, Buchmann P (2003) Multidimensional analysis of learning curves in laparoscopic sigmoid resection: eight-year results. Dis Colon Rectum 46:1371–1378 discussion 1378–1379
31. Schlachta CM, Mamazza J, Gregoire R, Burpee SE, Pace KT, Poulin EC (2003) Predicting conversion in laparoscopic colorectal surgery. Fellowship training may be an advantage. Surg Endosc 17:1288–1291
32. Birch DW, Asiri AH, de Gara CJ (2007) The impact of a formal mentoring program for minimally invasive surgery on surgeon practice and patient outcomes. Am J Surg 193:589–591 discussion 591–582
33. Choi DH, Jeong WK, Lim SW, Chung TS, Park JI, Lim SB, Choi HS, Nam BH, Chang HJ, Jeong SY (2009) Learning curves for laparoscopic sigmoidectomy used to manage curable sigmoid colon cancer: single-institute, three-surgeon experience. Surg Endosc 23:622–628
34. Crossley J, Davies H, Humphris G, Jolly B (2002) Generalisability: a key to unlock professional assessment. Med Educ 36:972–978
35. Joice P, Hanna GB, Cuschieri A (1998) Errors enacted during endoscopic surgery—a human reliability analysis. Appl Ergon 29:409–414
36. Talebpour M, Alijani A, Hanna GB, Moosa Z, Tang B, Cuschieri A (2009) Proficiency-gain curve for an advanced laparoscopic procedure defined by observation clinical human reliability assessment (OCHRA). Surg Endosc 23:869–875