Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Crowd-Sourced Assessment of Technical Skills: a novel method to evaluate surgical performance

Carolyn Chen, BA,[a] Lee White, PhD,[b] Timothy Kowalewski, PhD,[c] Rajesh Aggarwal, MD, PhD,[d] Chris Lintott, PhD,[e] Bryan Comstock, MS,[f] Katie Kuksenok, BA,[g] Cecilia Aragon, PhD,[g] Daniel Holst, BS,[a,] and Thomas Lendvay, MD[h]*

[a] *University of Washington, School of Medicine, Seattle, Washington*
[b] *Department of Bioengineering, University of Washington, Seattle, Washington*
[c] *Department of Mechanical Engineering, University of Minnesota, Seattle, Washington*
[d] *Department of Surgery, University of Pennsylvania, Philadelphia, Pennsylvania*
[e] *Department of Physics, University of Oxford, Oxford, United Kingdom*
[f] *Department of Biostatistics, University of Washington, Seattle, Washington*
[g] *Department of Computer Science Engineering, University of Washington, Seattle, Washington*
[h] *Department of Urology, University of Washington, Seattle Children's Hospital, Seattle, Washington*

ARTICLE INFO

Article history:
Received 31 July 2013
Received in revised form
6 September 2013
Accepted 18 September 2013
Available online 10 October 2013

Keywords:
Crowdsourcing
Robotic surgery
OSATS
GEARS
Education
Training

ABSTRACT

Background: Validated methods of objective assessments of surgical skills are resource intensive. We sought to test a web-based grading tool using crowdsourcing called Crowd-Sourced Assessment of Technical Skill.

Materials and methods: Institutional Review Board approval was granted to test the accuracy of Amazon.com's Mechanical Turk and Facebook crowdworkers compared with experienced surgical faculty grading a recorded dry-laboratory robotic surgical suturing performance using three performance domains from a validated assessment tool. Assessor free-text comments describing their rating rationale were used to explore a relationship between the language used by the crowd and grading accuracy.

Results: Of a total possible global performance score of 3–15, 10 experienced surgeons graded the suturing video at a mean score of 12.11 (95% confidence interval [CI], 11.11–13.11). Mechanical Turk and Facebook graders rated the video at mean scores of 12.21 (95% CI, 11.98–12.43) and 12.06 (95% CI, 11.57–12.55), respectively. It took 24 h to obtain responses from 501 Mechanical Turk subjects, whereas it took 24 d for 10 faculty surgeons to complete the 3-min survey. Facebook subjects (110) responded within 25 d. Language analysis indicated that crowdworkers who used negation words (i.e., "but," "although," and so forth) scored the performance more equivalently to experienced surgeons than crowdworkers who did not ($P < 0.00001$).

Conclusions: For a robotic suturing performance, we have shown that surgery-naive crowdworkers can rapidly assess skill equivalent to experienced faculty surgeons using

* *Corresponding author*. Department of Urology, University of Washington, Seattle Children's Hospital, 4800 Sand Point Way NE, Seattle, WA, PO Box 359300. Tel.: +1 307 752 1996; fax: +1 206 987 3925.
E-mail address: dholst12@gmail.com (D. Holst).

http://dx.doi.org/10.1016/j.jss.2013.09.024

Crowd-Sourced Assessment of Technical Skill. It remains to be seen whether crowds can discriminate different levels of skill and can accurately assess human surgery performances.

1. Introduction

The annual mortality because of medical errors may be as high as 98,000 patients in the United States [1]. Even more patients experience morbidity yielding consequences both clinically and economically [1]. An extra 2.4 million hospital days and $9.3 billion are incurred annually because of medical errors [2]. Efforts to reduce surgical complication rates have included incorporation of simulation training for learning and recertification of surgical skills [3]. Global surgical performance-rating scales, such as the Objective Structured Assessment of Technical Skills (OSATS), have been widely adopted for the assessment of surgical skill and the determination of trainee advancement [4,5]. These methods, although validated, are time-intensive and rely on real-time or video-recorded analysis by surgical experts who first need to demonstrate inter-rater reliability. Increasing responsibilities of surgical educators and the trend toward standardization of training dictate a need for a cheaper, faster, less biased method of rating surgical performance.

