Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology

Timothy M. Kowalewski, PhD,[a,*] *Lee W. White, PhD,*[b] *Thomas S. Lendvay, MD, FACS,*[c] *Iris S. Jiang, BS,*[b] *Robert Sweet, MD, FACS,*[d] *Andrew Wright, MD,*[e] *Blake Hannaford, PhD,*[f] *and Mika N. Sinanan, MD, PhD, FACS*[e]

[a] *University of Minnesota Mechanical Engineering, Minneapolis, MN*
[b] *University of Washington Bioengineering, Seattle, WA*
[c] *University of Washington Urology and Seattle Children's Hospital, Seattle, WA*
[d] *University of Minnesota Urology, Minneapolis, MN*
[e] *University of Washington Surgery, Seattle, WA*
[f] *University of Washington Electrical Engineering, Seattle, WA*

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form
12 May 2014
Accepted 27 May 2014
Available online 4 June 2014

Keywords:
Laparoscopy
Simulation
Education
Objective metrics
FLS
Validation
Grasping force

ABSTRACT

Background: Laparoscopic psychomotor skills are challenging to learn and objectively evaluate. The Fundamentals of Laparoscopic Skills (FLS) program provides a popular, inexpensive, widely-studied, and reported method for evaluating basic laparoscopic skills. With an emphasis on training safety before efficiency, we present data that explore the metrics in the FLS curriculum.
Materials and methods: A multi-institutional ($n = 3$) cross-sectional study enrolled subjects ($n = 98$) of all laparoscopic skill levels to perform FLS tasks in an instrumented box trainer. Recorded task videos were postevaluated by faculty reviewers ($n = 2$) blinded to subject identity using a modified Objective Structured Assessment of Technical Skills (OSATS) protocol. FLS scores were computed for each completed task and compared with demographically established skill levels (training level and number of procedures), video review scoring, and objective performance metrics including path length, economy of motion, and peak grasping force.
Results: Three criteria used to determine expert skill, training and experience level, blinded review of performance by faculty via OSATS, and FLS scores, disagree in establishing concurrent validity for determining "true experts" in FLS tasks. FLS-scoring exhibited near-perfect correlation with task time for all three tasks (Pearson $r = 0.99$, 1.00, 1.00 with $P < 0.00000001$). FLS error penalties had negligible effect on FLS scores. Peak grasping force did not correlate with task time or FLS scores.
Conclusions: FLS technical skills scores presented negligible benefit beyond the measurement of task time. FLS scoring is weighted more toward speed than precision and may not significantly address poor tissue handling skills, especially regarding excessive grasping force. Categories of experience or training level may not form a suitable basis for establishing proficiency thresholds or for construct validity studies for technical skills.

* *Corresponding author*. University of Minnesota, Minneapolis, MN 55455. Tel.: +1 612 626 0054; fax: +1 612 625 4344.
E-mail address: tmk@uw.edu (T.M. Kowalewski).
0022-4804/$ – see front matter
http://dx.doi.org/10.1016/j.jss.2014.05.077

1. Introduction

The Accreditation Council of Graduate Medical Education specifically identifies "technical competence in conducting surgical procedures" as a component in its core competencies under patient care and practice-based learning [1]. Professional surgical organizations responsible for certification of technical knowledge and proficiency and/or competency also need means for accurately assessing surgical skills [2]. Skill assessment systems such as Objective Structured Assessment of Technical Skills (OSATS) or the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills and its evolution into the Fundamentals of Laparoscopic Skills (FLS) have provided a framework to analyze surgical skill [3–6]. In 2004, the FLS assessment certification process was launched and after 5 y of its existence, almost 3000 clinicians have become certified [7]. We seek to build on this success by evaluating the metrics within existing FLS scoring and defining options for possible improvement.

The psychomotor skills aspect of FLS measures competency in the domain of laparoscopic surgical skills with several different measures. The FLS score for each task is derived from a formula based on task time and numbers of errors (Table 1), and is weighted such that time is heavily weighted and usually a dominant factor in the formula [8]. One of the limitations of the technical skills portion of FLS is that there are few objective metrics. A heavy focus on task time may mask other important measures of safety and proficiency. Task time alone is unlikely to accurately reflect the quality of the surgical performance and rushing to complete a task in the fastest possible time may negatively impact quality. Meanwhile, tissue handling as measured by grasping force on tissue as a metric is not measured. It directly correlates with tissue damage. This can lead to such poor clinical correlates such as a tear in a critical anatomic structure, a leaking anastomosis, or a physiologically significant stricture.

Although the discriminating power of FLS clearly separates experienced and inexperienced subjects [9–12], more granular measures of surgical performance such as economy of motion (EoM), grasp forces, and tool motion characteristics may also be important independent measure of performance but have not been evaluated. Physical model-based laparoscopic box trainers are not capable of capturing or reporting these metrics routinely. Moreover, there is no rigorous and repeatable method established to carry out identical performance measures on different platforms [1,13–16].

To address these issues, we created a laparoscopic box trainer that uses real tools and physical models in dry or wet laboratory tasks but also contains instrumentation for automated capture and analysis of tool motion, EoM, and grasping forces [17]. Building on prototype models (the red [18], and blue dragon [19]) this instrumented trainer box is a scaled-down, table-top version of the larger scale platforms, which were successfully used as a surgical skills research tool in live porcine surgery. De *et al.* [20] used a related grasper mechanism to establish peak grasp force thresholds that result in tissue damage. The red dragon prototype, along with the established hidden Markov model-based scoring methodology [21] as licensed and commercialized by Simulab Corporation (Seattle, WA), was incorporated into the EDGE (Fig. 1). To our knowledge, EDGE is the only dry laboratory reality-based box trainer that can obtain high-accuracy tool motion (position and orientation) measurements along with grasping force and synchronized video [22]. We used EDGE as a research platform to evaluate different measure of quantitative skill in FLS tasks.

Table 1 – Equations used to compute FLS scores [23,24,28].

