Surg Endosc (2017) 31:3883–3889
DOI 10.1007/s00464-017-5417-0

Video assessment of laparoscopic skills by novices and experts: implications for surgical education

Celine Yeung[1] · Brian Carrillo[2,4] · Victor Pope[3] · Shahob Hosseinpour[1] · J. Ted Gerstle[2,4] · Georges Azzie[2,4]

Received: 13 March 2016 / Accepted: 20 January 2017 / Published online: 15 February 2017

Abstract Background Previous investigators have shown that novices are able to assess surgical skills as reliably as expert surgeons. The purpose of this study was to determine how novices and experts arrive at these graded scores when assessing laparoscopic skills and the potential implications this may have for surgical education.
Methods Four novices and four general laparoscopic surgeons evaluated 59 videos of a suturing task using a 5-point scale. Average novice and expert evaluator scores for each video and the average number of times that scores were changed were compared. Intraclass correlation coefficients were used to determine inter-rater and test–retest reliability. Evaluators were asked to define the number of videos they needed to watch before they could confidently grade and to describe how they were able to distinguish between different levels of expertise.
Results There were no significant differences in mean scores assigned by the two evaluator groups. Novices changed their scores more frequently compared to experts, but this did not reach statistical significance. There was excellent inter-rater reliability between the two groups (ICC = 0.91, CI 0.85–0.95) and good test–retest reliability (ICC > 0.83). On average, novices and experts reported that they needed to watch 13.8 ± 2.4 and 8.5 ± 2.5 videos, respectively, before they could confidently grade. Both groups also identified similar qualitative indicators (e.g., instrument control).
Conclusion Evaluators with varying levels of expertise can reliably grade performance of an intracorporeal suturing task. While novices were less confident in their grading, both groups were able to assign comparable scores and identify similar elements of a suturing skill as being important in terms of assessment.

Keywords Video assessment · Suturing skill · Laparoscopic · Novice evaluators

The objective structured assessment of technical skills (OSATS) is the current gold standard for measuring surgical performance [1]. While this evaluation method has been widely adopted, it is time-intensive and expensive as only senior surgeons are able to evaluate trainees [2]. With the introduction of competency-based model of education, there is an increasing demand to assess the technical skills of trainees, and in the future, such assessment could extend to recertification of licensed surgeons [3, 4]. The shortage of expert surgeons, coupled to the challenge of teaching surgical trainees and providing objective assessment with finite faculty time has also become increasingly important [5]. In order to address this issue, previous investigators have explored the use of crowdsourcing platforms in the assessment of surgical skills. Crowdsourcing is the practice of obtaining ideas or services by enlisting contributions from a large group of people, who are generally non-experts from the online community. This practice has been applied in a variety of ways to solve different problems, such as deciphering medical cases through an online

✉ Georges Azzie
georges.azzie@sickkids.ca

1 Faculty of Medicine, University of Toronto, Toronto, ON, Canada

2 Division of General and Thoracic Surgery, The Hospital for Sick Children, Toronto, ON, Canada

3 Division of Otolaryngology, The Hospital for Sick Children, Toronto, ON, Canada

4 Centre for Image-Guided Innovation and Therapeutic Intervention, Toronto, ON, Canada

website [1], or training image classifiers in diagnostic radiology [6]. Applying crowdsourcing to the assessment of surgical skills has led to the advent of a novel evaluation method. This technique involves a large group of novices evaluating videotapes of surgical skills using online grading tools such as "Crowd-sourced assessment of technical skills (CSATS)" [1].

Blinded video assessment of surgical skills has also been gaining attention independent of crowd-sourced assessments. Previous investigators have found video assessment to be cost-effective, efficient, and a less biased means of assessing surgical skills [7, 8] relative to traditional methods that rely on the subjective opinions of faculty surgeons [6, 8]. Birkmeyer et al. [9] utilized this evaluation method and also objectively showed that a higher skill rating was directly proportional to better patient outcomes (e.g., fewer postoperative complications, readmission rates, death). The authors demonstrated a wide variation in technical skill among practicing expert surgeons and suggested that video assessment may be used to provide surgeons with an objective measure of their expertise level as well as anonymous, constructive feedback to help experts refine surgical technique [9]. These findings lend greater support for the use of video analysis in crowd-sourced assessments of surgical skills among trainees and toward evaluating skills among practicing surgeons for the purpose of maintaining certification.

