Crowdsourced Assessment of Inanimate Biotissue Drills: A Valid and Cost-Effective Way to Evaluate Surgical Trainees

MaryJoe K. Rice, MS,[*] Mazen S. Zenati, MD, MPH, PhD,[†] Stephanie M. Novak, MS,[‡] Amr I. Al Abbas, MD,[§] Amer H. Zureikat, MD,[§] Herbert J. Zeh, III, MD,[‖] and Melissa E. Hogg, MD, MS[‡]

[*]University of Maryland School of Medicine, Baltimore, Maryland; [†]Department of Surgery, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania; [‡]Department of Surgery, Northshore University HealthSystem, Chicago, Illinois; [§]Division of Surgical Oncology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania; and [‖]Department of Surgery, University of Texas Southwestern Medical Center, Dallas, Texas

OBJECTIVE: Providing feedback to surgical trainees is a critical component for assessment of technical skills, yet remains costly and time consuming. We hypothesize that statistical selection can identify a homogenous group of nonexpert crowdworkers capable of accurately grading inanimate surgical video.

DESIGN: Applicants auditioned by grading 9 training videos using the Objective Structured Assessment of Technical Skills (OSATS) tool and an error-based checklist. The summed OSATS, summed errors, and OSATS summary score were tested for outliers using Cronbach's Alpha and single measure intraclass correlation. Accepted crowdworkers then submitted grades for videos in 3 different compositions: full video 1× speed, full video 2× speed, and critical section segmented video. Graders were blinded to this study and a similar statistical analysis was performed.

SETTING: The study was conducted at the University of Pittsburgh Medical Center (Pittsburgh, PA), a tertiary care academic teaching hospital.

PARTICIPANTS: Thirty-six premedical students participated as crowdworker applicants and 2 surgery experts were compared as the gold-standard.

RESULTS: The selected hire intraclass correlation was 0.717 for Total Errors and 0.794 for Total OSATS for the first hire group and 0.800 for Total OSATS and 0.654 for Total Errors for the second hire group. There was very good correlation between full videos at 1× and 2× speed with an interitem statistic of 0.817 for errors and 0.86 for OSATS. Only moderate correlation was found with critical section segments. In 1 year 275 hours of inanimate video was graded costing $22.27/video or $1.03/minute.

CONCLUSIONS: Statistical selection can be used to identify a homogenous cohort of crowdworkers used for grading trainees' inanimate drills. Crowdworkers can distinguish OSATS metrics and errors in full videos at 2× speed but were less consistent with segmented videos. The program is a comparatively cost-effective way to provide feedback to surgical trainees. (J Surg Ed 76:814–823. © 2018 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Crowdsource, Surgical education, Robotic surgery, Biotissue

COMPETENCIES: Practice-Based Learning and Improvement

ABBREVIATIONS: OSATS, Objective Structured Assessment of Technical Skills HJ, hepaticojejunostomy IHJ, interrupted hepaticojejunostomy GJ, gastrojejunostomy PJ, pancreaticojejunostomy CS, critical section HPB hepatobiliary

INTRODUCTION

In 2013, Birkmeyer et al. conducted a landmark study linking surgeon technical skill in the operating room to patient outcomes.[1] As a result of this study, there

Funding: This work was funded by an industry supported educational training grant from Intuitive Surgical (EIN 23-0965480) (Sunnyvale, CA) to train Society of Surgical Oncology fellows from multiple institutions in the use of robotic hepato-pancreatico-biliary surgery. All data is housed at the University of Pittsburgh and access to this data is not available to anyone outside the IRB approved researchers.

Disclosures: MEH also has a grant from the Society of American Gastrointestinal and Endoscopic Surgeons (5712015) and receives funding in the way of salary support from the Veterans Affairs.

Correspondence: Inquiries to Melissa E. Hogg, MD, MS, Department of Surgery, Walgreens Building – Floor 2, 2650 Ridge Road, Evanston, IL 60201; e-mail: MHogg@Northshore.org

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2018.10.007

has been an increased interest in assessing intraoperative technical skill and operating room proficiency by video review.[2-4] With emerging surgical technologies, increasing fellowships, and resident work hour restrictions,[5] there are less opportunities for teaching trainees[6] in the operating room. There has correspondingly been an increase in the use of simulation[7-10] and inanimate training drills[11,12] to augment skill acquisition. Multiple studies have now correlated higher simulation training scores with better intraoperative technical skill.[13,14] A critical component in simulation training is providing feedback to trainees on their progress with areas to target for improvement.