FLS task	FLS score
PegTx	$FLS_{Peg} = (300 - t - 17E_{dr})/237$
Cutting	$FLS_{Cut} = (300 - t - 2E_a)/280$
Suturing	$FLS_{Sut} = (600 - t - E_{pd} - E_g - E_q)/520$

The goal of this work is to evaluate the training goals and scoring methods of FLS psychomotor skill scoring and compare them with a broader set of performance criteria available in EDGE. We hypothesize that existing FLS scoring with its emphasis on task time may unintentionally promote speed over careful tissue handling and lead to less reliable or less accurate global measures of skill. We further propose that

Fig. 1 – The EDGE Platform was developed by Simulab Corporation and is based on a mechanism developed by the University of Washington Biorobotics laboratory. It consists of a pair of interchangeable surgical tools whose motion is constrained to rotate about a fulcrum the same way laparoscopic instruments are constrained by their access ports. (Color version of figure is available online.)

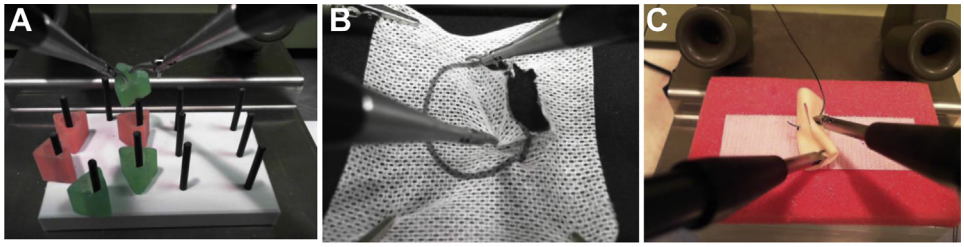

Fig. 2 – Screenshots of synchronized videos recorded during performance of the three FLS tasks: (A) PegTx, (B) cutting, and (C) suturing. (Color version of figure is available online.)

using demographic categories such as training level or experience alone may be insufficient when used to establish skill categories for construct validity analysis or to set performance thresholds for training and high-stakes testing.

2. Methods

2.1. *EDGE platform description and analysis methods*

This study used the EDGE platform to collect tool motion and task video data from volunteer participants consisting of surgical faculty, surgical residents, and fellows at multiple surgical centers in the United States.

EDGE hardware consists of an instrumented mechanism, which holds laparoscopic tools about a fixed pivot point. The tools are either Stryker Endoscopy (San Jose, CA) tools with interchangeable tool tip inserts (5 mm diameter, 33 cm length with 250-080-282 Maryland Grasper or 250-080-267 Endo Metzenbaum Curved Scissor inserts) or Karl Storz (Tuttlingen, Germany) curved needle drivers (26,173 KAL and KAR). Position sensors (potentiometers and optical encoders) measure tool Cartesian position (x, y, z in cm), tool roll (degrees), and grasper jaw angle (degrees). A calibrated strain-gauge measures grasping force (Newtons). EDGE includes custom software that time-stamped all sensors measurements for each hand at 30 Hz with synchronized video capture of the task. Videos are compressed real-time with MPEG4 codecs at ~10 MB per minute of video. Timing is automatically tracked and recorded. Timing begins and ends when tools are moved in and out of a fixed "homebase." FLS task blocks are attached within the workspace, rigidly and repeatedly aligned to a fixed coordinate system centered about the task.

EDGE uses a laptop computer running Microsoft (Redmond, WA) Windows XP (University of Washington sites only) or Windows 7 (all other sites) interfaced via universal serial bus 2.0 to the box trainer during all recordings. EDGE's custom software stores tool data and task video files and labels them with unique codes to identify subject, site, task, date, and time. All tool motion and demographic data for all recordings are compiled into a single database, sorted, and analyzed by custom scripts in MATLAB software (Mathworks Inc, Natick, MA). Pearon r and Spearman ρ with corresponding (P value) were adopted as measures of correlation for linearity and monotonicity, respectively.

2.2. *Subject pool description and data collection sites*

Subject enrollment was approved and registered under Western institutional review board 19125-A/B. This multicenter effort included surgeons of various skill levels from the University of Washington, the University of Minnesota, and three sites in the city of New Orleans. The subject pool spanned General Surgery, Urology, and Gynecology specialties. It consisted of active surgical faculty, surgical fellows, residents, and experienced practicing surgeons. Medical students pursuing surgical practice or FLS-experienced technicians were also enrolled in the study. Participation was voluntary and subjects were included as study participants if they met the previously mentioned criteria and excluded otherwise. Subjects were instructed to follow the FLS instructions for each task that is, to minimize completion time and minimize task-

Table 2 – Psychomotor OSATS (p-OSATS) grading scale used to evaluate and numerically code psychomotor skill [4,6].

Score	Bimanuality	Motion quality
1	One arm paralyzed, offering no help to complete the step.	Unnecessary, hesitant or awkward movements of tools.
2		
3	Using both arms most of the time, but clear perceivable bias of accomplishing most of the task with dominant hand.	Reasonably efficient movements of tools but frequent non effective moves.
4		
5	Both arms naturally complementing each other. Optimal use of nondominant hand.	Elegant, fluid, and efficient movements of tools.

specific errors related to precision. These errors are described in the following, respectively.

Three EDGE platforms, one dedicated to each task, were deployed at each site to maximize subject throughput and allow for simultaneous subjects. An approved study administrator set up the equipment at each site on a daily basis and subjects were invited to voluntarily participate in the study whenever their schedules allowed. Subjects were allowed to complete the study over multiple sessions. Data collection at each site lasted from 10 d–3 wk. Collections were established either adjacent to surgical simulation training centers or in the vicinity of the operating room.

Subject demographics were recorded after consent was obtained. A de-identified questionnaire included subject's gender, age, handedness, training level, surgical specialty, laparoscopic experience, approximate number of relevant procedures done (where a subject completed more than half of the case), time since last laparoscopic procedure, total number of FLS tasks done in past, and FLS certification status. A post-task questionnaire invited feedback regarding the acceptability of EDGE based on categorical Likert scales and written general comments. Open-ended subjective observations relative to the platform were collected by the study administrator.