Previously, there was no consensus as to the level of expertise necessary among evaluators to assure reliable assessment. Chen et al. [1] and Holst et al. [10, 11] have shown that crowds consisting of novices are as effective as expert surgeons in the evaluation of surgical skills. While these results are promising, key questions remain such as (1) whether novices and experts assess the same elements of a performance, enabling novices to pair their quantitative assessments with expert-level qualitative feedback; (2) the number of videos evaluators need to watch before they can confidently compare performers; (3) the number of changes evaluators make during evaluation; and (4) the test–retest reliability of video assessment. This information could be useful for peer–peer or near–peer formative assessments, and support the use of evaluation methods like crowdsourcing, which may ultimately help address the shortage of expert evaluators available for training residents and recertifying practicing surgeons. For example, this may involve non-expert students providing interim evaluations to their peers throughout the year to improve their surgical proficiency, thereby decreasing educational demands on expert surgeons.

In this study, we first determined whether novices could grade performance of a defined suturing task within a simulator as reliably as experts. We then sought to gain insight into how novices and experts arrive at their scores when assessing laparoscopic skills, and to explore the potential implications of this evaluation method for surgical education.

Materials and methods

Evaluators

Four novices and four general surgeons assessed 59 videos of a standardized intracorporeal suturing task. The four novice evaluators did not have any medical training, varied in age (range 18–35), and had different occupational backgrounds (high school student, science graduate students, and software developer) while the expert evaluators consisted of a group of four general surgeons trained in laparoscopic surgery. Multiple evaluators were used as it improves reliability and decreases error related to the halo effect (an observer bias that occurs when a good or bad performance affects the observer's overall impression of the performer, such as their assessments of other performance domains) [12]. Evaluators were blinded to the performers' identity and were asked to grade overall level of expertise using a 5-point scale.

Description of videos

Fifty-one candidates performed a defined intracorporeal suturing task using a laparoscopic simulator at the 2012 combined International Pediatric Endosurgery Group and Society of American Gastrointestinal and Endoscopic Surgeons meeting. Fifty-nine videos were generated as four videos were repeated three times for the purpose of calculating test–retest reliability.

The videos began when the participant started attempting to pick up the needle (15 cm 3–0 suture) that was placed on a predetermined location of the simulator. The task also involved passing the needle through two target marks on either side of a slit penrose drain, completing a square double-throw knot, and tying two square single-throw knots. The video ended when the second single-throw knot was completed.

The videos were initially categorized based on the performers' self-reported number of laparoscopic procedures performed per year (0–9 procedures = novice; 10–50 procedures = intermediate; >50 procedures = expert) and were randomized into three different sets (two sets of 20 videos and one set of 19 videos). The order of the videos within each set was randomized and all 59 videos were given to each evaluator.

Laparoscopic video grader

A custom program was written in Java, using the Java FX video library (laparoscopic video grader) and given to each evaluator alongside an instruction manual on how to use the program. Each evaluator used the laparoscopic video (LV) grader to view and grade the videos at any point during each performance (Fig. 1). Evaluators could not speed up the performance or skip ahead to the next video until a score was assigned. However, they were allowed to re-watch videos and re-assign scores at any given time.

The program recorded the times at which evaluators graded each video and when evaluators decided to change their scoring. From this, the number of changes per evaluator across the 59 videos was obtained.

Follow-up questionnaire

After assessing all the videos, evaluators were asked to complete a short questionnaire. This involved asking them to approximate the number of videos they needed to watch in order for them to confidently judge the level of expertise. This question was used to complement the number of times scores were changed while grading. This was then followed by an open-text question, asking evaluators to list five points, describing how they were able to differentiate between novices, intermediates, and experts.

Study design and statistical analysis

A mixed methods study design was used. Institutional ethics was received for the data collected. Quantitatively, average evaluator scores, the number of changes in scoring, and the number of videos evaluators needed to watch before assigning grades were calculated. The latter two were classified as indirect measures of evaluator confidence in assigning scores. Qualitatively, evaluators were asked to describe how they were able to differentiate between varying levels of expertise.

Independent Student's *t* tests were used to compare mean scores between the novice and expert evaluator groups, the average number of changes between the two groups, and the number of videos evaluators needed to watch before they could confidently assign scores. Intraclass correlation coefficients were calculated to determine inter-rater reliability (degree to which evaluators coincide with their ratings) and test–retest reliability (degree to which evaluators are able to reproduce similar scores when assessing repeated performances) [13]. All statistical tests were performed using a SPSS statistical software package (version 22.0; SPSS, Chicago, IL, USA).