Inspired by these studies, the Surgical Oncology Division at the University of Pittsburgh Medical Center (UPMC) has implemented mandatory robotic training for fellows. The training consists of a simulation curriculum[8] followed by weekly inanimate biotissue drills[11] which mimic the 3 major anastomoses of the pancreaticoduodenectomy: the hepaticojejunostomy (HJ), gastrojejunostomy (GJ), and pancreaticojejunostomy (PJ). The HJ includes 2 techniques, a running HJ and an interrupted HJ (IHJ), resulting in 4 total practice drills. To start, drills were graded by 2 internationally trained hepatobiliary surgeons with high correlation.[3,11] However, as the number of weekly drills increased with inclusion of more centers, it became infeasible to provide expert surgeon review on all videos.

Crowdsourcing is the use of nonexpert workers, "the crowd," to complete labor intensive tasks.[15] Over the past 8 years, crowdsourcing has been increasingly used in medicine[16] from the public health surveillance of malaria[17] to solving complex protein structures.[18] More recently, crowdsourcing has expanded to surgical skill evaluation with validation studies performed for simulation drills,[19] inanimate drills,[20-23] and surgeries.[24-26] It was determined that a similar crowdsourced method of grading surgical videos could be used to provide feedback to trainees at UPMC. Due to the consistent, high volume of videos from weekly inanimate drills, a steady group of crowdworkers was deemed optimal for the site.

The goal of this study is to evaluate whether a selected group of nonexpert crowdworkers can reliably and consistently grade surgical videos. A secondary aim is to establish a statistical method to objectively select prospective crowdworkers. The last study objective is to evaluate the optimum composition of videos to send videos to trainees balancing both quality and cost. We hypothesize that using statistical selection to exclude outliers will result in minimally trained, nonexpert crowdworkers able to reliably grade inanimate surgical training video.

METHODS

This work reports the design, assessment, and implementation of a crowdsourced video grading program from January to December 2017. The Institutional Review Board at the University of Pittsburgh approved this study (PRO15040497).

Initial Hiring of Crowdworkers

Premedical students at the University of Pittsburgh were targeted to form a consistent group of weekly crowdworkers. A job listing to the career site was posted. Interested applicants attended a mandatory informational session with staff describing the position. The information session included a powerpoint presentation with background on the robotic training program at UPMC, steps of the robotic whipple, and the grading system. Applicants were given an overview of the 5 step robotic training program and an in-depth explanation of step 2: biotissue training.[11] To provide additional context, applicants learned the major steps of the robotic whipple procedure, with specific focus on the 3 anastomoses they would be scoring: the HJ, GJ, and PJ. Lastly, an in-depth explanation of the metrics and grading system used with nontechnical references and specific examples were given.

To test the applicants understanding of the material, they were then required to grade 9 test videos and submit their grades for analysis. The training videos were selected from a pool of biotissue videos previously graded by 2 expert surgeons. Videos were selected to include a range of low performers, average performers, and high performers for 3 different anastomoses—the HJ, GJ, and PJ. The applicants were given 1 week to submit grades on the training videos. All applicants were paid $10/hour during training regardless of final selection.

Grading System

The Objective Structured Assessment of Technical Skills (OSATS) grading system,[27] previously described and validated,[6,11] was selected. The OSATS method is a generic scale to assess technical skill without the need for procedure specific knowledge.[28] Five OSATS metrics are graded on a 5-point Likert scale: (1) gentleness, (2) time and motion, (3) instrument handling, (4) flow of operation, and (5) tissue exposure. Additionally, a similarly graded sixth metric, the summary score, was used to evaluate overall skill during the drill session. In addition to the OSATS grading method, crowdworkers used an error-based checklist, recording each damaged material,

broken suture, or air knot. Both the OSATS metrics and error-based grading methods were covered within the informational session.

For hiring selection analysis, a database was created after collection of all grades in excel format (Fig. A1). The 5 OSATS metrics were summed (maximum value 25) into Total OSATS, the Summary Score metric (maximum value 5) was kept, and the 3 errors were summed into Total Errors. Applicant grades were added to grades from 2 experienced and previously validated graders[11] for selection. The database was then anonymized and sent to a biostatistician for analysis.

Statistical Selection of Crowdworkers

Applicants were tested before long-term positions were offered to ensure understanding and competence in the use of the OSATS and error-based grading systems. Statistical selection was used to identify and eliminate outliers resulting in a homogenous final cohort of crowdworkers.