2.3. *Surgical task description, iterations, and FLS scoring*

An iteration was defined as one complete execution of a single task and its associated data. Three of the five FLS tasks were chosen to assess relevant skills of the most common surgical tasks in the shortest subject participation time. The tasks used in the study were Peg transfer (PegTx), cutting, and intracorporeal suturing. Descriptions of each task appear in the following along with representative screen-shots in Figure 2. Subjects were asked to complete three iterations of the PegTx, two iterations of the cutting task, and two iterations of the suturing task, in that order. Subjects were also invited to perform additional iterations of any task if they were willing to do so. Each subject was introduced to each task via printed instructions. For each iteration, time (t) started on removing either tool tip from a fixed "home-base" position and stopped when both tools are returned to that position. The published FLS scoring methodology [23,24] was reviewed and equations were extracted to calculate FLS scores for each task iteration. Explicit equations used for computation are shown in Table 1 and each task's error variables E_{Error} are described in the following. Negative scores were not truncated to zero for correlation computations or plots.

The PegTx, also called block transfer, used two curved Maryland-type laparoscopic graspers. Instructions were to transfer six blocks as fast as possible but without errors, from one side to another and back again without regard for order or color of blocks, to transfer each block mid-air between hands without dropping blocks. EDGE automatically computed task time t. Each video was later manually reviewed to count the total number of nonrecovered drops considered to be errors E_{dr}.

The Circle Cutting task (cutting), also called pattern cutting, used a curved Maryland grasper in the nondominant hand and a curved shears in the dominant hand for the duration of each task. Instructions were to cut gauze along a marked circular pattern (diameter = 4 cm) in minimum time and with minimal error and to begin by either making a puncture anywhere on the circle or cutting in from the gauze edge. Task time t was automatically computed. The percentage of accumulated area cut beyond the marked circle boundary composed the error count E_a. To minimize subjective grader error and exploit automation, cut circles were flattened and electronically scanned, and cutting error was automatically computed via ImageJ (NIH, Bethesda, MD), a public domain image processing suite [25]. The wand tool automatically outlined and measured out-of-bound areas through an edge-finding algorithm that measures pixel values. Scans included a printed reference line to scale the image from pixels to real-world units (millimeters).

The intracorporeal suturing task (suturing), also called knot tying with intracorporeal suture, used two curved needle drivers. Subjects had a choice of a ratcheting or nonratcheting mechanism at the beginning of each task. Instructions were to complete the task as fast as possible but also without errors. The task was to puncture a Penrose drain at marked entry and exit dots with a 2-0 V-20 tapered half circle needle and 12.5 cm of suture and tie a three throw surgeon's knot. Errors were quantified by standard FLS methodology included distance away from puncture dots in mm (E_{pd}), gap of sutured slit in mm (E_g), and knot quality (E_q) where 0 indicated a secure knot, 10 a slipping knot, and 20 a knot that came apart. Task time t

Table 3 – Intra-rater and inter-rater reliability of p-OSATS scores. Coders only scored the Ground Truth expert candidates subset ($n = 56$). Intra-rater reliability compares coder A's scores with himself on a given examination; inter-rater compares coder A's mean scores with coder B's on that examination. Pearson *r* and Spearman ρ with corresponding (*P* value) are shown.

	Correlation coefficient	PegTx	Cutting	Suturing
Intra-rater				
Bimanuality	*r*	0.57 (0.01)	0.49 (0.03)	0.76 (0.00)
	ρ	0.45 (0.04)	0.51 (0.02)	0.69 (0.00)
Motion	*r*	0.74 (0.00)	0.34 (0.14)	0.85 (0.00)
quality	ρ	0.70 (0.00)	0.32 (0.17)	0.90 (0.00)
Inter-rater				
Bimanuality	*r*	0.55 (0.01)	0.04 (0.86)	0.70 (0.00)
	ρ	0.45 (0.05)	0.04 (0.88)	0.66 (0.01)
Motion	*r*	0.63 (0.00)	0.27 (0.26)	0.78 (0.00)
quality	ρ	0.60 (0.00)	0.21 (0.38)	0.80 (0.00)

Table 4 – Overview of all collected data totals broken down by site and categories.

Category	Site 1	Site 2	Site 3	Total
Faculty subjects	6	8	3	17
Total subjects	32	35	31	98
PegTx iterations	78	88	27	193
Cutting iterations	61	53	51	165
Suturing iterations	0	59	30	89
Total iterations	139	200	108	447
Total time (h)				22.7

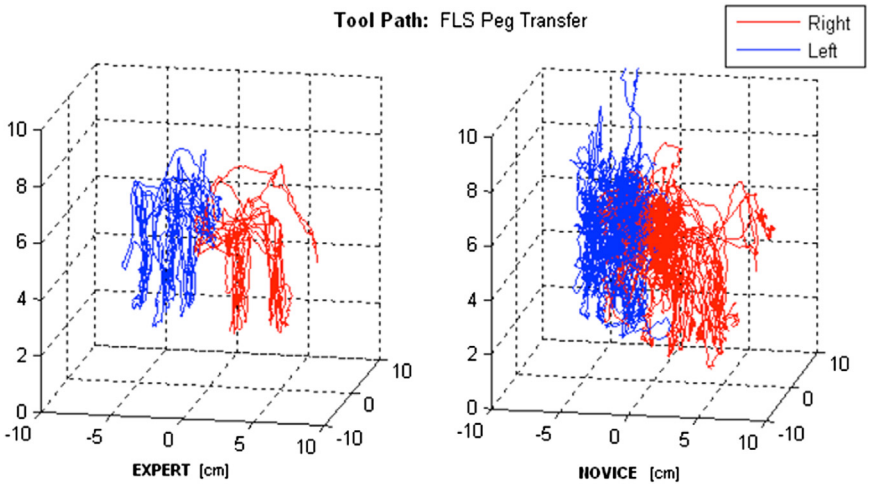

Fig. 3 – 3D Tool path plots for the left (blue) and right (red) hands of one iteration of a PegTx task for a faculty surgeon expert (left) and novice (right). (Color version of figure is available online.)

was automatically computed, and errors were manually determined by two FLS-certified graders.

2.4. *Electronically measured objective performance metrics*

Automatically computed tool-tip motion metrics such as tool path length (PathLength, the sum of left, and right tool-tip distance traveled) and (EoM, the ratio of path length and task time) were provided by EDGE software along with measured grasp force exerted by a surgeon at the handle. The maximum of the peak force (F_{peak}) exerted by either hand during an iteration was used to characterize force behavior. Each of these automated metrics was calculated in the same way for all iterations, independent of task.