Crowdsourcing is a relatively recent trend that uses an anonymous crowd to complete small, well-defined tasks [6]. The crowd must be diverse, decentralized, and independent, and the generated data need to be able to be aggregated [7]. Ongoing research in the area investigates how to define tasks in a way that enable the crowd to accomplish complex and expert-level work. Various workflows [8] can be used to break a complex piece of work into approachable parts and can also use the crowd to check the quality of its own work [9]. Crowdsourcing has been used to help blind mobile phone users navigate their environment [10], decipher complex protein folding structures with the online game called *Foldit* [11], and solve medical cases through the website CrowdMed.com [12]. These applications all use online work marketplaces, such as Amazon Mechanical Turk [13] to quickly and cheaply recruit an anonymous crowd of nonexperts. We hypothesize that crowd-sourced surgery performance rating is equivalent to ratings done by experienced surgeons. We also explored a link between the language of the crowd and more accurate ratings of surgical performances.

2. Materials and methods

After Institutional Review Board approval (IRB #42,811), three groups of subjects were recruited for this study: Amazon.com Mechanical Turk users, Facebook users, and teaching surgeons whose expertise and practice involve robotic surgery. Recruitment emails to the experienced surgeons were sent and Mechanical Turk and Facebook announcements were posted on the respective websites. Five hundred one subjects were recruited through the Amazon.com Mechanical Turk crowdsourcing platform (https://www.mturk.com/mturk/welcome) (Fig. 1A). Eligible subjects were active Mechanical Turk users who had completed 50 or more Human Intelligence Tasks, the task unit used by Mechanical Turk, and had achieved a greater than 95% approval rating. Each Mechanical Turk subject was compensated $1.00 for participating. In the second group, 110 subjects were recruited using Facebook (Fig. 1B). The control group consisted of 10 experienced robotic surgeons, who have all practiced as attending surgeons for a minimum of 3 y with predominantly minimally invasive surgery practices and who were familiar with evaluating surgical performances by video analysis (Fig. 1C). Neither the Facebook subjects nor the surgeon raters received monetary compensation. All subjects were required to be older than 18 y.

A surgical skill assessment survey was adapted from the Global Evaluative Assessment of Robotic Skills (GEARS) validated robotic surgery rating tool [14] and hosted online (Fig. 2). Each of the subjects from the three groups completed the same survey. The survey consisted of two steps. First, the subjects were asked to answer a qualification question in which a pair of videos of surgeons performing a Fundamentals of Laparoscopic Surgery block transfer task were displayed side by side on the screen [15] (Fig. 3). These videos were obtained from a previous study [16]. The left video demonstrated a surgeon performing with high skill, whereas the right video presented a surgeon performing with intermediate skill based on published benchmark metrics for this particular task [17,18]. Subject assessors were directed to indicate which video showed the surgeon of higher skill. This question was used to assess the subject's discriminative ability. After the qualification question, the criterion test involved rating a less than 2-min robotic surgery suture knot-tying video of an above average performance (Fig. 4) based on existing benchmark data [17,18]. No subject-identifying features were visible. After watching the video, each reviewer rated the suturing performance on three domains: depth perception, bimanual dexterity, and efficiency (Fig. 2). The domains were chosen from the six domains included in the GEARS tool and were rated on a Likert scale from 1–5 [14]. The global performance rating was obtained by summing the ratings of the three domains with a scale of 3–15. An attention question was also embedded within the criterion test to ensure that the assessor was actively paying attention and if the question was answered incorrectly, the subject was excluded from the study.

