Crowd-Sourced Assessment of Technical Skills for Validation of Basic Laparoscopic Urologic Skills Tasks

Timothy M. Kowalewski,* Bryan Comstock,† Robert Sweet,‡ Cory Schaffhausen, Ashleigh Menhadji, Timothy Averch,§ Geoffrey Box, Timothy Brand, Michael Ferrandino,‖ Jihad Kaouk,¶ Bodo Knudsen,** Jaime Landman, Benjamin Lee,††,‡‡ Bradley F. Schwartz,§§ Elspeth McDougall and Thomas S. Lendvay†,‖‖

From the Department of Mechanical Engineering (TMK, CS) and Department of Urology (RS), University of Minnesota, Minneapolis, Minnesota, Department of Biostatistics, University of Washington (BC), Department of Urology, University of Washington and Seattle Children's Hospital, Seattle (TSL), Madigan Army Medical Center, Uniformed Services University of the Health Sciences, Tacoma (TB), Washington, Boston University School of Medicine, Boston, Massachusetts (AM), University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania (TA), Department of Urology, Ohio State University, Columbus (GB, BK), Cleveland Clinic, Cleveland (JK), Ohio, Department of Urology, Duke University, Durham, North Carolina (MF), Department of Urology, UC Irvine, Orange, California (JL), Department of Urology and Oncology, Tulane University, New Orleans, Louisiana (BL), Division of Urology, Southern Illinois University, Springfield, Illinois (BFS), and Department of Urologic Sciences, University of British Colombia, Vancouver, British Columbia, Canada (EM)

Purpose: The BLUS (Basic Laparoscopic Urologic Skills) consortium sought to address the construct validity of BLUS tasks and the wider problem of accurate, scalable and affordable skill evaluation by investigating the concordance of 2 novel candidate methods with faculty panel scores, those of automated motion metrics and crowdsourcing.

Materials and Methods: A faculty panel of surgeons (5) and anonymous crowdworkers blindly reviewed a randomized sequence of a representative sample of 24 videos (12 pegboard and 12 suturing) extracted from the BLUS validation study (454) using the GOALS (Global Objective Assessment of Laparoscopic Skills) survey tool with appended pass-fail anchors via the same web based user interface. Pre-recorded motion metrics (tool path length, jerk cost etc) were available for each video. Cronbach's alpha, Pearson's R and ROC with AUC

Abbreviations and Acronyms

BLUS = Basic Laparoscopic Urologic Skills

C-SATS = Crowd-Sourced Assessment of Technical Skills

GOALS = Global Objective Assessment of Laparoscopic Skills

NPV = negative predictive value

OSATS = Objective Structured Assessment of Technical Skills

PPV = positive predictive value

Accepted for publication January 7, 2016.

No direct or indirect commercial incentive associated with publishing this article.

The corresponding author certifies that, when applicable, a statement(s) has been included in the manuscript documenting institutional review board, ethics committee or ethical review board study approval; principles of Helsinki Declaration were followed in lieu of formal ethics committee approval; institutional animal care and use committee approval; all human subjects provided written informed consent with guarantees of confidentiality; IRB approved protocol number; animal approved project number.

Supported by the American Urological Association.

* Financial interest and/or other relationship with C-SATS Inc. and Simulab Inc.

† Financial interest and/or other relationship with C-SATS Inc.

‡ Financial interest and/or other relationship with American Urological Association, Society of Laparoendoscopic Surgeons, Department of Defense, USAMRMC, and RDEcom.

§ Financial interest and/or other relationship with Bard.

‖ Financial interest and/or other relationship with TransEnterix.

¶ Financial interest and/or other relationship with Endocare.

** Financial interest and/or other relationship with Boston Scientific.

†† Current address: University of Arizona, Tucson, Arizona (e-mail: benrlee@yahoo.com).

‡‡ Financial interest and/or other relationship with Porges/Coloplast.

§§ Financial interest and/or other relationship with Cook Medical.

‖‖ Correspondence: Department of Urology, Seattle Children's Hospital, Seattle, Washington (e-mail: thomas.lendvay@seattlechildrens.org).

