JOURNAL OF ENDOUROLOGY
Volume 29, Number 10, October 2015

Pp. 1183–1188
DOI: 10.1089/end.2015.0104

Crowd-Sourced Assessment of Technical Skills: Differentiating Animate Surgical Skill Through the Wisdom of Crowds

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. Sorensen, MD,[5] Mireille Truong, MD,[6] Khara Simpson, MD,[6] Alyssa Tanaka, PhD,[6] Roger Smith, PhD,[6] and Thomas S. Lendvay, MD[5]

Abstract

Background: Objective quantification of surgical skill is imperative as we enter a healthcare environment of quality improvement and performance-based reimbursement. The gold standard tools are infrequently used due to time-intensiveness, cost inefficiency, and lack of standard practices. We hypothesized that valid performance scores of surgical skill can be obtained through crowdsourcing.

Methods: Twelve surgeons of varying robotic surgical experience performed live porcine robot-assisted urinary bladder closures. Blinded video-recorded performances were scored by expert surgeon graders and by Amazon's Mechanical Turk crowdsourcing crowd workers using the Global Evaluative Assessment of Robotic Skills tool assessing five technical skills domains. Seven expert graders and 50 unique Mechanical Turkers (each paid $0.75/survey) evaluated each video. Global assessment scores were analyzed for correlation and agreement.

Results: Six hundred Mechanical Turkers completed the surveys in less than 5 hours, while seven surgeon graders took 14 days. The duration of video clips ranged from 2 to 11 minutes. The correlation coefficient between the Turkers' and expert graders' scores was 0.95 and Cronbach's Alpha was 0.93. Inter-rater reliability among the surgeon graders was 0.89.

Conclusion: Crowdsourcing surgical skills assessment yielded rapid inexpensive agreement with global performance scores given by expert surgeon graders. The crowdsourcing method may provide surgical educators and medical institutions with a boundless number of procedural skills assessors to efficiently quantify technical skills for use in trainee advancement and hospital quality improvement.

Introduction

The healthcare environment is shifting toward performance-based reimbursement and focusing on quality improvement. A 2000 study from the Agency for Healthcare Research and Quality showed that the surgical mortality rate is among the top 10 causes of death in the United States.[1] While not all deaths from surgery were due to technical errors in this particular report, a different study, which focused on the role of surgical trainees, showed that 56% of malpractice claims unearthed errors in the manual technique.[2]

Recent literature has shown that blinded video assessments of technical performances among experienced laparoscopic surgeons directly correlate with patient outcomes.[3] Subsequently, efforts have been made to adopt methods for evaluating technical skill with tools such as GEARS (Global Evaluative Assessment of Robotic Skills) and GOALS (Global Objective Assessment of Laparoscopic Skills). Both are surgical performance scales that have been extensively validated for use in grading surgical technical skill.[4,5] They are gold standard methods for evaluating surgical performances objectively, but are often burdensome and require too much time and too many resources, yielding these methods impractical for frequent use. In addition, scaling these methods to much larger studies is not practical and, in many cases, not possible.

[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, Palo Alto, California.
[4]Department of Urology, Madigan Army Medical Center, Tacoma.
[5]Department of Urology, University of Washington, Seattle, Washington.
[6]Florida Hospital Nicholson Center, Orlando, Florida.

Crowdsourcing is the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, especially from the online community rather than from traditional employees or suppliers.[6] The advent of the Internet has enabled the global labor market ready to perform various tasks/surveys to help solve problems. These problems differ widely in scope, yet crowdsourcing is a common denominator used in helping to solve them. Examples include an app used to help solve protein-folding problems and another to help blind users find their mobile phone.[7,8] In recent studies by Chen et al. and Holst et al., crowds have been shown to be as effective as expert surgeons at evaluating surgical technical skill in a dry-laboratory setting.[9,10] Not only did the crowds perform as effectively as the expert surgeons in providing skill assessment but also the cost efficiency and practicality of use were all improved with crowd graders compared to expert surgeon graders. The major limitation of these studies was that the surgical tasks being assessed were dry-laboratory tasks. Thus, no real tissue was being manipulated in the study, leaving questions regarding whether nonexperts can appreciate the subtlety of real surgery. In this study, we hypothesize that crowdsourcing can be used to obtain valid performance grading of surgical technical skill on real, living viable tissue.