2.5. *Establishment of the "true expert" set*

Three methods commonly used in the surgical literature to define "expertise" to establish construct validity were used. All were then combined to establish a set of individual iterations (not subjects) that exemplified "true expertise." First, a subject's self-reported demographic criteria were considered. These included training level (none, medical school, post graduate years of residency 1–5, fellowship, and practicing

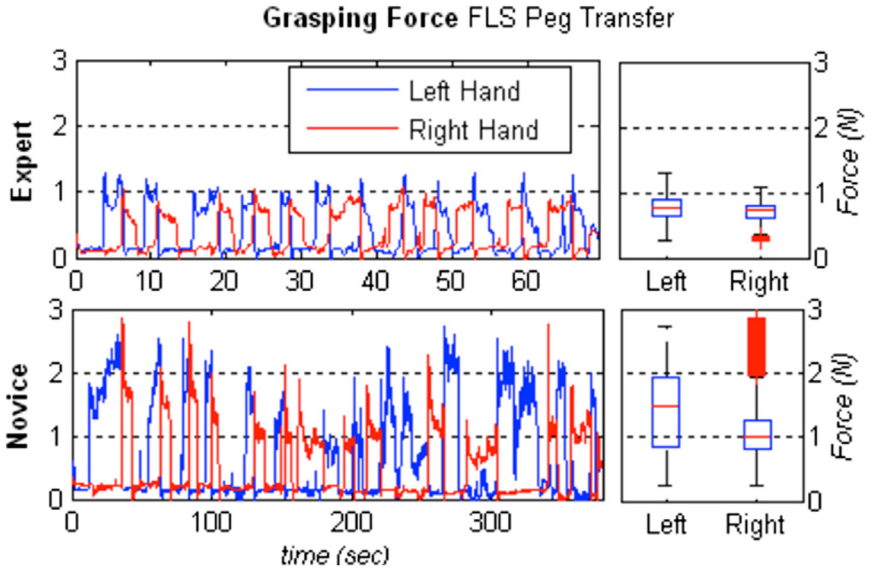

Fig. 4 – . The grasping force *versus* time for left (blue) and right (red) hands for the same PegTx task iteration shown in Fig. 3. Grasping force units are in tens of Newtons. (Color version of figure is available online.)

Table 5 – Correlation between FLS score and demographically based possible construct validity categories by task for the entire database ($n = 447$). Spearman ρ and (*P* value) are shown.

FLS *versus*	PegTx	Cutting	Suturing
LAPR cases	0.50 (0.00)	0.59 (0.00)	0.33 (0.00)
Training level	0.48 (0.00)	0.63 (0.00)	0.41 (0.00)
Age	0.17 (0.02)	0.32 (0.00)	0.07 (0.52)

surgeon) as well as the self-reported estimate of all-time total number of laparoscopic cases performed. Only practicing laparoscopic surgeons and fellows who had completed >100 laparoscopic cases (where they did ≥50% of the case) were considered candidate subjects for the "true expert" candidate subject pool. Second, the highest FLS-scoring iteration for each task of each expert candidate was taken as a set of ;true expert; candidate iterations. Finally, the third approach used an Objective Structured Assessment of Technical Skill (OSATS) protocol [4,6], which was modified to focus exclusively on psychomotor skills, denoted p-OSATS and shown in Table 2. Only the "true expert" candidate iterations were considered for p-OSATS review. The videos of these iterations were randomly renamed and reordered before evaluation. Two faculty surgeons (coders A and B) served as p-OSATS reviewers. Reviewers were blind to the identity and demographics of the subjects whose videos they reviewed and to the scores of the other review sessions. Only the ;true expert candidate; iterations that received p-OSATS scores ≥3 in all domains were included in the "true expert" set. This set consists of single iterations, not individuals. In this way, the three criteria for identifying expert skill such as training or experience level, validated performance measures, and OSATS review were combined.

To address intersession reliability, coder A reviewed all videos again, but in a different randomized order approximately 10 d after first review. Coder B's scores were compared with coder A's combined scores for inter-rater reliability. Intra-rater reliability also was examined by comparing a coder's score of a given video the first and second time. Both coders were asked to limit their coding to psychomotor characteristics of the tool motion in p-OSATS alone, but were repeatedly asked to verbalize their thoughts during review, which were subsequently recorded for each reviewed iteration.

3. Results

3.1. *Reliability of p-OSATS and overview of data collection*

When considering subjective evaluations, if Pearson $r > 0.5$ is considered a strong fit and $0.4 < r < 0.5$ is considered a moderate fit [26], then each p-OSATS domains exhibited acceptable reliability ($P < 0.05$) for the PegTx and suturing tasks, with strongest reliability for the "motion quality" p-OSATS domain. Intersession and inter-rater reliability were weakest for the cutting task. See Table 3 for details.

Not all subjects completed all requested iterations in the study. Some subjects voluntarily completed additional iterations. Incomplete iterations or iterations with corrupted data such as missing video, which prevented post-task scoring were excluded from analysis. An overview of the collected data appears in Table 4 (the suturing task was not available at the UW site). Representative plots of 3D tool path of an FLS novice (low-scoring, no FLS experience) and proficient subject (high FLS-scoring, faculty surgeon) appear in Figure 3 along with corresponding scores for a single PegTx iteration. Figure 4 shows the left and right hand grasping force plotted in time for the same iteration as well as the corresponding computed values.