Editor's Note: This article is the fifth of 5 published in this issue for which category 1 CME credits can be earned. Instructions for obtaining credits are given with the questions on pages 1960 and 1961.

0022-5347/16/1956-1859/0
THE JOURNAL OF UROLOGY®

http://dx.doi.org/10.1016/j.juro.2016.01.005
Vol. 195, 1859-1865, June 2016
Printed in U.S.A.

statistics were used to evaluate concordance between continuous scores, and as pass-fail criteria among the 3 groups of faculty, crowds and motion metrics.

Results: Crowdworkers provided 1,840 ratings in approximately 48 hours, 60 times faster than the faculty panel. The inter-rater reliability of mean expert and crowd ratings was good (α=0.826). Crowd score derived pass-fail resulted in 96.9% AUC (95% CI 90.3−100; positive predictive value 100%, negative predictive value 89%). Motion metrics and crowd scores provided similar or nearly identical concordance with faculty panel ratings and pass-fail decisions.

Conclusions: The concordance of crowdsourcing with faculty panels and speed of reviews is sufficiently high to merit its further investigation alongside automated motion metrics. The overall agreement among faculty, motion metrics and crowdworkers provides evidence in support of the construct validity for 2 of the 4 BLUS tasks.

Key Words: crowdsourcing, validation studies, urologic surgical procedures, clinical competence, laparoscopy

Surgical skills directly impact patient outcomes.[1] With the rapid adoption of new technologies, the decreased amount of time trainees spend in the hospital and a shift to performance based reimbursement, we need to operationalize scalable surgical skills evaluation of even a fraction of the 51 million surgeries performed annually in the United States. Our profession needs methods which accurately and objectively provide timely and meaningful feedback while minimizing the review time burden.[2]

The current state of skills assessment requires hours of direct performance review and is usually such a formidable task that it is not done outside of research endeavors. This evaluation bottleneck may be overcome by leveraging crowdsourcing, which is the process of using large groups of decentralized, independent people providing aggregated feedback.[3] Health care applications of crowdsourcing include discovering protein folding patterns, assisting disabled patients, locating automatic defibrillators within cities and annotating electronic medical records.[3] Crowdsourcing to assess surgical skill is a method by which surgical technique can be assessed by crowds of presumably nonmedically trained reviewers (see Appendix for definitions of terms).[4–7]

In urology laparoscopic skills remain an essential component in surgery. All general surgery trainees are required to complete and pass a Fundamentals of Laparoscopic Surgery certification.[8–10] Yet in urology no such certification process exists. Recognizing this gap in trainee assessment, the American Urological Association pursued validation of a basic laparoscopic training curriculum called Basic Laparoscopic Urologic Skills.[11] The goal was to address urology appropriate cognitive and technical laparoscopic skills. More than 450 performance videos of psychomotor skills were obtained in an initial validation project which would have required substantial and arguably impossible degrees of expert assessment. This underscores a deeper, more pervasive problem facing the discipline. Such gold standard methods involving expert surgical reviewers are prohibitively resource intensive and cannot handle the high throughput of training and credentialing programs.[1] Specifically there is a need for novel methods of skill assessment that are sufficiently accurate (eg statistically concordant with faculty panel reviews) and can scale to high volumes (eg hundreds or thousands of videos) yet remain affordable (a substantially lower cost than using panels of expert surgeons for review).

We address the degree to which 2 novel candidate methods meet the requirements of automated electronic data collection and analysis described in prior work (EDGE device, Simulab Corp., Seattle, Washington)[12] and crowdsourced assessment of technical skills described in prior research.[4–7] Prior work supported the use of automated metrics via EDGE,[13] whereas in this study we investigated the use of crowdsourcing. We hypothesized that crowdsourcing could provide skills assessments concordant with those from a panel of blinded expert reviewers of a representative sample of BLUS psychomotor performances across a range of surgical skill.