The assessor was asked to describe his or her grading rationale in a free-text box after rating for each domain. We focused on using the occurrence of style words, which are words that do not carry content individually, such as "the," "and," "but," and "however," to identify more accurate responses. Chung and Pennebaker distinguished between content and style words in text analysis, and found that noncontent words in English can help identify aspects of the writer's mood, expertise, and other characteristics [19]. In an exploratory step, we split all qualifying responses into two groups: those closer to the expert answers, and those farther

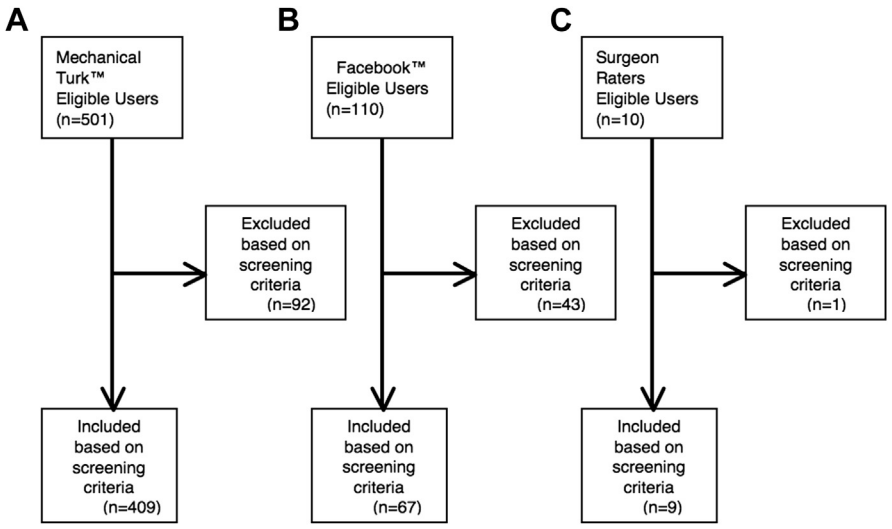

Fig. 1 – (A) Mechanical Turk subjects inclusion/exclusion diagram. (B) Facebook subjects inclusion and exclusion diagram. (C) Medical expert subjects inclusion/exclusion diagram.

away (Fig. 5). We achieved separation by computing the distance between each response and the expert average, and separating the responses along the median distance into roughly equal-sized parts: *better* and *worse* responses. Because Likert scale responses cannot be presumed to be interval-valued [20], and finding the distance between a response and the expert average does so, splitting the responses into two coarse categories serves to reduce the effect of this assumption [20].

Grades were obtained from 10 available experienced surgeons to establish a gold standard or ground truth grade for the video. A minimum of 400 ratings was determined *a priori* for the Mechanical Turk group to show equivalency with the average (mean) expert grade with >90% power, assuming a standard deviation in grades of three [21]. To establish equivalency, the entire 95% confidence interval (CI) for the mean Mechanical Turk grade had to be contained within the equivalence margin surrounding the gold standard grade. The *a priori* determined equivalence was +/− 1 point, assuming average rating differences of no greater than 0.5 points. The present study aimed to obtain a minimum of 100 Facebook user ratings to test the feasibility of alternative recruitment methods. All CIs were two-sided and not adjusted for multiple testing of groups. Statistical analyses were conducted using the R (v2.15; Institute for Statistics and Mathematics of WU [Wirtschaftsuniversität Wien]) statistical computing environment [22]. Explanations for the ratings for each of the domains were also collected. Four hundred seventy-six participants from Mechanical Turk and Facebook provided text responses.

3. Results

After eliminating subjects based on our screening criterion, we were left with nine experts (90% of the initial responses)

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. 2 – Assessment survey of skills (adapted from GEARS [14]).

