JOURNAL OF ENDOUROLOGY
Volume XX, Number XX, XXXXXX 2015

Pp. ■■■–■■■
DOI: 10.1089/end.2014.0616

Original Research

Crowd-Sourced Assessment of Technical Skills: An Adjunct to Urology Resident Surgical Simulation Training

Daniel Holst, BS,[1] Timothy M. Kowalewski, PhD,[2] Lee W. White, PhD,[3] Timothy C. Brand, MD,[4] Jonathan D. Harper, MD,[5] Mathew D. Sorenson, MD,[5] Sarah Kirsch, BS,[1] and Thomas S. Lendvay, MD[5]

Abstract

Background: Crowdsourcing is the practice of obtaining services from a large group of people, typically an online community. Validated methods of evaluating surgical video are time-intensive, expensive, and involve participation of multiple expert surgeons. We sought to obtain valid performance scores of urologic trainees and faculty on a dry-laboratory robotic surgery task module by using crowdsourcing through a web-based grading tool called Crowd Sourced Assessment of Technical Skill (CSATS).

Methods: IRB approval was granted to test the technical skills grading accuracy of Amazon.com Mechanical Turk™ crowd-workers compared to three expert faculty surgeon graders. The two groups assessed dry-laboratory robotic surgical suturing performances of three urology residents (PGY-2, -4, -5) and two faculty using three performance domains from the validated Global Evaluative Assessment of Robotic Skills assessment tool.

Results: After an average of 2 hours 50 minutes, each of the five videos received 50 crowd-worker assessments. The inter-rater reliability (IRR) between the surgeons and crowd was 0.91 using Cronbach's alpha statistic (confidence intervals = 0.20–0.92), indicating an agreement level between the two groups of ''excellent.'' The crowds were able to discriminate the surgical level, and both the crowds and the expert faculty surgeon graders scored one senior trainee's performance above a faculty's performance.

Conclusion: Surgery-naive crowd-workers can rapidly assess varying levels of surgical skill accurately relative to a panel of faculty raters. The crowds provided rapid feedback and were inexpensive. CSATS may be a valuable adjunct to surgical simulation training as requirements for more granular and iterative performance tracking of trainees become mandated and commonplace.

Introduction

As we enter a healthcare environment that is striving for quality improvement and shifting toward performance-based reimbursement, it is imperative that a feasible method of objective quantification of surgical skills be developed. In a recent report from Birkmeyer et al., the technical skill of practicing bariatric surgeons as assessed by blinded surgeon peers predicted patient outcomes. This was the first empiric evidence to demonstrate that video assessment of a surgeon's performance in their general practices correlated with postoperative complications, death, operative times, and readmission rates.[1] When looking at the role of surgical trainees in surgical malpractice claims data, Rogers et al. showed that errors in manual technique were present in 56% of cases.[2] In addition, surgical trainees contributed to errors in 46% of all cases and in 53% of those cases, a trainee was considered most responsible for the error.[2] Given such evidence suggesting that there is much room for improvement regarding surgical error, efforts have been made to adopt tools that objectively quantify surgical performance using global surgical performance-rating scales, like the Objective Structured Assessment of Technical Skills (OSATS) tool and its derivatives like the Global Evaluative Assessment of Robotic Skills (GEARS) or Global Operative Assessment of Laparoscopic Skills (GOALS) tools.[3,4] While

[1]University of Washington School of Medicine, Seattle, Washington.
[2]Department of Mechanical Engineering, University of Minnesota, Minneapolis, Minnesota.
[3]Stanford University School of Medicine.
[4]Department of Urology, Madigan Army Medical Center, Tacoma, Washington.
[5]Department of Urology, University of Washington, Seattle, Washington.

these methods are validated and considered the gold standard for evaluating surgical technical skill, they are infrequently used due to time-intensiveness, cost-inefficiency, and lack of standard practices. Subsequently, assessment of surgical trainees relies heavily on direct supervision and feedback given by preceptors, which may be subjective, unreliable, and inefficient.[5,6] A cheaper, faster, and less biased means of assessing technical skill is needed.