MATERIALS AND METHODS

This study used a representative sample (12 pegboard and 12 suturing, established in prior work[13]) derived from 454 recordings of the 4 BLUS tasks (peg transfer, pattern cutting, suturing and clip applying) spanning medical students and urology residents, fellows and faculty surgeons from 8 academic urology training centers in the United States.[13] This representative sample was chosen to capture the typical variability of skill among training levels (inexperienced medical students, early residents, late residents and experienced attending surgeons) and within training levels (fastest, median and slowest task times).[13]

Each trial consisted of video with synchronized tool motion derived metrics ranging from 1:09 to 6:00 minutes (median 2:24). Crowdsourced data collection was adopted from the study and methods described by Chen et al.[7] The custom-built web based video display and survey tool (Zoho Corp., Pleasanton, California) allowed controlled review of videos via surgical expert and external raters using an identical web interface (fig. 1). External raters were crowdworkers recruited from the Amazon Mechanical Turk (mTurk, Amazon.com Inc., Seattle, Washington) crowdsourcing platform.

For experts and crowdworkers we captured 5-point rating assessments (1—worst to 5—best) on 4 domains (depth perception, bimanual dexterity, efficiency and tissue handling) to generate a GOALS rating.[14] We appended an overall pass-fail final question that originally appeared on the OSATS survey tool to serve as an absolute reference for a passing threshold to the numeric scores of GOALS.[15] Since the OSATS and GOALS scales typically assume expert assessors, we separately defined the crowd based ratings and corresponding scale as C-SATS. For all reviewers (faculty and crowdworkers) the order of the 24 videos was randomized into a repeating playlist, and each video evaluator was started at a random position in the playlist and asked to review the sequence of 24 videos from that point in the playlist. All reviewers were blind to the identity of the performers and scores of other reviewers. For our primary comparison group we sought at least 60 evaluations on each of the same 24 videos from crowdworkers.

Cronbach's alpha statistic was adopted to measure internal reliability between raters and between domains on a questionnaire. We assessed the reliability of outcomes between crowdworkers and experts using cited thresholds of α 0.9 or greater (excellent, high stakes testing), α 0.7 to less than 0.9 (good, low stakes testing) and α 0.6 to less than 0.7.[16] The faculty panel mean combined global GOALS ratings were taken as the ground truth measure of technical skill exhibited in each video. We only included performance ratings from crowdworkers who passed a calibration test and an attention test as described by Chen et al.[7]

The 4 domains from expert and eligible crowdworker ratings were summed into a single numeric summary score ranging from 4 (worst) to 20 (best). Since multiple ratings per expert and crowdworker were allowed across video performances (only 1 rating per video), we determined the mean crowd based C-SATS rating and 95% CI for each video/task using a linear mixed effects model