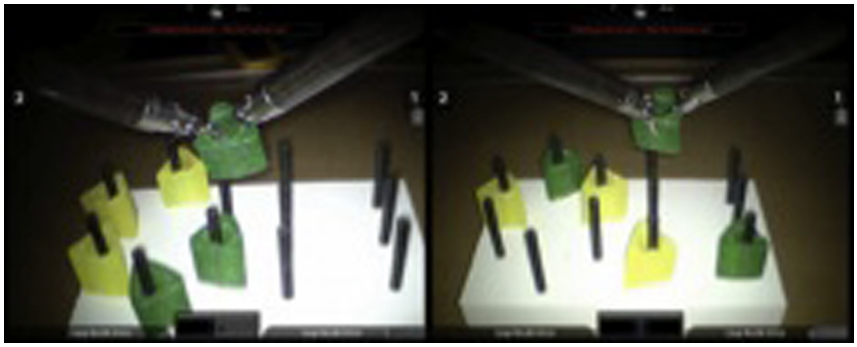

Fig. 3 – Fundamentals of Laparoscopic Surgery block transfer task side-by-side video used to screen subjects. The expert surgeon is on the left. (Color version of figure is available online.)

(Fig. 1C), 409 Mechanical Turk subjects (82% of the initial responses) (Fig. 1A), and 67 Facebook subjects (63% of the initial responses) (Fig. 1B). Surgeon raters graded the skills assessment video with a mean score 12.11, yielding an equivalence window of 11.11–13.11. Mechanical Turk and Facebook graders rated the video with mean scores of 12.21 (95% CI, 11.98–12.43) and 12.06 (95% CI, 11.57–12.55), respectively (Table 1 and Fig. 6). CIs for both crowd-sourced rating groups were contained entirely within the window of equivalence with experts. Bias from the gold standard rating was small in both crowd-sourced groups, with rating differences of +0.10 and −0.05 points for Mechanical Turk and Facebook users, respectively.

Response time from the different groups varied greatly. Table 2 shows the number of days required to achieve the responses from Mechanical Turk and faculty surgeon subjects. Figure 7 indicates the participation rate of each group over time.

With the Mechanical Turk and Facebook groups combined, 476 survey participants provided justification for their selections regarding all three domains. We considered the number of times each frequently occurring style word was observed in any of the explanations in the better *versus* worse responses. The probability of a word to occur given a good or bad response is related to the probability of a response being good or bad given the word occurring, according to the Bayes theorem [23]. We found that the word "but" was much more likely to occur in the better set of responses and therefore, focused on "but," and related negation words "however," "despite," "although," and "though." We used the existence of these words to split all qualifying responses into new *predicted-better* and *predicted-worse* categories. The predicted-better set contained 277 (58%) of the responses. As shown in Figure 8, the differences between predicted-better and predicted-worse are numerically small, but statistically significant using nonparametric Mann–Whitney *U* test ($P < 0.00001$) for each of the three dimensions of rating. The distance between the predicted-better responses is also closer to the expert average (as there were only nine expert responses, no statistical test was run).

Fig. 4 – Criterion video: intracorporeal robotic suturing video graded by subjects in this study. (Color version of figure is available online.)

4. Discussion

Development of an accurate crowd-sourced assessment of skills would address many challenges surrounding current methods of surgical performance assessment. Using C-SATS would be a faster, cheaper, less biased method of technical skill assessment compared with current methods. Although C-SATS will not replace conventional one-on-one instruction, it may be used within discrete elements of procedural education that can be outsourced to ensure objectivity and efficiency. OSATS is the current gold standard method for measuring surgical performance and relies on ratings generated by expert surgeons, which is time and resource intensive [4]. C-SATS could provide initial categorization of skills among trainees and re-evaluation of skills among experienced surgeons for maintenance of certification. It could also serve as an adjunct assessment alongside traditional rating methods, such as OSATS. If skill deficiencies can be identified early in a surgeon's training, additional focused training can be initiated.