Materials and Methods

After IRB approval, two groups of reviewers were recruited for this study. Representing the crowd were Amazon's Mechanical Turk™ users. These users are anonymous crowd workers from diverse backgrounds who complete small tasks for remuneration on a Mechanical Turk website, and the recruiting process was completed through this website. The second group consisted of expert faculty surgery graders, recruited through email. Six hundred prequalified Mechanical Turkers™ were recruited for the study (Fig. 1). Crowd workers must have met criteria as described by Chen et al. to qualify for the study, including having previously completed 100 or more Human Intelligence Tasks (HITs), and they must have achieved a greater than 95% approval rating as qualified by the Mechanical Turk at the time of the study.[9] A HIT is simply shorthand for a single task, which is hosted on the Mechanical Turk interface. The crowd workers' identities are anonymous and users can only be identified by unique user ID codes generated by the website. Gender, age, sex, and ethnicity were not available to the authors for this study. Each crowd worker was compensated 0.75 USD for assessing an individual performance. Seven experienced robotic surgeons, each of whom rated all videos once, made up the expert group. All the surgeons are part of practices in which minimally invasive surgery is the primary technique and they all had previous experience evaluating videos of surgical performance. The surgeon group was not compensated for participating in this study.

An online survey was developed and hosted on a secure server accessible only by recruited Mechanical Turk users. The survey contained an initial qualification question in which the crowd reviewers were shown two videos, displayed side by side, of a pair of surgeons performing a Robotic Fundamentals of Laparoscopic Surgery (RFLS) block transfer task (Fig. 2). The video on the left side of the screen showed a surgeon performing the tasks with a high level of proficiency compared to the video on the right side of the

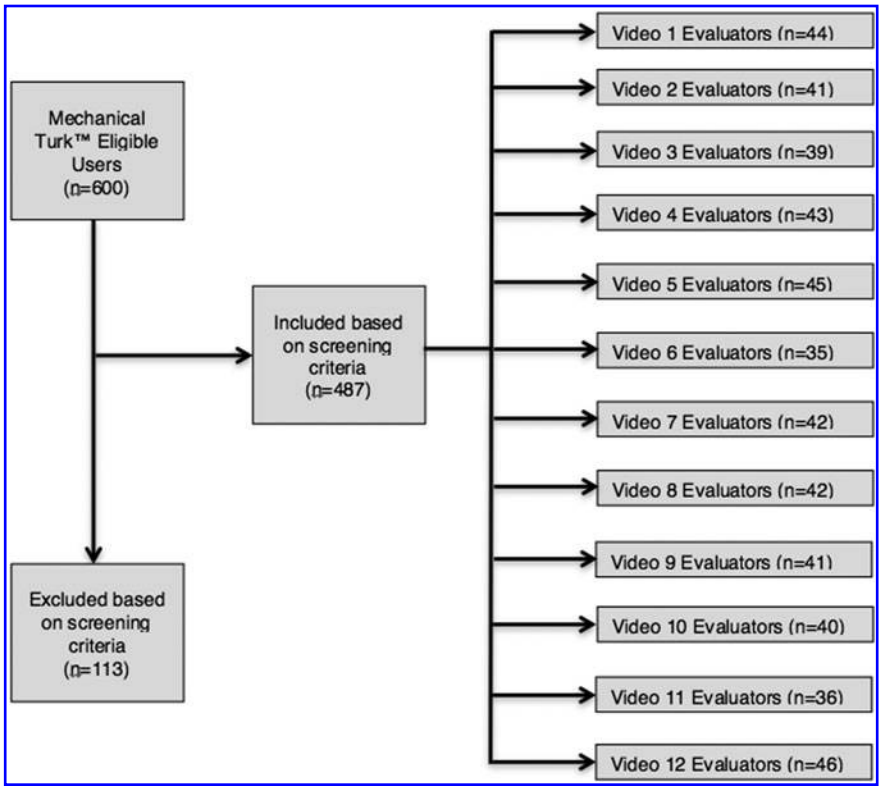


FIG. 1. Flowchart showing the breakdown of included Mechanical Turk™ graders randomly assigned to each of the 12 videos.