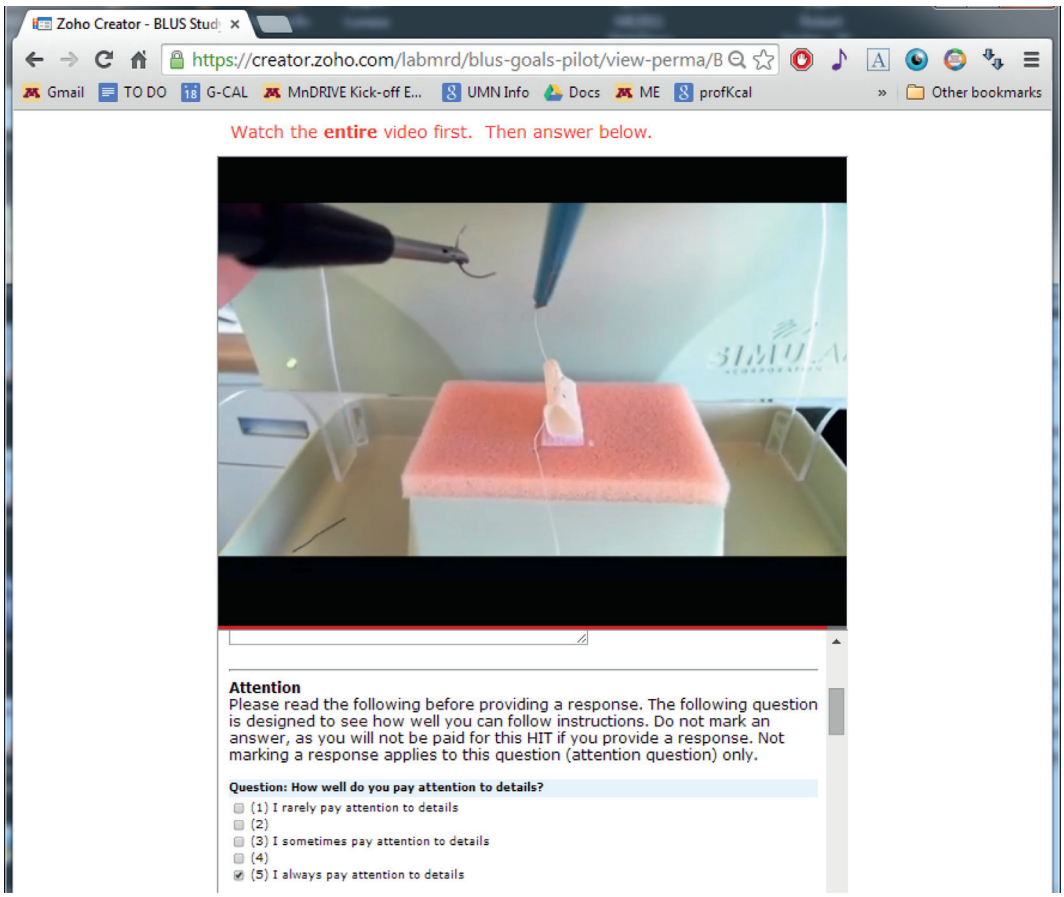

Figure 1. Screenshot of on-line presentation of GOALS assessment tool with embedded video image that crowd and expert reviewers used to evaluate performances. *HIT*, human intelligence task.

clustering on rater ID.[7,17] The mean C-SATS score was compared to the mean expert ratings graphically, overall and separately by task, as well as through the use of Cronbach's alpha statistic.[16,18] Pearson's R correlation statistic was also used to evaluate the correlation between motion metrics (task time, path length, movement count etc) and crowd and faculty scores.

Crowd based passing percentage cutoff scores were established by the C-SATS cutoff score that maximizes sensitivity and specificity, as displayed in a ROC curve, for predicting videos that a simple majority of experts rated as passing (eg greater than 50%). We estimated area under the ROC curve as a global measure of accuracy for C-SATS score at predicting an expert based passing performance. It also gave estimates of PPV and NPV for the estimated cutoff. A probability of passing performance was estimated by comparing the mean and 95% CI to percentiles from a normal distribution centered at the pass-fail cutoff score.

RESULTS

During a period of approximately 48 hours we received 1,840 ratings from Amazon Mechanical Turk crowdworkers, of which 1,438 (78.2%) passed all eligibility criteria and were used for further analysis. We captured all 120 ratings from the faculty experts during a period of 10 days (no faculty scores were excluded in the analysis). On average the crowds provided 30 eligible evaluations per hour, 60 times faster than the faculty panel average of 0.5 evaluations per hour.

The inter-rater reliability ratings among the 5 faculty experts was excellent (α=0.954). For each video and task type supplementary table 1 (http://jurology.com/) shows the mean rating and 95% CI for the experts as well as the fraction of overall passing ratings sorted by highest to lowest performance. Supplementary table 2 (http://jurology.com/) shows the same information for the crowds. The mean rating for the experts (13.07 pegboard, 14.52 suturing) was higher than for the crowds (11.60 pegboard, 12.37 suturing). Using the majority threshold for passing (greater than 50%), 16 of 24 video performances (67%) were subjectively rated as passing by experts compared to 15 of 24 (63%) for the crowds. The rating data in supplementary table 1 (http://jurology.com/) serve as the ground truth for subsequent comparisons to crowdworker ratings.