One limitation of OSATS is the potential for bias. OSATS assessments are often performed in-person and it is difficult to blind assessors to the identity of the subject. Furthermore,

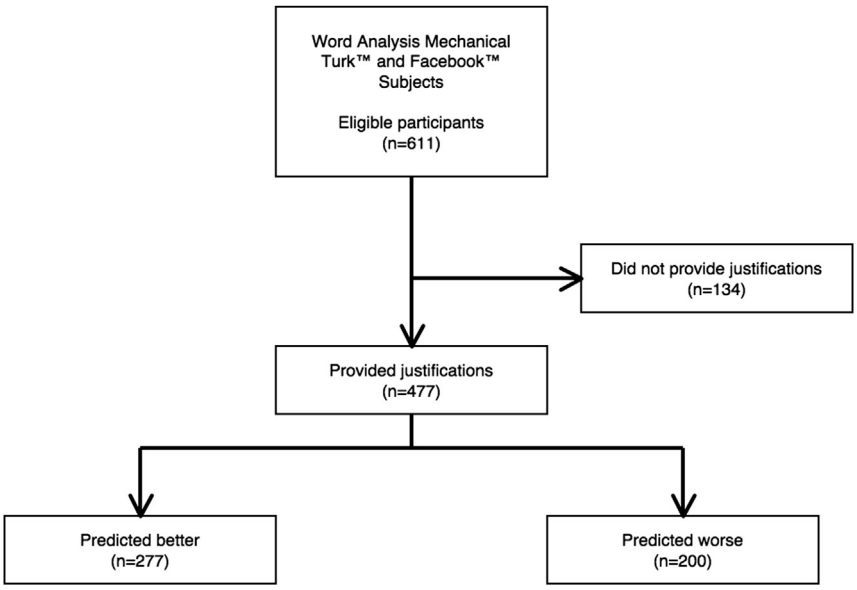

Fig. 5 – Word analysis diagram.

blinded or not, raters tend to be from the same training programs as their students and are confronted with the conflict of not advancing their own trainees if 'objective' assessments identify unsafe or ineffective practice. The crowd-sourced method is blind and thus the ratings are truly objective.

We found that applying assessment tools to more than just a very small number of performances, necessitated the creation of an online assessment suite for surgeon graders to use. The conventional method to distribute this task to a panel of surgeons would be to mail a scoring tool page and digital video disk of performance videos to all graders, have the graders watch a set of performance videos while assigning scores by hand, scan, and email or mail their scores, and finally have another individual collate those score pages into a spreadsheet or other database for assessment. This approach is actually very time-consuming both for the graders and other personnel involved in the evaluation of scores. Furthermore, this limits the locations where surgeons can perform the scoring task to those where they have access to equipment to play a digital video disk. An attractive alternative approach is to create a simple website with embedded performance videos accessible from anywhere on the Internet. This is the approach we elected to take. Scores were collected in a format natively compatible with assessment tools, such as Excel, SPSS, R, Matlab, and so forth. The grading infrastructure we built includes a very simple hypertext markup language–based survey whose results are sent directly to our server. Maintenance of the server

Table 1 – Summary of grades assigned by each subject group.

Score given	Group		
	Mechanical Turk	Facebook	Faculty surgeons
Initial *N*	501	107	10
Qualified *N*	409 (82%)	67 (63%)	9 (90%)
C-SATS			
Mean (SD)	12.21 (2.35)	12.06 (2.01)	12.11 (1.45)
95% CI	11.98, 12.44	11.56, 12.55	11.00, 13.22
Grade, *n* (%)			
3	0	0	0
4	1 (0.2)	0	0
5	2 (0.5)	0	0
6	10 (2.4)	3 (4.5)	0
7	11 (2.7)	0	0
8	14 (3.4)	0	0
9	17 (4.2)	4 (6.0)	0
10	26 (6.4)	3 (4.5)	0
11	36 (8.8)	11 (16.4)	4 (44.4)
12	78 (19.1)	17 (25.4)	3 (33.3)
13	76 (18.6)	10 (14.9)	0
14	73 (17.9)	16 (23.9)	1 (11.1)
15	65 (15.9)	3 (4.5)	1 (11.1)