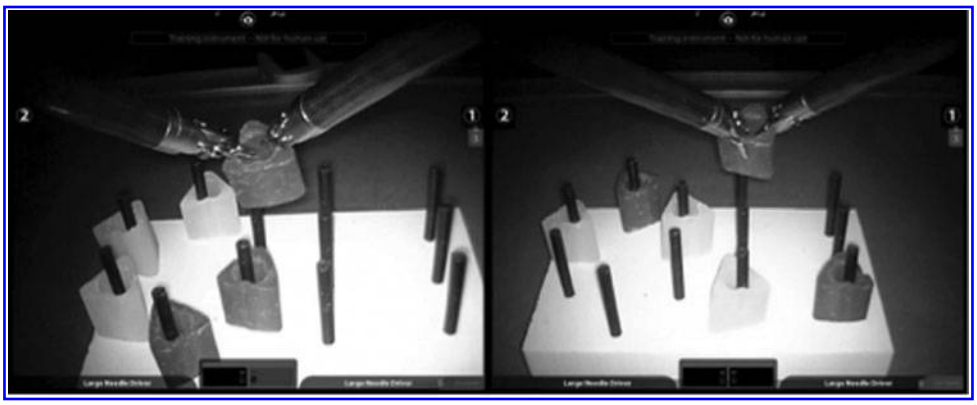


FIG. 2. Robotic fundamentals of the laparoscopic surgery (RFLS) block transfer task side by side video used to screen subjects.

screen, which showed a surgeon performing the task with an intermediate level of proficiency. These proficiency levels are based on published metric benchmarks for this particular task.[11,12] Crowd workers were asked to pick the video with the higher level of proficiency, prompting exclusion of those who answered incorrectly from the data analysis. Those excluded from the analysis were still remunerated. In addition, embedded in the survey was an attention question, which was designed so that only users who were actively paying attention to the survey would be able to correctly answer the question. Any crowd workers who answered the question incorrectly were screened out of the study and excluded from analysis (Fig. 1).

As part of the survey, we obtained recorded videos of 12 different surgeons of varying skill levels performing live porcine robot-assisted urinary bladder closures (Fig. 3). No identifying information of the surgeons performing the bladder closures was present in the videos. The length of the videos ranged between 2 and 11 minutes, with the average length being 4 hours 30 minutes. The videos were uploaded to the online survey, and evaluators were asked to evaluate the videos across five GEARS domains—bimanual dexterity, depth perception, efficiency, force sensitivity, and robotic control (Fig. 4). GEARS is an already validated tool used to assess robotic surgery.[13] Fifty unique Mechanical Turk crowd workers and seven expert surgeons evaluated each video based on the five GEARS domains. Crowd workers were only allowed to assess each performance once, but could assess more than one video if they chose to. The reason for having 50 crowd workers grade each video as opposed to larger or smaller numbers was based on a previous internal analysis of data (Chen et al.), which found 30–50 crowd responses sufficient to achieve satisfactory agreement with expert grades.[9]