Figure 2 displays a scatterplot of mean expert ratings against mean crowdworker ratings. Superimposed onto the plot are 1) a diagonal black line (slope=1) indicating a 1:1 correspondence of expert ratings to crowd ratings, 2) the best fitting least squares line for pegboard performances (red broken line) and 3) the best fitting least squares line for suturing performances (blue broken line). The slope of the best fitting line of pegboard performances

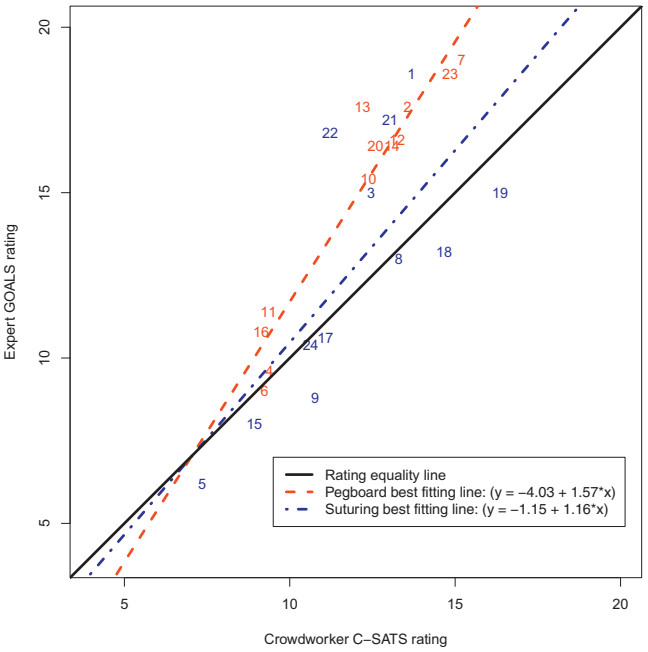

Figure 2. Correlation plot showing crowd scores using GOALS tool in x-axis compared to expert GOALS scores in y-axis. Blue broken line represents suturing skill correlation and red broken line represents pegboard block transfer correlation.

(1.57) was similar to that of suturing performances (1.16), and both indicate that while inter-rater consistency is good, experts tend to rate performances numerically higher than crowdworkers.

Across all 24 videos (combined pegboard and suturing scores) the inter-rater reliability of the mean expert rating and the mean crowdworker rating was good (α=0.826). Cronbach's alpha was similarly good within the tasks of pegboard (α=0.792) and suturing (α=0.919). The correlations between faculty scores and task motion metrics presented in table 1 show good to excellent agreement in the majority of cases. Table 2 presents correlations between crowd scores and EDGE motion metrics.

Table 1. *Correlations in faculty and crowd scores*

	Pearson's R (p value)	
Faculty Scores vs	Peg Transfer	Suturing
Crowd scores (C-SATS)	0.95 (0.00)	0.70 (0.01)
EDGE overall score (composite)*	0.95 (0.00)	0.95 (0.00)
Time*	−0.91 (0.00)	−0.95 (0.00)
Tool path length*	−0.95 (0.00)	−0.81 (0.00)
Jerk cost*	−0.63 (0.03)	−0.82 (0.00)
Movement count (speed)*	−0.94 (0.00)	−0.72 (0.01)
Economy of motion*	0.62 (0.03)	Not significant†
Errors	−0.68 (0.02)	0.62 (0.03)

*Previously published faculty correlations to task metrics, reproduced for comparison.[13]
†p >0.05.

Table 2. *Correlations of crowd scores to motion metrics*

	Pearson's R (p value)	
Crowd Scores (C-SATS) vs	Peg Transfer	Suturing
EDGE overall score (composite)	0.85 (0.00)	0.73 (0.01)
Time	−0.82 (0.00)	−0.67 (0.02)
Tool path length	−0.87 (0.00)	−0.66 (0.02)
Jerk cost	−0.70 (0.01)	−0.78 (0.00)
Movement count (speed)	−0.88 (0.00)	−0.64 (0.02)
Economy of motion	Not significant*	Not significant*
Errors	−0.71 (0.01)	0.67 (0.02)

*$p > 0.05$.

On the subjective overall pass-fail performance rating using a simple majority threshold (greater than 50%), experts and crowdworkers agreed on 92% of the performances. In 2 video performances (11 and 22) experts rated the performance as passing whereas crowdworkers did not. In a more sophisticated and objective analysis of passing performance, we examined whether the crowd based C-SATS rating could be used to distinguish among passing performances as rated by faculty experts. Figure 3 shows the resulting ROC curve and corresponding AUC of 96.9% (95% CI 90.3−100), indicating excellent discrimination between passing and failing video performances. By comparison, discrimination rates appear for crowd scores and motion metrics in table 3 by task.