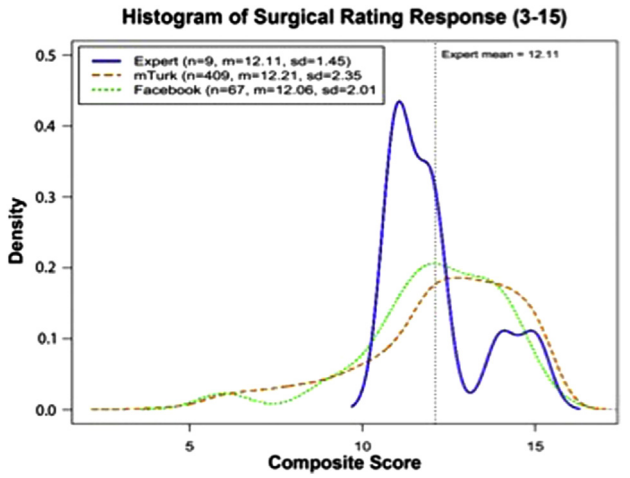

Fig. 6 – Graph showing a scoring density composite of all three assessor groups. (Color version of figure is available online.)

Table 2 – Time to receive full responses from each subject group.

Group	Days
Mechanical Turk	5
Facebook	25
Faculty surgeons	24

and generation of the survey do require a certain level of computer aptitude but in our experience has not necessitated the use of outside contractors. Once assumed that this level of infrastructure is needed for the assessment of either traditional GEARS or C-SATS, the marginal effort needed to set up a crowd-sourced assessment is very low. Essentially the same surveys presented to the surgeon graders can be added to the Amazon Mechanical Turk interface following the simple guidelines available on that site to make the surveys available to vast numbers of crowd graders.

The use of a crowd-sourced assessment is time-efficient. The study was able to generate 409 usable responses in a 24-h period (Fig. 7). In contrast, it took 25 d to generate 67 Facebook responses and 24 d to receive nine surgeon responses. One limitation to the study was that only the Mechanical Turk subjects were compensated. Perhaps more Facebook responses could have been generated at a quicker rate with compensation; however, it is unlikely that a $1.00 offer would have accelerated surgeon participation.

Our approach of using writing style cues to identify better responses is similar to the approach of using behavioral patterns for the same purpose [24]. We were able to isolate meaningfully different ratings using writing style cues alone, as evidenced by significant differences between predicted-better and predicted-worse sets. Furthermore, it is possible that these writing style cues can help identify more accurate responses, as the predicted-better responses were closer to the expert average. They were also more critical than the predicted-worse responses, which may be because negation

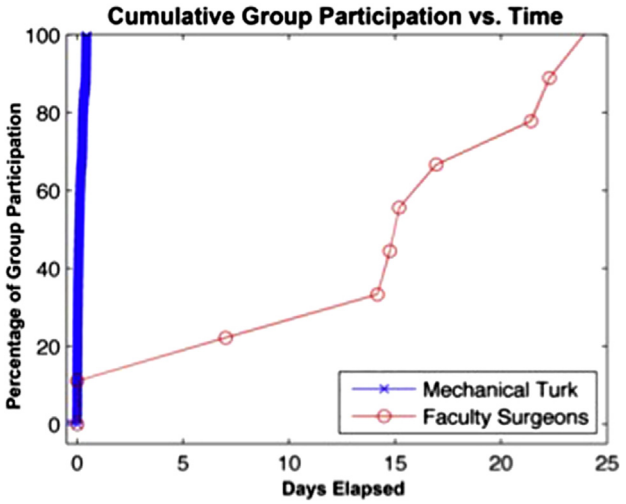

Fig. 7 – Elapsed time plot. The percent of submitted evaluations per group over time. Only participants who passed the qualification step are shown. (Color version of figure is available online.)