Crowdsourcing is a means of accomplishing tasks through the work of decentralized independent groups of people who are generally nonexperts in terms of the specific task.[7] Crowdsourcing has been used in a variety of ways to help solve different problems, from helping blind mobile phone users navigate their surroundings[8] to discovering complex protein folding structures.[9] In Chen et al., surgery-naive crowd-workers accurately assessed the technical skill of a single surgeon on a dry-laboratory robotic suturing task equivalent to assessments provided by a panel of experienced faculty surgeons.[10] The study was limited in that only one performance was assessed, thereby preventing any conclusions regarding variance in terms of surgical training level. We hypothesize that crowdsourcing can be used to obtain valid performance grading of urologic trainees of different levels and faculty on basic robotic skills tasks. Furthermore, we hypothesize that such assessments can be performed in a near-immediate time frame, thus elevating the value of the performance scoring feedback.

Materials and Methods

Evaluation of five unique performances

After the Institutional Review Board approval, two subject pools of video reviewer groups were recruited for this study: Amazon.com Mechanical Turk™ (Amazon.com, Seattle, WA) users and expert faculty surgeon graders, whose expertise and practice involve robotic surgery. Recruitment emails to the surgeons were sent and announcements were posted to the Mechanical Turk crowdsourcing platform. Two hundred fifty subjects were recruited through the Amazon.com Mechanical Turk platform and served as our crowd-workers. (Fig. 1). To qualify for the study, the crowd-workers had to have completed 100 or more Human Intelligence Tasks (HITs), the task unit used by Mechanical Turk, and must have had a greater than 95% approval rating as qualified by the Amazon.com site, described in Chen et al.[10] These workers were identified only by a unique anonymous user identification code provided by Amazon .com and no other information was known about them (gender, age, sex, ethnicity, etc). Each Mechanical Turk subject was compensated 0.50 USD for participating. The expert faculty surgeon grader group consisted of three experienced robotic surgeons, who have all practiced as attending surgeons for a minimum of 3 years with predominantly minimally invasive surgery practices and familiar with evaluating surgical performances by video analysis. (Fig. 1) The expert faculty surgeon graders did not receive monetary compensation. All graders were required to be older than 18 years of age.

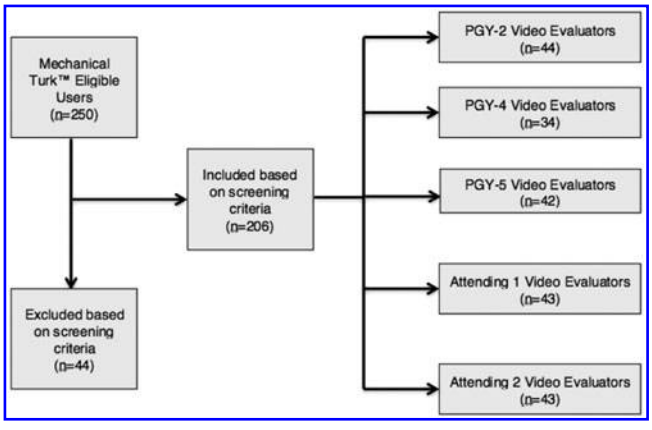

FIG. 1. Flowchart of how Amazon Turk users were distributed across the five videos.

A surgical skill assessment survey was developed and hosted online on a secure server. The survey consisted of two main parts. First, the subjects were given a qualification question in which they were shown two videos of a pair of surgeons performing a Robotic Fundamentals of Laparoscopic Surgery (RFLS) block transfer task displayed side-by-side. (Fig. 2). The surgeon in the left video performed the task with a high level of skill, while the surgeon in the right video performed the task with an intermediate level of skill. The skill level was based on published benchmark metrics for this particular task.[11,12] The crowd-workers were asked to select the performance that they thought showed the higher level of skill. Crowd-workers who failed to pick the correct video were excluded from the analysis, but were still remunerated. The expert faculty surgeon graders did not participate in this qualifying question. The survey also contained an attention question to ensure that the assessor was actively paying attention. Assessors who incorrectly answered the attention question were also excluded from the analysis.

For the second step of the survey, five videos were recorded of three urology residents (PGY-2, -4, -5) and two urology faculty performing a robotic FLS intracorporeal suturing module as part of a standard residency training session. (Fig. 3) The length of each video was 5'00,'' 2'05,'' 3'06,'' 1'33,'' and 1'00,'' respectively. (Table 1) These videos were deidentified and uploaded to the online survey, which incorporated three domains from the GEARS validated robotic surgery rating tool[13] (Fig. 4). Both crowd-workers and expert faculty surgeon graders were blinded toward the identity and experience level of the surgeon in each video performance. Fifty unique Mechanical Turk crowd-workers and three expert faculty surgeons graded each video based on the three technical skills domains (Table 1). Based on our experience in analyzing the response performance of crowds in comparison to expert reviews, we have determined that 30–50 valid crowd responses are sufficient to achieve satisfactory agreement. An internal analysis of the data collected in a previous study by our team (Chen et al. 2014) simulated the effect of smaller crowd sizes and found that this number of crowd responses generated a crowd score that fell within the agreement criteria used a majority of the time.[10] Since then, this number of crowd responses has been used as we have observed that it balances the cost of performing the crowd assessment with the composite agreement between the crowd and the experts. In addition, crowd-workers were able to assess as many videos as available to them. That is, a single crowd-worker may have been able to assess up to all five of