We used the ROC curve to select a C-SATS based pass-fail cutoff value of 11.1 that maximized performance measures of sensitivity and specificity, resulting in a PPV of 100% and NPV of 89%.[13] Using this cutoff we estimated video specific passing probabilities in figure 3, which shows that 6 performances clearly did not pass, 13 performances clearly passed and the middle 5 performances were close to the pass-fail cutoff of 11.1. Table 3 also presents pass-fail cutoffs by task.

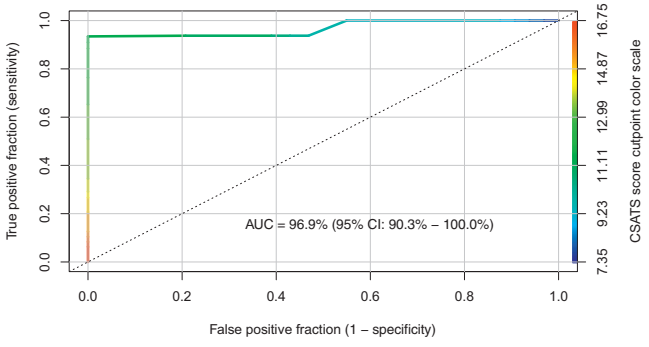

Figure 3. ROC between crowd derived cutoff score (optimal cut point 11.1) and faculty derived pass-fail rating.

Table 3. *ROC metrics based on faculty derived majority pass-fail rating for faculty scores, crowd scores C-SATS scores and motion metrics per task*

	AUC (optimal cut point)	
Faculty Pass-Fail Consensus vs	Peg Transfer	Suturing
Faculty ratings	1.0 (11.4)	1.0 (13.0)
Crowd ratings (C-SATS)	1.0 (10.1)	0.97 (12.6)
EDGE overall score (composite)	1.0 (11.4)	1.0 (12.2)
Time (sec)	1.0 (183)	1.0 (263)
Tool path length (cm)	1.0 (941)	1.0 (859)
Jerk cost (cm/sec^2)	0.89 (2.3E8)	1.0 (7.82E7)
Movement count (speed)	1.0 (122)	1.0 (72)
Force variance (N^2)	0.74 (13)	0.74 (13)

DISCUSSION

The high rate of reviews by crowdworkers strongly suggests that crowdsourcing can handle high volumes of video reviews well beyond what faculty panels can practically achieve. The ranking of performances by the crowd nearly match those of the panel, especially for significantly different performances (supplementary tables 1 and 2, http://jurology.com/). It was also surprising to note that the crowds were more critical, failing nearly 10 performances compared to the expert panel's 8. Crowdworkers provided statistically good concordance with faculty panels in terms of aggregated numeric domain scores. We observed that experts consistently rate performances higher than random crowdworkers and this is most apparent near the upper end of the C-SATS scale across numerous crowd based studies of C-SATS.[4−7] This may be because crowds are more critical than faculty experts, or perhaps it is due to regression to the mean.

Crowds provided excellent discernibility between passing and failing performance in concordance with experts (fig. 3). Furthermore, the crowds effectively erred on the side of caution since no poor performers passed (100% PPV) and only a questionably good performance failed (89% NPV). This strongly suggests that crowds may have an effective role in screening out high volumes of obviously failing or obviously passing performances. The resulting few performances closer to the performance thresholds could be forwarded to alternate review methods like expert panels for closer scrutiny.

The correlations in tables 1 and 2 indicate that, on average, crowds and motion metrics generate good correlations. Faculty and crowds appeared to reward errors in the suturing task.[13] This may underscore that the BLUS version of the suturing task places high penalties on errors or that some of the error accounting by the on-the-ground assessors may have been inadequately captured (visually tracked by proctors assessing errors over the shoulder of subjects).