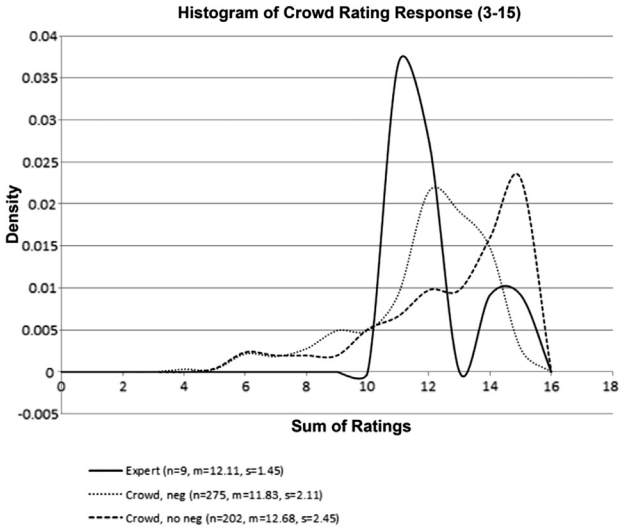

Fig. 8 – Graph showing a composite of scoring density based on assessors' free-text use or nonuse of negation words.

words serve to identify more critical responses. It is also possible that the overall crowd is more lenient than experts and identifying more critical responses implies identifying more accurate ones. For example, one subject justified rating depth perception as a 'four' (which was equivalent to the rating given by the experts for depth perception) and stated that, "Making the knots seemed at first choppy, but looked better the second time a knot was made." Using additional text cues may provide the ability to hone the crowds for specific tasks.

A limitation of this study is that a single video was assessed, so we cannot make conclusions with regard to variance. Future studies aiming to include videos across a range of surgical domains (robotic, laparoscopic, open, and so forth), surgeon skills, and human surgery will help build strength to demonstrate discrimination of variance. The present study is also limited, in that it is difficult to obtain a detailed understanding of the background of the subjects who participated in the study and the length of time that any subject spent on the survey or if the subject skipped ahead in the video. In future studies, collecting information about the assessor population, such as location, occupational history, and the length of time spent on the survey, may provide valuable information. It is possible that some of the crowdworkers were medical professionals who use the Mechanical Turk venue, which could skew the data. However, it is not likely that a large proportion of the Mechanical Turk users were medical professionals. Ideally, honing the crowd's discriminative abilities through a series of videos would allow us to identify and use individuals who show a record of accurate task assessment.

In this study, we used the Mechanical Turk platform because it was an easily accessible venue to distribute our video and survey content rapidly. There exist a number of free crowd-sourcing venues (such as CitizenScienceAlliance.org) that are strictly voluntary. These platforms are populated by "workers" who choose to participate because they are interested in advancing science, hence the name of the workers is 'citizen

scientists.' We acknowledge that there are potential ethical considerations, and our future C-SATS studies seek to enroll subjects through free platforms thereby excluding the potential for appearing coercive because of the remunerative gain.

C-SATS represents a novel methodology for rapid and efficient assessment of technical skills. Future studies may determine the minimum number of crowd-sourced users to reliably match the assessment of experienced surgeons and determine the optimal remuneration strategies that balance cost against the time it takes to collect a complete set of assessment responses. Use of a crowd to rate technical skills could be expanded to any type of medical procedure anywhere in the world. One can envision procedural training in remote centers globally that use online crowdsourcing to rapidly and objectively quantify and perhaps even qualify performance so that expertise for evaluation of skills does not need to be 'on the ground.' Furthermore, methods to use crowds for real-time intraoperative feedback may be possible to help improve performance and patient outcomes.

We report the development of a crowdsourcing method (C-SATS) for evaluating surgical performance. The crowd-sourced method provides an objective and feasible way to evaluate surgical trainees and re-evaluate experienced surgeons. With further validation, this method could be implemented to provide formative feedback for training and to create checkpoints during residency and postresidency to monitor surgical performance and acquisition of surgical skills. Further investigation is required to validate C-SATS as an adjunct to the gold standards for evaluation of procedural skills.