Each applicant's grades were compiled to analyze the Total Errors, Total OSATS, and Summary Score. A 1-way ANOVA F test with post hoc multiple comparison Tukey test was performed. A test of homogeneity of variance using Levenes test was used, and reliability analysis was performed using Cronbach's alpha and a single measure intraclass correlation. Linear regression method was used to fit the association lines between the average of 2 expert graders and individual crowd sourcers' outputs. All statistical tests used in the analysis were of 2 sided nature with significance level of $\alpha = 0.05$. All statistical calculations were performed in IBM SPSS Statistics 25 (IBM Corp; Armonk, NY) and Stata 15 (StataCorp LLC; College Station, TX).

Speed Analysis

After 2 weeks of grading videos, a study was designed to minimize the cost of the program without sacrificing the quality of the grades. Over the course of 3 weeks, 85 videos were sent to approved crowdworkers in 3 different compositions: full video at normal speed, full video at 2× speed, and a critical section (CS) "short" video at normal speed. For the CS videos, critical segments of each anastomotic drill were sent instead of the full video. The crowdworkers were unaware of the study or that they graded the same videos multiple times in different speeds. Similar to the initial hire statistical analysis, the grades at the 3 different compositions were tested using Levene's test for homogeneity of variance, reliability analysis using Cronbach's Alpha and single measure interclass correlation. A 1-way ANOVA with post hoc multiple comparison Tukey test to indicate differences between groups was also performed.

Vimeo as a Video Host Platform

Vimeo (New York, NY) is an online video hosting website offering a simple and inexpensive video distribution platform. For a yearly subscription to Vimeo Pro Unlimited ($399/year), 3 TB of space with unlimited weekly uploads is provided. Additionally, built-in interfaces streamline the creation of distinct sites to send to crowdworkers each week. They can watch the videos online through the platform (Fig. A2) and the site does not allow for fast forwarding of videos.

Recurring Tasks

A full-time employee is needed to coordinate the crowdworkers, prepare inanimate drills, maintain a database of the grades, and to give feedback to trainees. Biotissue training days occur twice per week and require preparation of drills the previous day with video download and editing the following day (Fig. 1). The biotissue videos are then exported to the desired speed and uploaded to the Vimeo sites sent to crowdworkers. This weekly video link is sent to the crowdsource cohort on Monday afternoons and the crowdworkers have 1 week to return their grades. Once grades are returned, they are downloaded by staff, compiled, and averaged to obtain final grades across all metrics.

Each resident, fellow, or visitor that completes biotissue training at UPMC or remotely receives a personal Vimeo page. Videos and their corresponding grades are uploaded onto these private sites as a way of providing personalized feedback on their performance.

Survey Data

A survey was sent to all crowdworkers at the end of 2017. The purpose of the survey was to identify factors making it difficult for crowdworkers to assess videos and the perceived difficulty of the various metrics. The survey was created using Qualtrics (Provo, UT) and was left open for 2 weeks to allow adequate response time.

RESULTS

Selection of Crowdworkers

During the first hiring period, 17 applicants submitted training video grades while the second round of hires had 19 applicants complete training video grades. The spread of total errors and total OSATS fitted to the average of 2 expert graders is shown in Figure 2.

Of the 17 applicants from the first hire period, 9 were determined to be within acceptable limits by having top associations, 3 were borderline, and 5 were outliers. These groupings were determined using single measure

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Training drill video download and export into 2x speed 12 PM: Grades due 5 PM: New videos sent to graders	Compilation and analysis of previous weeks grades Inanimate drill preparation	UPMC Robotic Training Inanimate Practice	Training drill video download and export into 2x speed	Inanimate drill preparation Reminder sent to graders	UPMC Robotic Training Inanimate Practice	

FIGURE 1. Weekly timeline for inanimate biotissue drill preparation and and crowdworker tasks.