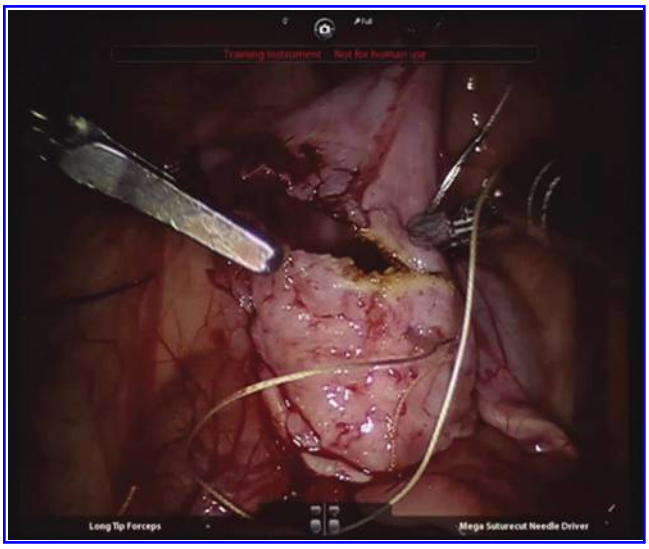


FIG. 3. Image from one of the suturing performances that was graded by both expert surgeons and Amazon's Mechanical Turk crowd.

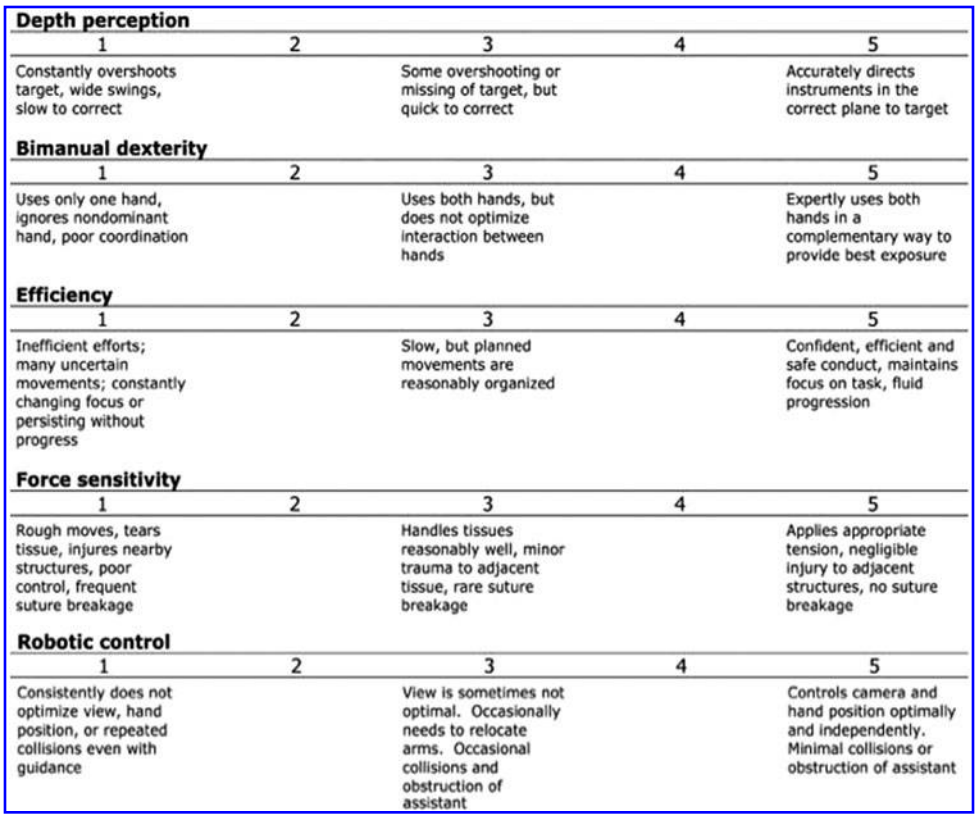
Each grader's Likert ratings across each of the five GEARS domains were summed to acquire composite performance scores for each video. This yielded a composite score scale of 5–25. Means of the crowd composite scores were assessed for concordance using Cronbach's Alpha statistic (Table 1). Cronbach's Alpha scores above 0.9 indicate excellent agreement, scores from 0.9 to 0.7 indicate good agreement, and scores below 0.5 indicate poor and unacceptable levels of agreement.[14]