Acknowledgment

Katie Kuksenok was supported by an NSF Graduate Research Fellowship in Computer Science, grant number DGE-0718124. Rajesh Aggarwal was funded by a Clinician Scientist Award from the National Institute of Health Research, UK grant number NIHR/CS/099/001.

Carolyn Chen, Thomas Lendvay, Timothy Kowalewski, Katie Kuksenok, Cecilia Aragon, Lee White, Bryan Comstock, and Daniel Holst contributed to the literature search, figures, study design, data collection, and data analysis. Rajesh Aggarwal and Chris Lintott performed valuable contextual analysis and input. All the authors contributed to the data interpretation and writing of the manuscript.

The authors have no conflict of interest.

REFERENCES

[1] Kohn LT, Corrigan JM, Donaldson MS. To err is human: building a safer health system. Washington, DC: National Academies Press; 2000.
[2] Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA 2003;290:1868.
[3] Scalese RJ, Obeso VT, Issenberg SB. Simulation technology for skills training and competency assessment in medical education. J Gen Intern Med 2008;23(Suppl 1):46.
[4] van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LP, Dankelman J. Objective assessment of technical surgical skills. Br J Surg 2010;97:972.
[5] Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. Am J Surg 2006;192:372.
[6] Quinn AJ, Bederson BB. Human computation: a survey and taxonomy of a growing field. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY: ACM, pp. 1403–1412, 2011.
[7] Surowiecki J. The wisdom of crowds. 1st Anchor books ed. New York, NY: Anchor Books; 2005.
[8] Bernstein MS, Little G, Miller RC, et al. Soylent: a word processor with a crowd inside. Presented at the Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, 2010.
[9] Zaidan OF, Callison-Burch C. Crowdsourcing translation: professional quality from non-professionals. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - vol 1. Stroudsburg, PA: Association for Computational Linguistics, pp. 1220–1229, 2011.
[10] Bigham J, Jayant C, Ji H, et al. VizWiz: nearly real-time answers to visual questions. Presented at the Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology 2012.
[11] Foldit Wiki. http://foldit.wikia.com/wiki/Foldit_Wiki [accessed 18.04.13].
[12] CrowdMed. CrowdMed Beta, https://www.crowdmed.com/; 2012 [accessed 18.04.13].
[13] Amazon.com I. Amazon Mechanical Turk: artificial artificial intelligence. 2005-2013. https://www.mturk.com/mturk/ [accessed 18.04.13].
[14] Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 2012;187:247.
[15] Fried GM. FLS assessment of competency using simulated laparoscopic tasks. J Gastrointest Surg 2008;12:210.
[16] Lendvay TS, Brand TC, White L, et al. Virtual reality robotic surgery warm-up improves task performance in a dry laboratory environment: a prospective randomized controlled study. J Am Coll Surg 2013;216:1181.
[17] Tausch TJ, Kowalewski TM, White LW, McDonough PS, Brand TC, Lendvay TS. Content and construct validation of a robotic surgery curriculum using an electromagnetic instrument tracker. J Urol 2012;188:919.
[18] Lendvay TS, Hannaford B, Satava RM. Future of robotic surgery. Cancer J (Sudbury, Mass.) 2013;19:109.
[19] Chung CK, Pennebaker JW. The psychological function of function words. In: Fiedler K, editor. Social Communication. New York: Psychology Press; 2007. p. 343–59.
[20] Jamieson S. Likert scales: how to (ab) use them. Med Educ 2004;38:1217.
[21] Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006; 295:1152.
[22] Team RC. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07–0. Available from: http://www.R-project.org; 2008. 2011.
[23] Bertsekas D, Tsitsiklis J. Introduction to probability. 2nd ed. Belmont, MA: Athena Scientific; 2008.
[24] Rzeszotarski J, Kittur A. CrowdScape: interactively visualizing user behavior and output. In: Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology. New York, NY: ACM, pp. 55–62, 2012.