3.2. *Demographics and FLS scores*

Demographic categories typical for construct validity like training level and laparoscopic experience do not correlate well over the entire database ($n = 447$) Table 5. Spearman $\rho < 0.5$ for all but the cutting task *versus* training level and laparoscopic case count. The corresponding scatter plots

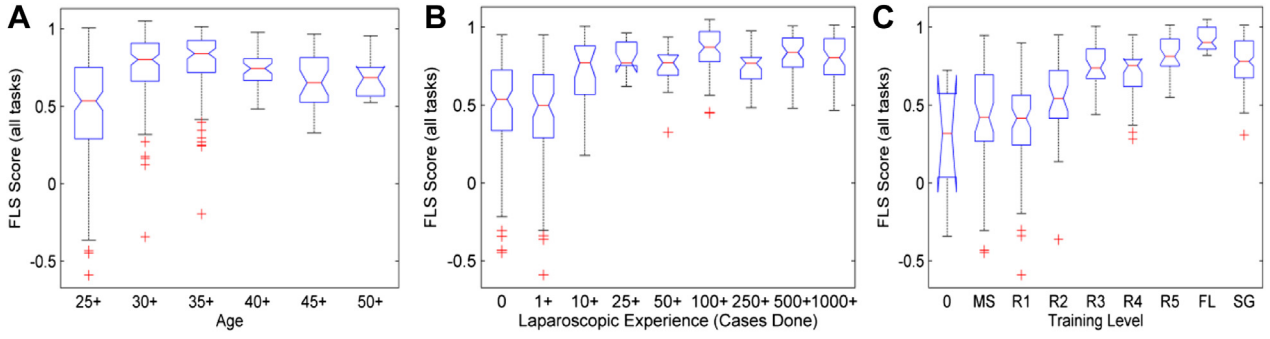

Fig. 5 – Box plots of FLS scores *versus* demographic categories. Such categories are often used to establish construct validity. (A) Age: 25 + indicates subjects who are aged ≥25 y, and so forth. (B) Laparoscopic Experience: the self-reported estimate of lifetime total number of laparoscopic cases performed. 1 + indicates subjects who had performed ≥1 cases, where they did ≥50 of the case. (C) Training Level (0 = none, MS = medical school, R1-5 = post graduate year of surgical residency, FL = fellow, SG = practicing surgeons). (Color version of figure is available online.)

Table 6 – Comparison of typical construct validity criteria among only the "True Expert" candidate iterations ($n = 56$). Pearson r and Spearman ρ with corresponding (P value) are shown.

Comparison	Correlation coefficient	PegTx	Cutting	Suturing
Laparoscopic experience *versus* p-OSATS	r	0.15 (0.52)	−0.01 (0.95)	−0.11 (0.68)
	ρ	0.25 (0.29)	0.11 (0.65)	−0.32 (0.22)
Laparoscopic experience *versus* FLS	r	0.09 (0.71)	−0.21 (0.38)	−0.31 (0.24)
	ρ	−0.09 (0.72)	−0.09 (0.72)	−0.16 (0.56)
p-OSATS *versus* FLS	r	0.57 (0.01)	0.57 (0.01)	0.79 (0.00)
	ρ	0.55 (0.01)	0.64 (0.00)	0.77 (0.00)
Time *versus* p-OSATS	r	−0.57 (0.01)	−0.56 (0.01)	−0.79 (0.00)
	ρ	−0.55 (0.01)	−0.64 (0.00)	−0.77 (0.00)
Time *versus* FLS	r	**−1.00 (0.00)**	**−1.00 (0.00)**	**−1.00 (0.00)**
	ρ	**−1.00 (0.00)**	**−1.00 (0.00)**	**−1.00 (0.00)**

Exceptionally high correlations are denoted in bold.

(Fig. 5) indicate significant variation of FLS scores in the subjects with most training and most experience: demographically identified experts exhibited FLS scores significantly lower than fellows and not significantly different than third- or fourth-year residents. FLS scores correlated most weakly or not at all across age.

3.3. *Comparison of criteria used to establish the "true experts" set*

All iterations of "true expert" candidates ($n = 52$) received p-OSATS scores in addition to their FLS scores and demographic ranking data. For this pool, no demographic category correlated well with either FLS or p-OSATS scores. However, FLS and p-OSATS correlated well ($P < 0.01$) for all tasks (Table 6). Additionally, each individual p-OSATS domain (not shown in table) correlated significantly with FLS ($P < 0.01$) but none did so with laparoscopic experience.

FLS and p-OSATS were further compared with simple task time among the "true expert" candidates. FLS score was found to be nearly identical to time ($r = -1.00$, $P < 0.0000001$) indicating that the "true expert" candidates made virtually no mistakes. p-OSATS correlated well with time, though exhibited some deviation. The correlation details for each task are listed in the lower section of Table 6. Moreover, Figure 6A illustrates the specific deviations between p-OSATS and FLS; Figure 6B shows the near-perfect equivalence of FLS and time; and Figure 6C illustrates the deviation between p-OSATS scores and time.

3.4. *Comparison of FLS and alternative objective metrics among all subjects*

All iterations from all skill levels ($n = 447$) were used to compare the automated objective performance metrics (task time, path length, EoM, and peak grasp force) and FLS scores. Correlation details appear in Table 7. Figure 7 indicates near-perfect correlation between FLS and task time. Path length correlates well with task time for all tasks (with correlation coefficients at ≥ 0.87 with significance $P < 0.0001$ or better). EoM shows substantially less correlation with FLS scores, especially for lower EoM values. Finally, peak grasping force F_{peak} shows no discernible overall trend across tasks.

A closer view of peak grasping force and FLS scores for only PegTx and suturing tasks appears in Figure 8. Iterations with the high FLS scores exhibited a wide range of peak grasping forces, especially for PegTx and suturing.

4. Discussion

FLS has become the common language that surgical educators use to measure laparoscopic surgical skills. The value of FLS has been extensively demonstrated in dry laboratory and operative training curricula, and was the first widely adopted system to attempt to quantitatively measure the quality of surgical skills [12]. One limitation of FLS, however, is that scoring is heavily based on time to complete the tasks. *In vivo* surgical performance should not just be graded by the time it takes to do a case or surgical step but also precision measures such as how carefully the surgeon interacts with tissues. By "teaching to the test," trainees may learn that time is more important than precision or quality of movement, particularly if penalties for lack of precision are low or do not exist. This may be evidenced by the fact that, on specific FLS tasks non-clinicians may attain better ;scores; than experienced laparoscopists not familiar with the tasks because the scoring