A researcher triangulation method was used to analyze the responses from the questionnaire. Three reviewers coded the questionnaire responses manually. The data were then analyzed using an inductive approach to determine whether there were any common, emerging themes between the two evaluator groups.

Results

For each performer group, there was no significant difference between the scores assigned by novice and expert evaluators (Fig. 2). The mean scores ± SEM assigned by

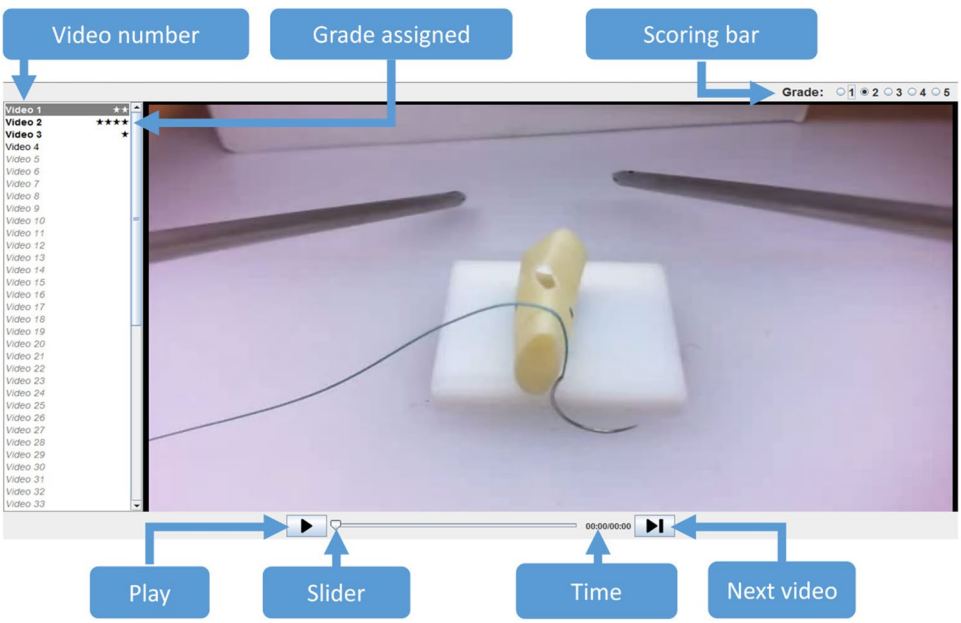

Fig. 1 A screenshot of the laparoscopic video grader program that evaluators used to assign grades

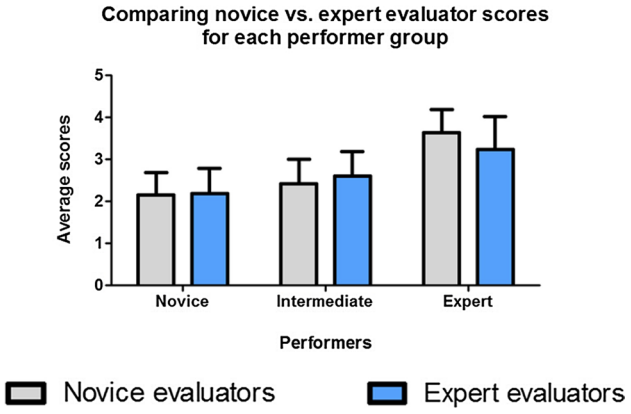

Fig. 2 There was no significant difference between the scores assigned by novice and expert evaluators for each performer group. The mean scores ± SEM assigned by novice evaluators were 2.2 ± 0.27, 2.4 ± 0.29, and 3.6 ± 0.28 for the novice, intermediate, and expert performer groups, respectively, while those assigned by surgeons were 2.2 ± 0.30, 2.6 ± 0.30, and 3.2 ± 0.39, respectively

novice evaluators were 2.2 ± 0.27, 2.4 ± 0.29, and 3.6 ± 0.28 for the novice, intermediate, and expert performer groups, respectively, while those assigned by surgeons were 2.2 ± 0.30, 2.6 ± 0.30, and 3.2 ± 0.39, respectively.