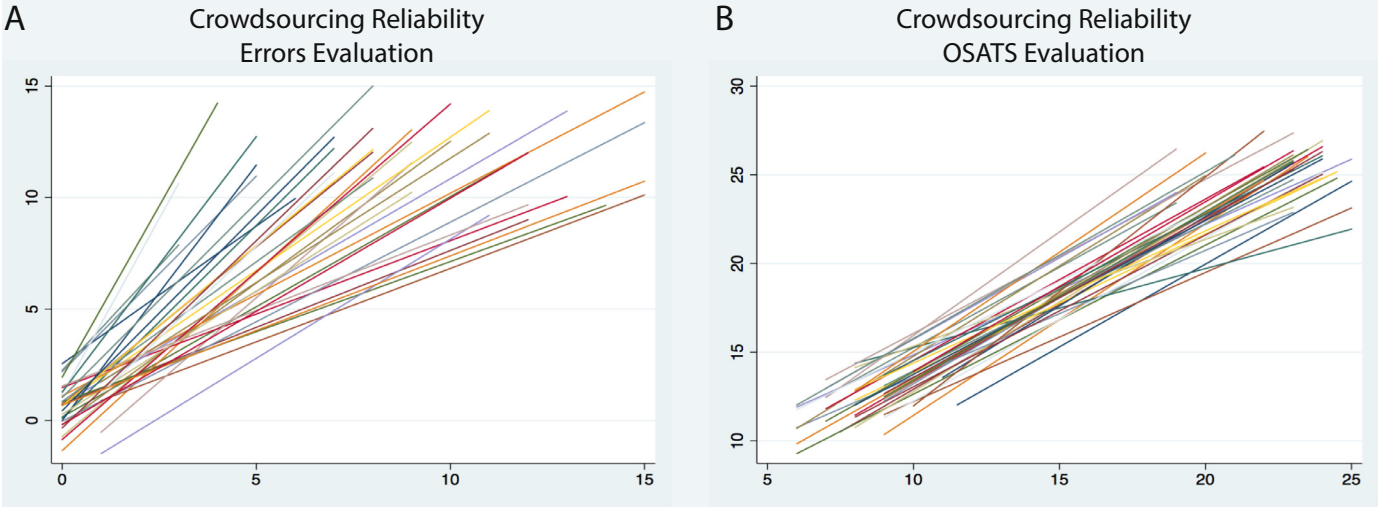

FIGURE 2. Fitted graphs where each line represents one applicant during selection process for (A) Total Errors which shows more variability at the upper end and (B) Total Objective Structured Assessment of Technical Skills (OSATS) which shows more overlap and less variability at the mean.

intraclass correlation. The single measure intraclass correlation for the selected crowdworkers was 0.717 ($p < 0.0001$) for total errors and 0.794 ($p < 0.0001$) for total OSATS (Table 1). All acceptable and borderline applicants were hired with a retraining session for borderline applicants. The second group of hires had 19 applicants of which 11 were found to be acceptable, 4 were borderline, and 4 were outliers. The final cohort intraclass correlation for errors was 0.654 ($p < 0.0001$) and for OSATs was 0.801 ($p < 0.0001$). In combination with continuing crowdworkers from the first hire batch, the second group in total had 24 crowdworkers. Table 1 shows the difference in intraclass correlation between the overall applicant group and selected crowdworkers for each applicant groups.

Speed Analysis

The speed analysis was conducted on the first hire group after 2 weeks of grading experience. Eighty-five videos were sent in 3 different speeds for a total of 255 videos graded. The videos included 27 GJs, 26 PJs, 23 HJs, and 9 IHJs. Very good correlation was found between the full video at normal (1×) speed and at double (2×) speed with an interitem statistic of 0.817 ($p < 0.0001$) for errors and 0.86 ($p < 0.0001$) for OSATS (Table 2). There was modest to good correlation for both the full video in 1× speed to CS segments and the full video in 2× speed to CS segments. With a high correlation between 1× speed and 2× speed, all inanimate drill videos were sent to the second hire group in 2× speed. However, training videos were kept at normal speed.

Summary Numbers

The crowdsourcing program was run from January to December 2017. Biotissue drills were performed by 6 general surgery residents, 19 surgical oncology and hepatobiliary fellows, and 11 surgical attendings. During this period 17,000 minutes of videos were graded correlating to more than 26,000 total minutes of video (Table 3). For the first hire group the cost per minute ($1.08) and cost per video ($27.96) were both higher than the second hire group cost per minute ($0.99) and cost per video ($19.03). The total cost of the program in 2017 was $17,479 including a $399/year vimeo subscription. Additionally, the estimated time for a research assistant to perform these duties is 0.25 full-time employee.