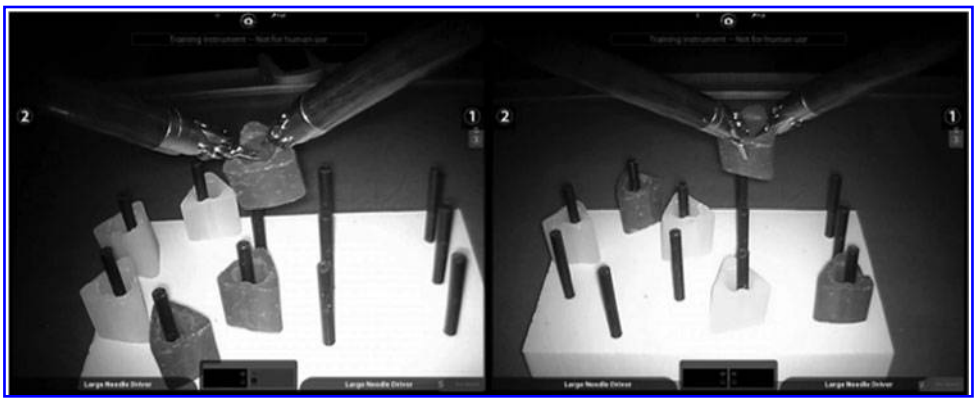

FIG. 2. Screenshot of RFLS block transfer task side-by-side video used to screen subjects. Subjects who could not correctly identify the best performance were screened out.

the videos or as few as one video depending on the number of remaining HITs available.

Videos were presented to each expert faculty surgeon grader in identical sequence; crowd-worker videos were not shown in a consistent sequence nor was a single crowd-worker guaranteed a chance to review all videos. Composite performance scores were tallied by summing the Likert ratings across the three domains with a scale of 3–15. The mean of the crowd-workers' composite scores was compared to that of the expert faculty surgeons' and assessed for correlation using Cronbach's alpha statistic to assess the degree of concordance. Following common practice, levels above 0.9 were adopted to indicate ''excellent agreement,'' down to 0.7 to indicate ''good agreement,'' and levels below 0.5 to indicate ''poor unacceptable'' levels of agreement.[14]

Cold vs *warm performance evaluation*

In addition to the above study in which crowd-workers and surgeons evaluated performances representing varying levels of training and skill, an adjunct study was performed in which the crowd evaluated a single urology resident performing an intracorporeal robotic suturing task without any warm-up and then after 10 minutes of practice with faculty-guided feedback. Performances were captured through video streamed from the robotic console. For the first performance, the resident completed the task cold (i.e., without any warm-up or instruction) and the deidentified video was posted to the Mechanical Turk site with the GEARS grading tool and discrimination questions. Immediately after 10 minutes of guided instruction, a criterion robotic suturing task was performed again, captured on video (warm), and posted to the Mechanical Turk site, as described above.

FIG. 3. Screenshot of criterion video: intracorporeal robotic suturing video graded by subjects in this study.

A total of ninety Mechanical Turk crowd-workers were recruited to evaluate the two resident videos and a third video of a faculty surgeon performing the same task. The faculty performance served as a calibration of the crowds to ensure that the crowds could discriminate higher levels of performance. Each of the three videos was evaluated by 30 crowd-workers and the same screening criterion was used as above.

Results

Evaluation of five unique performances

After eliminating crowd-workers who were excluded based on our screening criteria, we were left with grades from 206 Mechanical Turk crowd-workers (Fig. 1). We found that only 30 of the 206 responses were from crowd-workers who evaluated more than one video. In addition, only one crowd-worker evaluated all five videos. It took an average of 2 hours 50 minutes for each video to receive 50 crowd-worker scores. In comparison, it took the three expert faculty surgeon graders 26 hours to return their grades for the videos. The composite scores and frequency of scores given to each performance by both the crowd-workers and expert surgeon graders are shown in Table 1. The crowds rated the performances in the same order of skill level as the faculty. The crowds and faculty both graded the PGY-2 performance lowest, with mean composite scores of 8.27 and 7.00, respectively. Attending #2 scored highest by both crowds and faculty with a mean score of 13.0 from the crowd and 14.7 by the faculty on a scale up to 15. Interestingly, one senior trainee's performance (PGY-4) was rated higher than one of the attending surgeon's performance (attending #1) by both grading groups. Inter-rater reliability (IRR) between the surgeons and crowd was 0.91 using Cronbach's alpha statistic, which indicates an agreement level between the two groups as excellent. The linear relationship between the surgeon grades and crowd grades is shown in Figure 5.