Experts changed their scores fewer times than novice evaluators across the 59 videos (33.5 ± 6.5 for novices vs. 16.8 ± 4.2 for experts), though this difference did not reach significance ($p = 0.08$) (Table 1). When asked how many videos evaluators needed to watch before they could grade confidently, novice evaluators reported an average of 13.8 ± 2.4 videos (range 10–20), and experts reported an average of 8.5 ± 2.5 (range 5–15). The difference between novices and experts was not significant ($p = 0.38$) (Table 1).

Inter-rater reliability between the two evaluator groups was excellent (ICC = 0.91, CI 0.85–0.95), and exceeded the threshold of desired reliability for high-stakes testing (ICC > 0.80) [4, 10]. For the performances that were repeated, there was good test–retest reliability for both the novice (ICC = 0.85) and expert evaluators (ICC = 0.83), indicating that evaluators were able to reproduce similar scores after reviewing videos that were repeated.

Inductive data analysis of the questionnaires revealed differences in how the evaluators phrased their observations of the performances. Though the novices did not use the same words or phrases as the experts to comment on the performance, emerging themes shared by the novice and expert evaluators were evident. For example, both evaluator groups were able to differentiate the expertise level by assessing the quality of instrument control.

> Novice: "...subject showed awareness in the placement of the tools but projected jerky movements"
> Expert: "rotating the needle grasper with the needle to smoothly place the surgeon's knot" or "efficiency of movements"

Other themes that were shared by the novice and expert evaluators included evaluating accuracy of passing the needle through small black targets, the quality of the final knot, and whether the performers demonstrated respect for the penrose (Table 2). While there were notable similarities, novices concentrated more on the total time taken to complete the task and whether the performers were able to recover from their mistakes. Meanwhile, experts focused more on details such as the initial position of the needle relative to the needle driver.

Discussion

This study sought to understand how novices and experts assess performance of a defined intracorporeal suturing task. In recent studies, investigators showed that novices were able to evaluate surgical skills as reliably as expert surgeons. However, these studies were limited by the number of videos that were given to novices for evaluation (range 1–12 videos) [1, 2, 6, 10, 11, 14]. Only one recent study used over 20 videos for a more complex cricothyrotomy procedure [14]. In our study, evaluators were asked to watch 59 videos and discriminate varying levels of expertise between each performer. The large number of videos allowed novices and experts to approximate the number of videos they needed to compare before they

Table 1 Average number of changes made by novice and expert evaluators (mean ± SEM), and the number of videos the evaluators needed to watch before they felt confident with grading

		Novice evaluator	Expert evaluator	*p* value
Average # of changes across videos	Novice performers	10 ± 3	4.3 ± 1.1	0.13
	Intermediate performers	8.5 ± 0.6	5.3 ± 1.8	0.14
	Expert performers	15 ± 3.6	7.3 ± 2.8	0.14
	Overall average	33.5 ± 6.5	16.8 ± 4.2	0.08
Videos needed to be watched before grading confidently	Average	13.8 ± 2.4	8.5 ± 2.5	0.38
	Range	10–20	5–15	

Table 2 Qualitative indicators identified by novice and expert evaluators for an intracorporeal suturing task

In order of frequency	Novice	Expert
Most frequent	Time—"…shorter times associated with increasing levels of training"	Correct positioning—"how the needle is initially positioned in the needle driver of the dominant hand"
	Accuracy—"accuracy of the needle in reference to the black dots"	Final knot—"flat knot that is perpendicular to the axis of the penrose drain"
	Instrument control—"…showed awareness in the placement of the tools but projected jerky movements"	Respect for the tissue—"…bracing the penrose as the needle goes through; no excessive force"
	Final knot—"…finished knot is messy (e.g., bow tie knot rather than a clean knot)"	Accuracy—"accuracy passing needle through black dots"
Least frequent	Respect for the tissue—"…tugging string and tube" or "…pulling thread through thoroughly before beginning"	Instrument control—"Rotating the needle grasper with the needle to smoothly place a surgeon's knot"

could confidently grade. On average, novices indicated that they needed to watch 13.8 ± 2.4 videos before they could confidently assign scores. This number was higher than the average number indicated by experts (8.5 ± 2.5 videos). However, asking evaluators to approximate the number of videos they needed to watch is a subjective measure. Thus, this result was further supported by objective data obtained using the LV Grader.