Results

After excluding crowd workers who failed the attention or discrimination question, we were left with valid scores tfrom 487 of 600 Mechanical Turk crowd workers (Fig. 1). It took 4 hours 28 minutes to receive all crowd worker grades for the 12 videos. In comparison, it took 14 days to receive grades from all seven expert surgeons. Composite scores given by both the crowds and experts are shown in Table 1. Concordance between the surgeons and crowd was 0.93 using Cronbach's Alpha statistic, which indicates excellent agreement (Table 1). The linear relationship between the surgeon grades and crowd grades is shown in Figure 5. The R^2 value is 0.91. Standard error is shown in Figure 6.

Discussion

The current gold standard, an OSATS (Objective Structured Assessment of Technical Skills)-like method for objectively assessing surgical skill, continues to be underutilized due to cost, resource intensiveness, and the lag-time for return of results. Feedback is most effective if given immediately or near real time; therefore, existing OSATS practices tend to be deficient outside an academic research project.[15] Due to the significant variability in the absence of an agreement workshop and mentor bandwidth precluding



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

Force sensitivity				
1	2	3	4	5
Rough moves, tears tissue, injures nearby structures, poor control, frequent suture breakage		Handles tissues reasonably well, minor trauma to adjacent tissue, rare suture breakage		Applies appropriate tension, negligible injury to adjacent structures, no suture breakage

Robotic control				
1	2	3	4	5
Consistently does not optimize view, hand position, or repeated collisions even with guidance		View is sometimes not optimal. Occasionally needs to relocate arms. Occasional collisions and obstruction of assistant		Controls camera and hand position optimally and independently. Minimal collisions or obstruction of assistant

FIG. 4. The five Global Evaluative Assessment of Robotic Skills (GEARS) domains that were used to score the videos. Composite scores of the five domains were used to compare surgeon *vs* Turker grading.

frequent iterative trainee objective technical skills assessment, alternative methods to assist in these goals are required. In addition, video reviews may not be as objective when performed by reviewers who are within the same institution as the trainees.[16]

In Holst et al. and Chen et al., it was noted that a Crowd-sourced Assessment of Technical Skills (C-SATS) was not designed to replace one on one instruction and evaluation in the setting of residency training, but may provide an adjunct method of providing quick feedback and identifying trainees who are deficient in one area of training. Traditional methods of instruction and feedback are invaluable because they offer content expertise and transfer information about the nuances of surgery that could not be yielded by crowds; however, C-SATS may have a role in rapidly triaging trainees with deficiencies and allowing mentors to target valuable training resources to these deficiencies, as opposed to teaching all trainees with the same curricula. Feedback from crowds may be obtained rapidly enough to provide this guidance between surgical cases or between days in the operating room.

C-SATS has been used in a residency training environment, which is ideally suited to this method because of the

Table 1. Summary of Grades Assigned to Each of the 12 Video Performances

	Mechanical Turk™ graders				*Surgeon graders*		
	Initial, N	*Qualified,* N	*C-SATS mean (SD)*	*95% CI*	*Number of graders, N*	*C-SATS mean (SD)*	*95% CI*
Video 1	50	37	21.49 (3.42)	± 1.10	7	18.71 (1.67)	± 2.99
Video 2	50	41	20.95 (3.81)	± 1.17	7	18.00 (3.39)	± 2.96
Video 3	50	39	20.36 (3.51)	± 1.10	7	16.57 (5.39)	± 3.57
Video 4	50	43	18.02 (4.69)	± 1.40	7	15.85 (3.21)	± 3.01
Video 5	50	45	20.29 (3.28)	± 0.96	7	17.85 (5.10)	± 3.91
Video 6	50	35	20.37 (3.56)	± 1.18	7	18.14 (3.85)	± 2.58
Video 7	50	42	20.02 (4.04)	± 1.22	7	16.29 (5.72)	± 3.82
Video 8	50	42	21.45 (2.74)	± 0.83	7	17.85 (3.59)	± 2.07
Video 9	50	41	15.10 (4.87)	± 1.49	7	10.71 (1.92)	± 2.52
Video 10	50	40	17.13 (3.78	± 1.17	7	11.57 (2.05)	± 2.49
Video 11	50	36	18.47 (4.84)	± 1.58	7	14.57 (2.88)	± 1.76
Video 12	50	46	15.48 (4.43)	± 1.28	7	9.00 (1.67)	± 1.47
Cronbach's Alpha		0.93					