Cold vs *warm performance evaluation*

After exclusion of subjects who failed to pass our screening criteria, we were left with grades from 65 Mechanical Turk

Table 1. Summary of Grades Assigned to Each of the Five Video Performances (PGY-2, -4, -5, Attending 1, Attending 2)

Score given	*Performance by training level*				
Training level (video length)	*PGY-2 (5'00'')*	*PGY-4 (2'05'')*	*PGY-5 (3'06'')*	*Attending 1 (1'33'')*	*Attending 2 (1'00'')*
Mechanical Turk™ subjects					
Initial *N*	50	50	50	50	50
Qualified *N*	44	34	42	43	43
CSATS					
Mean (SD)	8.27 (2.31)	11.32 (2.31)	9.93 (2.48)	10.02 (2.48)	12.97 (1.91)
95% CI	7.54, 8.91	10.55, 12.10	9.18, 10.68	9.28, 10.77	12.41, 13.55
Faculty surgeon graders					
Initial N	3	3	3	3	3
Qualified N	3	3	3	3	3
CSATS					
Mean (SD)	7.00 (1.00)	11.00 (1.73)	8.33 (2.08)	10.33 (2.52)	14.67 (0.58)
95% CI	5.87, 8.13	9.04, 12.96	5.98, 10.69	7.49, 13.18	14.01,15.00

CI = confidence intervals; SD = standard deviation.

crowd-workers. Both videos received 30 crowd-worker grades each in 1 hour and 6 minutes. The mean composite score for the cold performance was 8.41 (95% confidence intervals [CI]: 7.33, 9.49) and the mean score for the warm performance was 9.61 (95% CI: 8.56, 10.86) demonstrating 14% improvement in performance (Table 2). A box plot of the estimated distribution of scores of the cold and warm performances is shown as well as the estimated distribution of scores from an expert faculty member performance for comparison (Fig. 6).

Discussion

Despite being the current gold standard for objective assessment of technical skill, global rating scales, such as OSATS and the numerous derivative tools it implies, like GEARS, remain underutilized in academic training centers. One major limitation of OSATS is that many feel that the method is not objective.[15] OSATS assessments are often performed in person by a small panel or an individual and unless videos are obtained, it is difficult or impossible to blind assessors to the identity of the subject performing the task. Even if blinded videos are used, raters tend to be from the same training programs as their learners, which inherently carry subjective biases. Efforts taken to blind OSATS evaluators to the performer often result in more inefficiency and require that faculty allocate additional time from their practices to review these videos. The anonymity of Crowd Sourced Assessment of Technical Skill (CSATS) makes this a unique alternative. In addition, using crowdsourcing enables to scale both large numbers of videos and large pools of reviewers to increase statistical power and IRR.

CSATS may be of value as an adjunct or screening tool used to identify trainees who could benefit from targeted faculty-guided instruction. Furthermore, as surgical training programs attempt to utilize proficiency-based benchmarking for advancement, more objective performance data will need to be analyzed. Standard training methods such as high-impact one-on-one instruction are invaluable and cannot be replaced,[16] however, methods to triage skill deficits rapidly may enable more efficient use of simulation education and curriculum design.

Depth perception				
1	2	3	4	5
Constantly overshoots target, wide swings, slow to correct		Some overshooting or missing of target, but quick to correct		Accurately directs instruments in the correct plane to target
Bimanual dexterity				
1	2	3	4	5
Uses only one hand, ignores nondominant hand, poor coordination		Uses both hands, but does not optimize interaction between hands		Expertly uses both hands in a complementary way to provide best exposure
Efficiency				
1	2	3	4	5
Inefficient efforts; many uncertain movements; constantly changing focus or persisting without progress		Slow, but planned movements are reasonably organized		Confident, efficient and safe conduct, maintains focus on task, fluid progression

FIG. 4. Objective structured assessment tool; Global Evaluative Assessment of Robotic Skills (GEARS) (adapted from Goh et al.[13]).