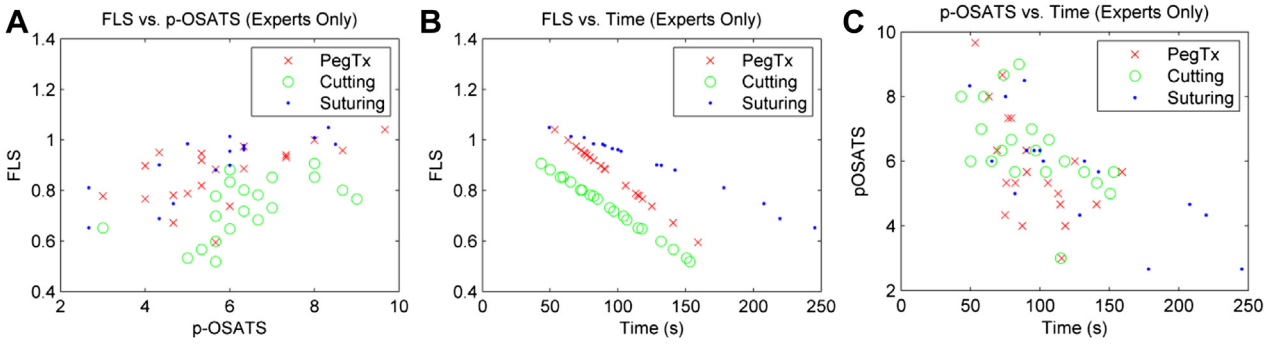

Fig. 6 – Scatter plots of only the iterations in the "true expert" set for (A) FLS scores *versus* p-global OSATS score, (B) FLS *versus* Time, and (C) p-OSATS *versus* Time. FLS shows little or no statistical deviation from time indicating that the experts performed minimal errors; however, p-OSATS varied substantially from time. (Color version of figure is available online.)

Table 7 – Correlation of FLS scores with objective performance metrics over all subjects of all skill levels. Pearson r and Spearman ρ with corresponding (P value) are shown.

FLS score *versus*	Correlation coefficient	PegTx	Cutting	Suturing
Time (s)	*r*	**−0.99 (0.00)**	**−1.00 (0.00)**	**−1.00 (0.00)**
	ρ	**−0.99 (0.00)**	**−1.00 (0.00)**	**−1.00 (0.00)**
Path length (cm)	*r*	−0.92 (0.00)	−0.87 (0.00)	−0.95 (0.00)
	ρ	−0.89 (0.00)	−0.88 (0.00)	−0.93 (0.00)
EoM (cm/s)	*r*	0.70 (0.00)	0.35 (0.00)	0.40 (0.00)
	ρ	0.82 (0.00)	0.42 (0.00)	0.61 (0.00)
Peak force (N)	*r*	**−0.18 (0.01)**	**−0.27 (0.00)**	**−0.16 (0.14)**
	ρ	**−0.08 (0.30)**	**−0.27 (0.00)**	**−0.26 (0.01)**

Extreme values are indicated in bold.

heavily weights task time. Furthermore, the advantage of adding additional performance metrics that correctly measure precision is that this may allow for more granular proficiency discrimination and provide educators with additional data to tailor surgical curricula.

4.1. *Establishment of "true experts" group: demographic criteria are insufficient*

In the surgical simulation literature, a single criterion is commonly used to group subjects into skill levels like "expert" and "novice". In this study, we used three criteria and combined them together with a goal of more accurately identifying performance level. The three criteria used were p-OSATS review, demographically identified skill via training level or laparoscopic experience, and validated FLS scoring methodology. In many cases, these measures did not correlate favorably (Table 5). Particularly, greater laparoscopic experience did not strongly imply better FLS or p-OSATS scores especially for PegTx and suturing. If demographic markers like experience and training level are taken as the most credible basis for measuring expert skill, this result suggests that the FLS tasks themselves may not relevantly test surgical expertise, at least not among higher skill levels. This may be due to the fact that FLS was designed to reflect basic proficiency rather than true expertise. As a primarily time-based measure, FLS likely has a ceiling effect. More-expert surgeons may not necessarily be faster surgeons. If FLS and p-OSATS are taken as ground truth, this suggest that perhaps demographic categories like experience or training level provide imperfect or even poor means to establish construct validity in simulation studies. However, high-stakes FLS pass or fail thresholds are set based on the performances of a group of "experts" defined by such demographic categories [27]. If this demographic selection is not optimal, the resulting high-stakes thresholds may not be optimal.

4.2. *FLS score effectively equivalent to task time alone*

According to the FLS score equations of Table 1, strict correlation between FLS and task time is required if and only if there are no errors. According to our results in Table 7, the correlation between FLS and task time is never substantially reduced by errors for any task for all iterations from all skill levels in the database. This shows that the existing error penalties have a negligible effect on the final score in practice. This implies that FLS scoring (Table 1) overemphasizes time to the detriment of precision. Also, the resource cost incurred by manually collecting and computing the penalties may not be justified given this negligible impact. Adjustment of the weights used to compute error penalties (coefficients of *E* in Table 1) may address these issues. Because we expect "true experts" to have a near-zero FLS error rates, this near-perfect correlation between task time and FLS score is not surprising among them (shown in Fig. 6B and Table 6) and indicates that with FLS scoring, only task time can discriminate skill among experts. However, p-OSATS scores of "true expert" iterations deviated substantially from task time (Fig. 6A,C and Table 6). If the p-OSATS scores are accurate, this suggests that they discriminate some dimension of psychomotor skill among experts not detectable with task time or existing FLS penalties.