C-SATS = Crowd-sourced Assessment of Technical Skills; SD = standard deviation; CI = confidence interval.

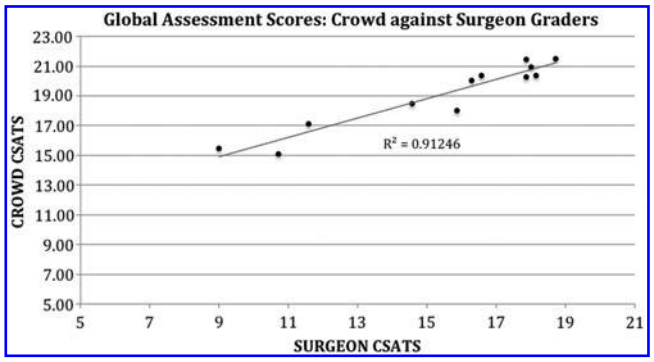


FIG. 5. Crowd-sourced Assessment of Technical Skills (C-SATS): Global performance Scores provided by the crowd against the global performance scores provided by the expert surgeon graders. The R^2 value of the best fit line is 0.91.

controlled learner-centered nature of residency. Holst et al. showed that crowds can identify differences in urology resident training levels and that crowdsourcing is a practical effective way of providing feedback in near real time.[10] The major limitation of that study, however, was that all tasks evaluated were dry-laboratory tasks. In a setting of resident work-hour restrictions, surgical trainees are spending more time in simulation laboratories to refine their technical skills, and thus, it is important that crowds can evaluate these dry-laboratory tasks quickly; however, it is vital to prove that crowds can also judge technical skill being performed on live tissue as opposed to dry-laboratory materials. Animate surgery better approximates real human surgery; thus, our hypothesis needed to be tested in this environment as a next step in validating C-SATS. With no knowledge of relevant anatomy, crowds provided extremely rapid and accurate feedback in comparison to expert graders.

A limitation of this study is that only one type of live-tissue performance was assessed and the surgery was still in a controlled environment through a porcine laboratory. In addition, all videos assessed were relatively short (averaging under 5 minutes in length). It remains to be seen if crowd evaluators can continue to provide effective grading across a range of live-tissue surgeries with varying lengths. Future studies aim to include videos across a range of surgical approaches, such as laparoscopic and open surgeries. While additional validation is needed before C-SATS is embedded into training centers, evidence that crowds can evaluate live-tissue surgery adds to the growing body of literature for the value of this adjunctive objective assessment tool.

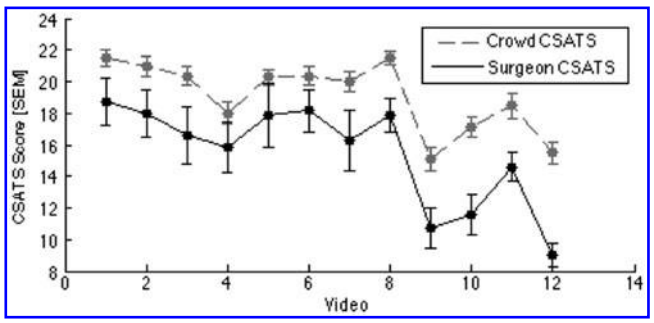


FIG. 6. The mean score of each video (*circle*) is provided for the crowd and surgeon C-SATS groups along with error bars for the standard error of the mean to indicate variation of the mean within our data.