The LV Grader was able to identify the number of times evaluators changed their grades, which served as an indirect measure of confidence, presuming that the greater the number of changes, the less confident the evaluator was in his or her ratings. In this study, novices made more changes than experts, consistent with the higher self-reported number of videos needed to be watched, though these differences did not reach significance. Changing scores throughout the grading process also suggests that evaluators assigned scores by comparing performers relative to one another. This concept is corroborated by Malpani et al. [6], who established a basis for assessing surgical skills using pairwise comparisons of different segments of a suturing task. They also showed that novices can assess segments of a performance as reliably as expert surgeons and suggest that novice evaluators can be trained to improve accuracy. Thus, both these findings, average number of videos needed to be watched and the number of changes, suggest that evaluators may be required to watch multiple performances prior to grading for more accurate and objective assessment. The results presented in this paper may help improve current crowd-sourced evaluation methods, providing insight into the number of videos that novices should watch before they assign grades (i.e., when novices are trained to assess surgical skills).

In this study, novices were also able to assign scores as reliably as expert evaluators (ICC = 0.91), which is consistent with previous studies that compared novice and expert evaluator scores [1, 2, 6, 10, 11, 14]. Unlike previous studies, test–retest reliability was also assessed; this was possible given the large number of videos used which allowed for the integration of repeated videos. With regard to the four videos that were repeated in this study, there was excellent test–retest reliability among the novices and the experts (ICC > 0.83). This demonstrates that both evaluator groups were able to reproduce similar scores for the same videos regardless of when the performances were reviewed. To further explore this phenomenon, a questionnaire was administered to elicit responses that may not have been captured by the descriptive statistics and reliability measures.

The participants' questionnaire responses may provide an explanation as to why no differences were seen quantitatively. The qualitative indicators that the novices used to support their evaluation were similar to those used by expert surgeons. Some of these indicators also appear on the validated OSATS scoring system [15], or are parameters that have been validated in previous studies (e.g., time and motion smoothness) [16]. Despite not having any surgical training, the novices were able to identify 3 of the 7 items on the OSATS (respect for the tissue, time and motion, and instrument handling) [15].

In addition, one astute novice evaluator described feeling confident when differentiating between novice and expert performers, but found it difficult to differentiate between performers with more similar skill sets.

> I felt confident differentiating between a novice and expert. However, as someone who has never seen this task before…it was very hard for me to differentiate between a 2 and a 3, a 4 and a 5, or a 2 and a 1.

This observation is consistent with the quantitative findings reported by Aghdasi et al. [14], who also found that novices had an easier time identifying highly skilled performers and greater difficulty differentiating between beginner or average-skilled performers.

This study provides insight into what novice and expert evaluators look for when evaluating an intracorporeal suturing task without any formal scoring system. While there were many similarities between the two groups, there were also notable differences. For example, novice evaluators

concentrated more on time taken to complete the suturing task, accuracy, and instrument control, whereas experts focused more on how the needle was loaded onto the needle driver, the quality of the final knot, and respect for the tissue. Novices also did not use the same surgical jargon as experts, which could affect the quality of feedback they provide; novices may not be able to help performers advance to the same degree as experts. Furthermore, if these comments were applied to practicing surgeons, they may be dismissed as it would be apparent that the evaluators are not expert surgeons. Future studies should assess what indicators expert evaluators value most or segments of the task to which they pay greater attention. Defining crucial components or segments of a task may help improve formative novice assessment (e.g., during near-peer teaching sessions) by training novices to focus on key components and provide feedback using the same surgical jargon as experts (e.g., clean knot = square knot). Future studies could also examine whether training novices to provide appropriate feedback could lead to performance improvement to the same extent as receiving expert feedback.

A limitation of this study is the use of a 5-point scale, which may not have been discriminating enough to separate subtle differences in evaluator responses. Use of a 7- or 10-point scale may have allowed for greater separation of values and a different mean for each evaluator group. However, regardless of scale size, qualitative indicators were remarkably similar between novice and expert evaluators, indicating that each group identified similar performance elements and supporting the consensus between group scores.

Another limitation of this study is that a single intracorporeal suturing task performed in a dry laboratory setting was tested. Surgery on real human tissue is much more complex. Having no knowledge of relevant anatomy or experience in different surgical approaches may limit novice evaluators. It is likely that, as the complexity of the procedure increases, expert evaluators may be required to discriminate between varying levels of expertise. Future studies should aim to identify ways to train novice evaluators to match the assessment skills of expert surgeons. Alternatively, finding ways to identify novice evaluators who demonstrate accurate scoring from among a larger group of novices may be worthwhile as there is currently no consensus on how much training is appropriate [17].