If two objective measures of psychomotor skill each discriminate skill but correlate very highly with each other, this suggests one is redundant and unnecessary: they both measure the same aspect of psychomotor skill. If they discriminate skill but do not correlate strongly, this suggests they measure different aspects of psychomotor skill and that both provide meaningful, unique information. According to Table 7, path length and to a lesser extent EoM correlated favorably with FLS (and time, implicitly). It is reasonable to expect that longer path lengths for a given procedure will result in longer times and hence be closely dependent. Thus, path length may provide little additional information to task

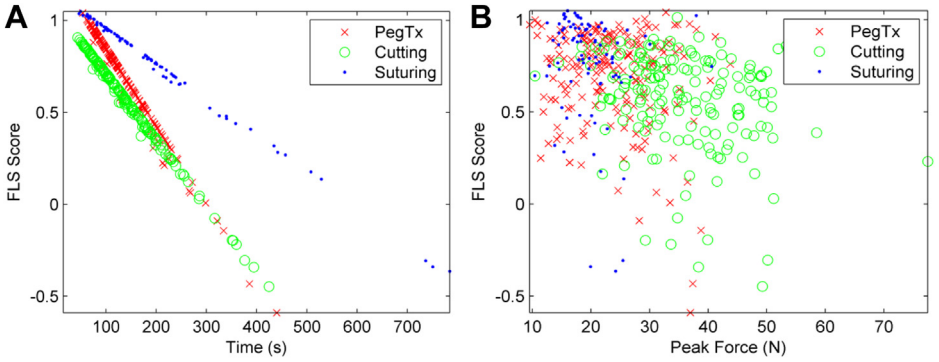

Fig. 7 – FLS scores *versus* (A) task time and (B) peak grasping force for each of the three tasks for all subjects of all skill levels. (Color version of figure is available online.)

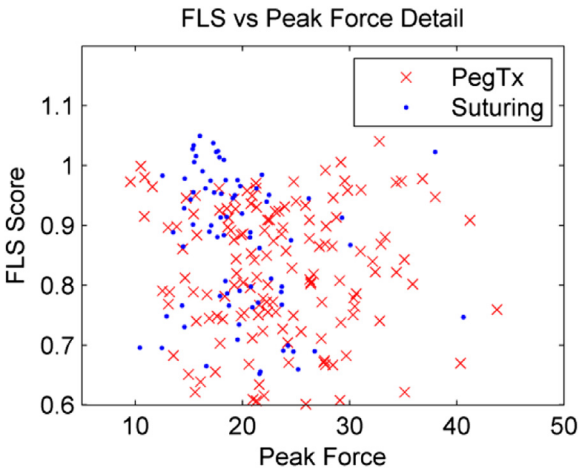

Fig. 8 – Peak grasping force behavior for PegTx (red x) and suturing tasks (blue dots) near highest FLS scores. (Color version of figure is available online.)

time. However, EoM, which is merely an average tool velocity in a task iteration (units: centimeter per second), specifically controls for task time. If it correlated perfectly with task time, this would imply that it is unnecessary. Instead it appears to measure, at least modestly, some other aspect of psychomotor skill, an aspect that may be illustrated qualitatively in the pattern differences captured in Figures 3 and 4.

In real surgery, excessive grasping force induces tissue damage [20] and thus should be considered a teaching goal of a basic laparoscopic psychomotor skills curriculum. Currently, the FLS methodology provides no way to deter or penalize such behavior, nor does the technical skills training outside the operating room or surgical laboratory (where tissue damage from excessive force is more evident) include targets or instructions regarding grasping force in relation to tissue handling skills. We expect this lack of emphasis on grasping force, at least in part; results in the widespread lack of correlation shown between peak grasp force and FLS scores in Figure 7 and Table 7. This lack of correlation, coupled with the clinical relevance of excessive peak grasp force, indicates that peak grasp force provides an additional, unique measure of psychomotor skill over task time. FLS currently provides an effective penalty for too little grasping force because using inadequate force may accrue error penalties such as drops or additional time due to insufficient tissue traction. However, FLS provides no penalty for excessive grasping force. We speculate that indiscriminate use of high forces may in some cases improve task times in FLS as one is penalized for "dropping" a block during Peg transfer and therefore squeezes with excessive force to prevent this error. Figure 7B shows incidence of high relative peak force among high FLS scores, for example, in "good" scores >0.8.

There remain several limitations to the use of grasping force as a skill indicator. Creating a penalty based on peak grasp force requires establishing a clinically relevant threshold of peak force to avoid tissue damage. Such values have been established in live porcine models [20] for various tissues. Incorporating such a peak force threshold for a penalty score in FLS will require extending this work to human tissues and a wider consensus on what target tissue should be used to determine the threshold. More importantly, the measurement of peak-grasping force is not a direct measure of tissue damage: it promises to establish a correlation with tissue damage, but not equivalence. Peak grasp force is favorable over directly measuring tissue damage only due to the lower resource cost and easier availability with exiting tools like EDGE or even surgical robots. Thus, although peak grasping force provides additional value to metrics like task time, it should only be used in combination with such measures, not exclusively.

A disadvantage of FLS methodology is that the testing procedure requires the presence of a trained proctor. The test is then scored by a proprietary formula, requiring the proctor to send the test materials to a central source for grading. Reporting of the score is therefore delayed, and of little value in providing proximate feedback to the learner [17]. Although surrogate measures have been published that can be used for immediate feedback and training, these surrogate measures are still based primarily on time and not quality. Computerized scoring has better accuracy and repeatability and is more efficient than FLS scoring because it eliminates evaluator time, bias, and human error and can present evaluation feedback instantly. Excessive grasping force behavior can be easily identified and appropriately addressed (e.g., a score penalty or automated audio feedback) along with precision variables such as EoM. Such accuracy is particularly required for the force measurements that are difficult to quantify from video. However, more research is required to establish clinically relevant grasping force targets for the different tasks or targeted tissues.

Several limitations are present in this study. Data collection occurred in variable environments. The FLS tasks themselves were deployed in the EDGE platform and not the box trainer provided by the FLS program, resulting in a modified setup and video lighting conditions. The cutting task varied slightly for one site in that a double ring was used for the UW site and a single ring was used everywhere else, this potentially resulted in different scoring criteria between sites for the cutting task. Although surgeons and surgical trainees tend to be naturally competitive, for this study they voluntarily participated when they had time and under different circumstances of operator fatigue, causing unmeasured variation in attention to the task and performance. Interobserver reliability between coders in the p-OSATS video review could be improved. Post analysis of recorded reviewer comments indicated differences in the scoring evaluation for handling of the gauze between coders.