Objective motion metrics provided nearly identical discrimination at passing and failing performances compared to crowds, and showed slightly increased agreement with experts in the suturing task in the ROC analysis. However, this stronger agreement may be due to the influence of task time on the composite EDGE score, and the unexpectedly strong correlation between faculty scores and task time in that task. An advantage to automated tool motion metrics for skills evaluation is that reports would be immediately available upon the completion of a task and that automated, high volume data collection is feasible across sites. Overall this indicates that objective, automated metrics from devices like EDGE may provide accurate skill evaluation and handle high volumes of reviews. The limitation of motion metric evaluation systems, apart from the high initial cost, is that not all qualities of surgical skill can be tracked through tool motions alone, such as tissue handling which, to date, requires the observation of tool-tissue interactions by a human eye. This underscores a potential benefit of the crowdsourcing method. Furthermore, in most medical centers tool motion metric capture is practically unavailable. Crowdsourcing, on the other hand, requires only an ability to capture surgical video.

There were several limitations in this study. Only 24 performances were reviewed from only 2 of the 4 tasks. While the subsamples were carefully chosen to be representative,[13] the conclusions from statistical analysis are limited. Most notably, the spread in skill level was significant over relatively few performances. Realistically a majority of performances will inhabit the moderate range of skill and not spread out evenly toward the extremes. This would indicate that over the entire database (454) the statistical concordance and ROC values might be lower.

Another limitation is that the demographics of the crowds were unknown. The Amazon Turk crowdsourcing platform precludes the identification of demographic information, such as education level, profession or other relevant demographic data that could influence the crowd's ability to assess surgical skill. Furthermore, the crowdworkers were remunerated (on average $0.67 per review) compared to voluntary expert faculty reviews, although it is unlikely that such low payments would have significantly incentivized faculty reviewers.

This study is limited to technical skill in dry lab laparoscopic performances. We do not submit that crowds can evaluate other critical surgical skills like decision making. While other studies suggest that crowds evaluate skills in animate settings, these data are early and further validation is required.[5,19]

CONCLUSIONS

We have established and described a methodology for obtaining and characterizing crowd based ratings of surgical task performance. We found that the crowd is capable of accurately sorting video performances in a manner that is on par with faculty experts. Therefore, crowd based ratings may be a useful and efficient method for discriminating between passing and failing performances, for cross-sectional performance evaluation, or for potentially measuring change in performance after remediation or training. In environments where surgical skill motions can only be captured through video and not by devices uniquely suited for motion metric acquisition, crowdsourcing offers a scalable and reasonably accurate solution to the evaluation bottleneck. The overall agreement among faculty ratings, automated motion metrics and crowd scores lends further evidence in support of the construct validity for 2 of the 4 BLUS tasks.

APPENDIX

BLUS Errors — Mistakes in the execution of BLUS Task instructions intended to mimic clinically-relevant situations: not following procedural sequence instructions, dropping objects, or tearing tissue analogues.

BLUS Motion Metrics — Quantitative measures derived from the motion and forces of laparoscopic tools during BLUS tasks such as Task time, Path length, EoM, grasp force, and EDGE SimScore.

BLUS Score — A mathematical combination of BLUS Motion Metrics and Errors.

BLUS Task — One of four, non-surgical (patient-less) tasks performed with real laparoscopic tools on physical objects demanding psychomotor and visuo-spatial laparoscopic skill with objectives to minimize completion time and avoid specific errors. These include Peg Transfer, Pattern Cutting, Suturing, and Clip Apply.

Construct Validity — Evidence that a test measures the intended construct: e.g. our test (BLUS task *and* scoring method) correctly distinguishes and classifies the skill level of a subject as compared to ground truth.

Crowdworker — A human user employed by an online crowd sourcing service like Amazon Mechanical Turk or CrowdSource.com that completes requested, pre-specified tasks typically requiring human intelligence.

C-SATS (Crowd-sourced Assessment of Technical Skills) — Crowd-sourced version of OSATS, shown to be statistically concordant in skill evaluation when replacing small panel of authoritative experts (faculty surgeons) with large anonymous crowd in certain situations.

GOALS (Global Objective Assessment of Laparoscopic Skills) — The OSATS tool modified for laparoscopic surgery, evaluation is performed by a panel of faculty surgeons with expert domain knowledge. We appended the Pass/Fail checkbox of OSATS to the GOALS survey in this study.