In conclusion, the group of novices in this study was as reliable as the expert evaluators in grading participants performing a defined intracorporeal suturing task. This was supported by similar qualitative indicators identified by both novice and expert evaluators. Novices seemed less confident in their scoring, which was demonstrated by the number of times they modified their grades and by the self-perceived number of videos needed to be watched before they could grade with confidence. The information gathered from this study may support the use of and further enhance evaluation methods like CSATS. For example, providing evaluators with a large number of videos and the ability to re-assign scores may help evaluators formulate scoring criteria and increase confidence among novices. Training novices using a standardized number of videos and qualitative indicators that have been validated by experts may also be the logical next step to improve current crowd-sourced assessment techniques. This may be beneficial not only for the purpose of formative assessment, but for teaching technical skills, especially in environments where expert input is a limited resource.

Acknowledgements This project was supported by the Comprehensive Research Experience for Medical Students (CREMS) program and the Department of Surgery at the University of Toronto. The authors would also like to acknowledge Dr. Paul Wales for providing us with his expertise in statistics and Dr. James Rutka for his continuous support of this project.

Funding This study was funded by the University of Toronto Comprehensive Research Experience for Medical Students and by the University of Toronto Department of Surgery.

Disclosures Celine Yeung, Dr. Brian Carrillo, Victor Pope, Shahob Hosseinpour, Dr. J. Ted Gerstle, and Dr. Georges Azzie have no conflicts of interest or financial ties to disclose.

References

1. Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, Kuksenok K, Aragon C, Holst D, Lendvay T (2014) Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 187(1):65–71
2. White LW, Kowalewski TM, Dockter RL, Comstock B, Hannaford B, Lendvay TS (2015) Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. J Endourol 29(11):1295–1301
3. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR (2010) The role of assessment in competency-based medical education The role of assessment in competency-based medical education. Med Teach 32(8):676–682
4. Meier AH, Gruessner A, Cooney RN (2016) Using the ACGME Milestones for resident self-evaluation and faculty engagement. J Surg Educ 73(6):e150–e157
5. Williams TE, Satiani B, Thomas A, Ellison EC (2009) The impending shortage and the estimated cost of training the future surgical workforce. Ann Surg 250(4):590–597
6. Malpani A, Vedula SS, Chen CCG, Hager GD (2015) A study of crowdsourced segment-level surgical skill assessment using pairwise rankings. Int J CARS 10(9):1435–1447
7. Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J, Reznick R, MacRae HM (2004) Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. Surg Endosc 18(12):1800–1804
8. Driscoll PJ, Paisley AM, Paterson-Brown S (2008) Video assessment of basic surgical trainees' operative skills. Am J Surg 196(2):265–272

9. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJO (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
10. Holst D, Kowalewski TM, White LW, Brand TC, Harper JD, Sorenson MD, Kirsch S, Lendvay TS (2015) Crowd-sourced assessment of technical skills: an adjunct to urology resident surgical simulation training. J Endourol 29(5):604–610
11. Holst D, Kowalewski TM, White LW, Brand TC, Harper JD, Sorensen MD, Truong M, Simpson K, Tanaka A, Smith R, Lendvay TS (2015) Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds. J Endourol 29(10):1183–1188
12. Thomas MR, Beckman TJ, Mauck KF, Cha SS, Thomas KG (2011) Group assessments of resident physicians improve reliability and decrease halo error. J Gen Intern Med 26(7):759–764
13. Gwet KL (2014) Handbook of inter-rater reliability. In: The definitive guide to measuring the extent of agreement among raters, 4th edn. Advanced Analytics, LLC, Gaithersburg
14. Aghdasi N, Bly R, White LW, Hannaford B, Moe K, Lendvay TS (2015) Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. J Surg Res 196(2):302–306
15. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchinson C, Brown M (1997) Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84(2):273–278
16. Hiemstra E, Chmarra MK, Dankelman J, Jansen FW (2011) Intracorporeal suturing: economy of instrument movements using a box trainer model. J Minim Invasive Gynecol 18(4):494–499
17. Scott DJ, Rege RV, Bergen PC, Guo WA, Laycock R, Tesfay ST, Valentine RJ, Jones DB (2000) Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv S 10(4):183–190