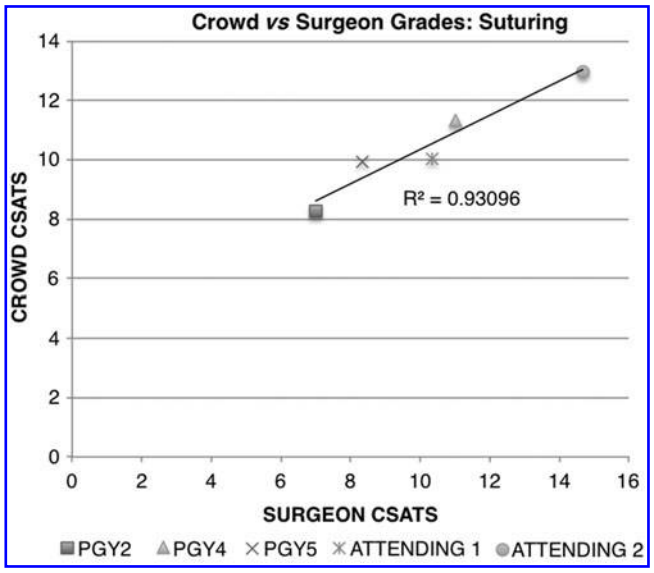

FIG. 5. The scatter plot shows the composite grades given by surgeons (*x*-axis) and crowds (*y*-axis) for each of the five videos. The linear relationship between crowd and surgeon grades for the different videos is shown and has an R^2 value of 0.93. CSATS = Crowd Sourced Assessment of Technical Skill.

In a previous study, Chen et al. showed that crowds can provide equivalent grading of suturing performances in a dry-laboratory setting; however, this study was limited in that only one video was evaluated. In this study, we showed that crowds could discriminate between varying levels of surgical skill—as effectively as expert faculty surgeon graders—in the five performance videos and with the expected ranking in the cold, warm, and expert case. In the case of the five performance videos, grades were obtained in less than 3 hours, and in the case of the warm versus cold videos, grades were obtained in about 1 hour. Feedback in training is effective if it is near real time,[17] yet many simulation training curricula call for independent practice of trainees in unsupervised environments due to limitations on faculty time. In this area, CSATS could provide an intermediate form of near real-time feedback that could be summarized for the performer and presented to the educator to help tailor training.

A limitation of this study is that only dry-laboratory performances were assessed. Real human surgery is far more complex than simple dry-laboratory suturing procedures, and it remains to be seen if crowd evaluators who have no knowledge of relevant anatomy can accurately assess animate surgery. Another limitation of this study is that only videos of robotic surgery were used; thus, it is difficult to make conclusions with regard to how crowds would evaluate performances across different surgical approaches (laparoscopic, open). Future studies aim to include videos across a range of surgical approaches as well as include human surgery. Methods to identify crowd-workers who demonstrate more accurate scoring as related to expert assessors will aid in reducing the volume of crowd-workers needed to perform the tasks—honing the crowd. Additional validation is required before CSATS is imbedded into training centers, yet forces in healthcare are driving educators to develop novel training methods to deal with the massive influx of objective surgical skills data. Scaling these data evaluations will not be feasible with the current state of faculty resources.

Table 2. Summary of Grades Assigned to Cold and Warm Resident Performances

	Cold trainee vs *warm trainee* vs *expert*		
Score given	*Cold*	*Warm*	*Expert*
Mechanical Turk™ subjects			
Initial *N*	30	30	30
Qualified *N*	22	21	25
CSATS			
Mean (SD)	8.41 (2.58)	9.71 (2.69)	12.8 (2.24)
95% CI	7.33, 9.49	8.56, 10.86	11.92, 13.68

CSATS = Crowd Sourced Assessment of Technical Skill.

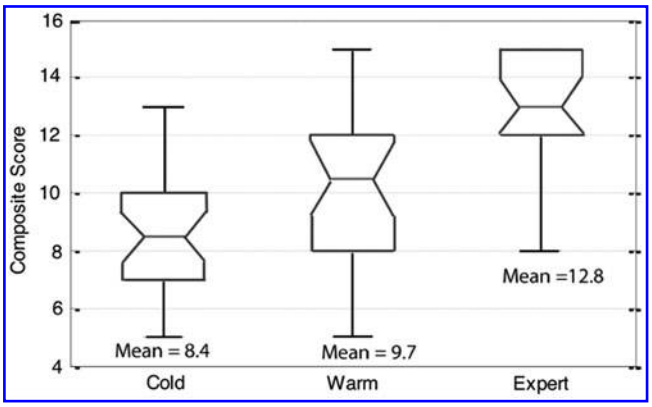

FIG. 6. Box plots showing crowd-worker composite scores for trainee's cold and warm performances and an expert performance of the same task. Notches indicate median 95% confidence intervals (CI).