Ground Truth — A statistical term that refers to the absolute truth of a value or target and it is used for comparing the veracity or accuracy of proposed metrics. It differs from 'gold standard' in that it cannot be arbitrarily set or changed. In this study, the ground truth of skill is provided by blinded review of videos by a panel of faculty raters. The accuracy of proposed metrics like crowd-sourced assessment or objective motion metrics is statistically compared to this ground truth.

OSATS (Objective Structured Assessment of Technical Skills) — Validated, standard tool of technical skills assessment for open surgery, employs structured surveys and panel of faculty surgeons with expert domain knowledge; includes Pass/Fail checkbox.

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.
2. Hampton T: Efforts seek to develop systematic ways to objectively assess surgeons' skills. JAMA 2015; **313:** 782.
3. Ranard BL, Ha YP, Meisel ZF et al: Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med 2014; **29:** 187.
4. Holst D, Kowalewski TM, White LW et al: Crowd-sourced assessment of technical skills: an adjunct to urology resident surgical simulation training. J Endourol 2014; **29:** 604.
5. Holst D, Kowalewski TM, White LW et al: Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds. J Endourol 2015; **29:** 1183.
6. White LW, Kowalewski TM, Dockter RL et al: Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. J Endourol 2015; **29:** 1295.
7. 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.
8. Derossis AM, Fried GM, Abrahamowicz M et al: Development of a model for training and evaluation of laparoscopic skills. Am J Surg 1998; **175:** 482.
9. Fried GM: FLS assessment of competency using simulated laparoscopic tasks. J Gastrointest Surg 2008; **12:** 210.
10. Peters JH, Fried GM, Swanstrom LL et al: Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery 2004; **135:** 21.
11. Sweet RM, Beach R, Sainfort F et al: Introduction and validation of the American Urological Association Basic Laparoscopic Urologic Surgery skills curriculum. J Endourol 2012; **26:** 190.
12. Kowalewski TM, White LW, Lendvay TS et al: Beyond task time: automated measurement augments fundamentals of laparoscopic skills methodology. J Surg Res 2014; **192:** 329.
13. Kowalewski TM, Sweet R, Lendvay TS et al: Validation of the AUA BLUS tasks. J Urol 2016; **195:** 998.
14. Vassiliou MC, Feldman LS, Andrew CG et al: A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 2005; **190:** 107.
15. Martin JA, Regehr G, Reznick R et al: Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997; **84:** 273.
16. Cronbach LJ and Shavelson RJ: My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas 2004; **64:** 391.
17. Liang KY and Zeger SL: Longitudinal data analysis using generalized linear models. Biometrika 1986; **73:** 13.
18. Hayes AF and Krippendorff K: Answering the call for a standard reliability measure for coding data. Commun Methods Meas 2007; **1:** 77.
19. Peabody J, Miller D, Lane B et al: Wisdom of the crowds: use of crowdsourcing to assess surgical skill of robot-assisted radical prostatectomy in a statewide surgical collaborative. J Urol, suppl., 2015; **193:** e655, abstract PD30-05.

EDITORIAL COMMENT

Surgeons have long espoused the principle that knowing when and on whom to operate is more important than the ability to operate. However, recent studies have demonstrated that individual technical skill can vary greatly and significantly impacts patient outcomes (reference 1 in article). With growing acknowledgment that surgeon skill is a highly modifiable factor, there will be increasing demand for valid, objective assessments of technical skill.

In this article the authors evaluate the use of a novel crowdsourcing platform (C-SATS) to assess technical ability on validated basic laparoscopic skill tasks (reference 7 in article). The C-SATS tool demonstrated construct validity and was concordant with expert faculty ratings, the current gold standard assessment method.

The idea that nonexperts can assess technical skill as well as accepted experts is clearly a contentious one, but C-SATS is an exciting tool that may allow educators and institutions to make objective assessments in a faster, more economical and less resource intense manner, with applicability to urology trainees and for the credentialing of practicing urologists.

Jason Y. Lee
Division of Urology
St. Michael's Hospital
and Department of Surgery
University of Toronto and
Li Ka Shing Knowledge Institute
Toronto, Ontario, Canada