Another limitation to this study is that the performances assessed were from a wide range of surgical skill levels from robotic faculty to novice trainees. Thus, the skill effect-size may have been disparate enough for lay people to easily see differences. It is arguable that if the cohort of performers were of more similar skill levels, it would require expert observers to discriminate the smaller technical skills differences. Resident training environments where the skills of the trainees vary significantly are ideally suited for using this methodology. Additional studies will be needed to test C-SATS on cohorts of surgeons who have similar skills.

Conclusion

We demonstrate that crowdsourcing basic surgical skills of animate surgery compares favorably to a panel of expert surgeon assessors and is faster than the experts—providing large-volume feedback in a matter of hours. Utilizing crowdsourcing as a means to assess technical surgical skills provides an inexpensive, scalable, rapid, and effective way to evaluate live-tissue procedures, paving the way for further validation in human surgery. Ultimately, C-SATS assessments will need to be linked to clinical outcomes to gain confidence that presumably nonmedically trained crowds of people can accurately ascribe surgical skill.

Author Disclosure Statement

Drs. L.W.W., T.M.K., and T.S.L. are now equity shareholders in the company CSATS, Inc., which is a company spun out of the University of Washington's Technology Transfer Office to commercialize CSATS. However, all material and data presented in this article took place before the formation of CSATS, Inc., and thus represent efforts made without the umbrella of financial incentive.

References

1. Zhan C, Miller MR. Excess length of stay, charges, and mortality attributable to medical injuries during hospitalization. JAMA 2003;290:1868–1874.
2. Rogers SO, Jr., Gawande AA, Kwaan M, et al. Analysis of surgical errors in closed malpractice claims at 4 liability insurers. Surgery 2006;140:25–33.
3. 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.
4. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, et al. Objective assessment of technical surgical skills. Br J Surg 2010;97:972–987.
5. 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.
6. Anon. Crowdsourcing. Merriam-Webster.com. Merriam-Webster, n.d. Web. Available at: www.merriam-webster.com/dictionary/crowdsourcing. Accessed January 17, 2015.
7. Bigham JP, Jayant C, Ji H, et al. VizWiz: Nearly real-time answers to visual questions. In: Proceedings of the 23rd

Annual ACM Symposium on User Interface Software and Technology. UIST'10. New York, NY: ACM. 2010 pp 333–342. Available at: http://doi.acm.org/10.1145/1866029.1866080. Accessed May 3, 2014.
8. Khatib F, Cooper S, Tyka MD, et al. Algorithm discovery by protein folding game players. Proc Natl Acad Sci 2011. Available at: http://www.pnas.org/content/early/2011/11/02/1115898108. Accessed May 3, 2014.
9. 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.
10. 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 [Epub ahead of print].
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 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. Moorthy K, Munz Y, Sarker SK, et al. Objective assessment of technical skills in surgery. BMJ 2003;327:1032–1037.
16. Darzi A, Smith S, Taffinder N. Assessing operative skill. Needs to become more objective. BMJ 1999;318:887–888.

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

C-SATS = Crowd-sourced Assessment of Technical Skills
GEARS = Global Evaluative Assessment of Robotic Skills
GOALS = Global Operative Assessment of Laparoscopic Skills
OSATS = Objective Structured Assessment of Technical Skills
HIT = Human Intelligence Task
RFLS = Robotic Fundamentals of Laparoscopic Surgery

This article has been cited by:

1. White Lee W., Kowalewski Timothy M., Dockter Rodney Lee, Comstock Bryan, Hannaford Blake, Lendvay Thomas S.. 2015. Crowd-Sourced Assessment of Technical Skill: A Valid Method for Discriminating Basic Robotic Surgery Skills. *Journal of Endourology* **29**:11, 1295-1301. [Abstract] [Full Text HTML] [Full Text PDF] [Full Text PDF with Links] [Supplemental Material]
2. Robert J. Smith, Raina M. Merchant. 2015. Harnessing the crowd to accelerate molecular medicine research. *Trends in Molecular Medicine* **21**, 403-405. [CrossRef]