Conclusions

We demonstrate that CSATS is a rapid accurate method for obtaining basic robotic dry-laboratory technical skills assessments for trainees and faculty. Furthermore, we demonstrate that untrained crowds could provide near-immediate global performance feedback using validated surgical skills assessment tools and identify improvement of skill over a short period of time of training. Utilizing crowdsourcing as a means to assess technical surgical skills provides an objective and efficient way to evaluate surgical trainees. CSATS allows for continuous iterative performance tracking with minimal resource utilization. Applications within surgical training programs need to be rigorously evaluated and ultimately linked to tracking longitudinal improvement among the trainees.

Disclosure Statement

Dr. Thomas Lendvay, Dr. Lee White, and Dr. Tim Kowalewski are cofounders in CSATS, Inc., a newly formed company dedicated to skills evaluation. All research undertaken in this article preceded the formation of CSATS, Inc. No competing financial interests exist for all other authors.

References

1. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369:1434–1442.

2. Rogers Jr SO, Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. Surgery 2006;140:25–33.
3. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, et al. Objective assessment of technical surgical skills. Br J Surg 2010;97:972–987.
4. Datta V, Bann S, Mandalia M, et al. The surgical efficiency score: A feasible, reliable, and valid method of skills assessment. Am J Surg 2006;192:372–378.
5. Darzi A, Smith S, Taffinder N. Assessing operative skill. Needs to become more objective. BMJ 1999;318:887–888.
6. Mandel LP, Lentz GM, Goff BA. Teaching and evaluating surgical skills. Obstet Gynecol 2000;95:783–785.
7. Surowiecki J. The Wisdom of Crowds. 1st Anchor books. New York, NY, USA: Anchor Books; 2005.
8. Bigham JP, Jayant C, Ji H, et al. VizWiz: Nearly Real-time Answers to Visual Questions. In: Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology. UIST'10. New York, NY, USA: ACM 2010; pp 333–342. Available at: http://doi.acm.org/10.1145/1866029.1866080, accessed May 3, 2014.
9. Khatib F, Cooper S, Tyka MD, et al. Algorithm discovery by protein folding game players. Proc Natl Acad Sci 2011. Available at: www.pnas.org/content/early/2011/11/02/1115898108, accessed May 3, 2014.
10. Chen C, White L, Kowalewski T, et al. Crowd-Sourced Assessment of Technical Skills: A novel method to evaluate surgical performance. J Surg Res 2014;187:65–71.
11. Tausch TJ, Kowalewski TM, White LW, et al. Content and construct validation of a robotic surgery curriculum using an electromagnetic instrument tracker. J Urol 2012;188:919–923.
12. Lendvay TS, Hannaford B, Satava RM. Future of robotic surgery. Cancer J Sudbury Mass 2013;19:109–119.
13. Goh AC, Goldfarb DW, Sander JC, et al. Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills. J Urol 2012;187:247–252.
14. Cronbach LJ, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas 2004;64:391–418.
15. Hiemstra E, Kolkman W, Wolterbeek R, et al. Value of an objective assessment tool in the operating room. Can J Surg 2011;54:116–122.
16. Drew PJ, Cule N, Gough M, et al. Optimal education techniques for basic surgical trainees: Lessons from education theory. J R Coll Surg Edinb 1999;44:55–56.
17. Schmidt RA, Bjork RA. New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. Psychol Sci 1992;3:207–217.

Address correspondence to:
Daniel Holst, BS
University of Washington School of Medicine
3824 Fremont Lane N.
Seattle, WA 98103

E-mail: dholst@uw.edu

Abbreviations Used

CI = confidence intervals
CSATS = Crowd-Sourced Assessment of Technical Skills
GEARS = Global Evaluative Assessment of Robotic Skills
GOALS = Global Operative Assessment of Laparoscopic Skills
HIT = Human Intelligence Task
HITs = Human Intelligence Tasks
IRR = inter-rater reliability
OSATS = Objective Structured Assessment of Technical Skills
RFLS = Robotic Fundamentals of Laparoscopic Surgery