To our knowledge, all prior studies indicate FLS to be valid and show positive transfer of skills to the operating room. We do not see any contradictions of this in our results. However, our work suggests that task duration is the predominant independent metric in the FLS scoring system, and that errors as a measure of "precision" contribute less value within the currently published FLS scoring system. Given the very visible success of the FLS program, these data suggest that alternative metrics beyond task time have potential to broaden the significance of quality testing in the FLS tasks. Engaging subjects to more accurately balance task time with technical precision, as is necessary in any clinical procedure, will improve the

relevance of FLS scoring. Based on the data presented, adjusting the weights in FLS scoring equations and using additional, nonredundant objective metrics such as peak grasping force all offer a more nuanced and likely, accurate assessment of laparoscopic skills that will help keep FLS current and relevant.

Acknowledgment

The authors are indebted to UMN SimPORTAL, Brian Ross, and John Paige and their staff for generously facilitating data collection at their simulation centers and surgical sites; and the many volunteer surgeons who participated in our study.

Authors' contributions: T.M.K., T.S.L., A.W., B.H., and M.N.S. contributed to the study design. T.M.K., L.W.W., T.S.L., I.S.J., R.S., and A.W. did the data collection. T.M.K., T.S.L., and B.H. did the data analysis. T.M.K., L.W.W., and I.S.J. did the writing of the article. T.M.K., L.W.W., T.S.L., R.S., A.W., B.H., and M.N.S. did the critical revisions of the article. B.H., A.W., T.M.K., R.S., M.N.S. did the funding of the resources.

Disclosures

Authors T.M.K., L.W.W., and I.S.J. have received partial support in the past from Simulab Corp through Prof B.H. research laboratory. Profs B.H. and M.N.S. are co-inventors of technology licensed by Simulab and may receive future royalties.

REFERENCES

[1] Heinrichs L, Lukoff B, Youngblood P, et al. Criterion-based training with surgical simulators: proficiency of experienced surgeons. JSLS 2007;11:273. Available from, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3015829/.

[2] Sturm L, Windsor J, Cosman P, Cregan P, Hewett P, Maddern G. A systematic review of skills transfer after surgical simulation training. Ann Surg 2008;248:166.

[3] Fried G, Feldman L, Vassiliou M, et al. Proving the value of simulation in laparoscopic surgery. Ann Surg 2004;240:518.

[4] Martin J, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997;84:273. Available from, http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2168.1997.02502.x/abstract.

[5] Peters J, Fried G, Swanstrom L, et al. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery 2004;135:21. Available from, http://www.astec.arizona.edu/sites/astec.arizona.edu//pdf_files/%20Program.pdf.

[6] 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. Available from, http://www.sciencedirect.com////.

[7] Okrainec A, Soper N, Swanstrom L, Fried G. Trends and results of the first 5 years of Fundamentals of Laparoscopic Surgery (FLS) certification testing. Surg Endosc 2011;25:1192.

[8] Derossis A, Bothwell J, Sigman H, Fried G. The effect of practice on performance in a laparoscopic simulator. Surg Endosc 1998;12:1117.

[9] Derevianko A, Schwaitzberg S, Tsuda S, et al. Malpractice carrier underwrites fundamentals of laparoscopic surgery training and testing: a benchmark for patient safety. Surg Endosc 2010;24:616.

[10] Kolozsvari N, Kaneva P, Brace C, et al. Mastery versus the standard proficiency target for basic laparoscopic skill training: effect on skill transfer and retention. Surg Endosc; 2011:1.

[11] Rosenthal M, Ritter E, Goova M, et al. "Proficiency-based fundamentals of laparoscopic surgery skills training results in durable performance improvement and a uniform certification pass rate". Surg Endosc 2010;24:2453.

[12] Sroka G, Feldman L, Vassiliou M, Kaneva P, Fayez R, Fried G. Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. Am J Surg 2010;199:115.

[13] Gallagher A, Ritter E, Champion H, et al. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. Ann Surg 2005;241:364.

[14] Pearson A, Gallagher A, Rosser J, Satava R. Evaluation of structured and quantitative training methods for teaching intracorporeal knot tying. Surg Endosc 2002;16:130.

[15] Satava R, Cuschieri A, Hamdorf J. Metrics for objective assessment. Surg Endosc 2003;17:220.

[16] Seymour N, Gallagher A, Roman S, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 2002;236:458. Available from, https://.ncbi.nlm.nih.gov////.

[17] Ritter E, Scott D. Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery. Surg innovation 2007;14:107.

[18] Gunther S. Red dragon: a multi-modality system for simulation and training in minimally invasive surgery. Master's thesis. University of Washington; 2006.

[19] Rosen J, Brown J, Chang L, Barreca M, Sinanan M, Hannaford B. The BlueDRAGON—a system for measuring the kinematics and dynamics of minimally invasive surgical tools in-vivo, in Robotics and Automation, 2002. *Proceedings. ICRA'02. IEEE International Conference on*, 2. IEEE, 2002, pp. 1876–1881, blue dragon.

[20] De S, Rosen J, Dagan A, Hannaford B, Swanson P, Sinanan M. Assessment of tissue damage due to mechanical stresses. Int J Robotics Res 2007;26:1159.

[21] Rosen J, Brown J, Chang L, Sinanan M, Hannaford B. Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. Biomed Eng IEEE Trans 2006;53:399.

[22] Wright AS, Kowalewski TM, Hannaford B. Novel laparoscopic box trainer with integrated force and positioning sensors. In: 12th World Congress of Endoscopic Surgery. National Harbor, MD: Emerging Technology Session; 2010.

[23] Derossis M, Anna M, Fried M, et al. Development of a model for training and evaluation of laparoscopic skills. Am J Surg 1998;175:482.

[24] Fraser S, Feldman L, Stanbridge D, Fried G. Characterizing the learning curve for a basic laparoscopic drill. Surg Endosc 2005;19:1572.

[25] Rasband W. Imagej, U. S. National Institutes of Health. 1997-2011, Bethesda, Maryland, USA, Available from: http://imagej.nih.gov/ij/.

[26] Cohen J. Statistical power analysis for the behavioral sciences. Lawrence Erlbaum; 1988.

[27] Fraser S, Klassen D, Feldman L, Ghitulescu G, Stanbridge D, Fried G. Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. Surg Endosc 2003;17:964.

[28] Scott D, Ritter E, Tesfay S, Pimentel E, Nagji A, Fried G. Certification pass rate of 100% for fundamentals of laparoscopic surgery skills after proficiency-based training. Surg Endosc 2008;22:1887.