Survey Data

A survey was completed by 20 out of 24 crowdworkers at the end of 2017. For OSATS metrics, crowdworkers ranked all 6 from easiest to hardest: gentleness, flow of operation, time and motion, summary score, instrument handling, and tissue exposure. Out of the 3 errors, crowdworkers

TABLE 1. Intraclass Correlation for All Applicants vs Selected Applicants: Total Error and Total Objective Structured Assessment of Technical Skills (OSATS) Single Measure

		Errors			OSATS		
		Number of Viewers	Intraclass Correlation Single Measure	p Value	Number of Viewers	Intraclass Correlation Single Measure	p Value
Hire 1	All cohort	18	0.663 (0.454, 0.882)	<0.0001	18	0.774 (0.595, 0.928)	<0.0001
	Selected cohort	11	0.717 (0.506, 0.907)	<0.0001	13	0.794 (0.615, 0.936)	<0.0001
Hire 2	All cohort	20	0.552 (0.339, 0.825)	<0.0001	20	0.752 (0.565, 0.919)	<0.0001
	Selected cohort	15	0.654 (0.439, 0.878)	<0.0001	17	0.801 (0.630, 0.938)	<0.0001

TABLE 2. Intraclass Correlation for Different Speeds and Lengths: Grading Correlation Between Full Videos at 1× Speed, 2× Speed, or Segmented Critical Section (CS) "Short" Videos. All p values <0.0001

			Interitem statistic			Intraclass Correlation Single Measure
	Cases	Cronbach's Alpha	1X:2X	1X:CS	2X:CS	Single Measure, 95%CI
Total errors	85	0.906	0.817	0.721	0.751	0.737, (0.650, 0.810)
Total OSATS	85	0.892	0.860	0.660	0.690	0.730, (0.641, 0.805)

TABLE 3. Summary Numbers from a Year of Crowdsource Grading

	Hire 1	Hire 2	Total
Total # GJs	87	138	225
Total # HJs	77	106	183
Total # IHJs	29	101	130
Total # PJs	85	144	229
Total # videos graded	278	489	767
Total time (minutes)	7185	9409	16,594
Total cost	$7772	$9308	$17,479
Cost per minute	$1.08	$0.99	$1.03
Cost per video	$27.96	$19.03	$22.27

found it easiest to recognize damaged material and the most difficult to distinguish air knots. Of the anastomoses, HJs and IHJs were rated to be the easiest though none of the anastomoses were found to be "Extremely Difficult." Fifteen respondents (75%) report watching the videos just once, 4 (20%) report watching the videos twice, and 1 respondent (5%) reports watching the videos more than twice. Seventeen graders (85%) stated they had to rewind videos to double check for errors. Crowdworkers rated blurry or out of focus videos, lighting, and "missing video" to be the factors most influencing difficulty of accurate grading. 19 respondents (95%) stated they took the job for experience/resume builder while 1 participant (5%) took the job for the money.

DISCUSSION

It was possible to use statistics to select a homogenous cohort of crowdworkers from the undergraduate population with good correlation to expert graders. Over 1 year the crowdworkers proved to be reliable and efficient in grading inanimate drills and providing critical feedback to 36 trainees. The correlation was very similar for the Total OSATS and the Summary Score while the Total Errors had consistently lower correlation (Fig. 2). The grades showed the highest correlation for videos in the middle of the spectrum (i.e., the average performers) with higher deviation toward the extremes on either side. This is true for Total OSATS, Summary Score, and Total Errors.

This is the first study to evaluate whether speed and composition of videos affects grades. The results indicate crowdworkers are able to clearly distinguish necessary characteristics for grading full videos in double speed. While there was still moderate correlation between the CS short videos, the lower correlation indicates some confusion with missing segments. Futhermore, survey results indicated "missing video" made videos more difficult to grade. These results suggest full videos should be used over selected segments whenever possible for the most accurate grades. Alternatively, further training may be necessary when only short segments are utilized.

Multiple studies have validated the use of crowdworkers to grade simulation drills,[19] inanimate drills,[20-23] and surgeries.[24-26] The majority of these studies have utilized Amazon Mechanical Turk (AMT; Amazon Inc., Seattle, WA), an online crowdsourcing platform, to recruit users. AMT simplifies the process needed to postvideos and receive grades without conducting a full hiring process. AMT additionally has a rapid response time, often with >500 responses gathered in <24 hours. However, the platform does not allow for continued users and works on a first-come, first-served basis. Even when a large number of videos are uploaded together, previous studies have shown 75% of crowdworkers grade less than 5 videos.[22] For a long term, sustained program, it is more desirable to have a consistent group of crowdworkers each week. Additionally, all previous studies have used video clips <10 minutes.[29] The biotissue drills in this study range from 10 to 90 minutes depending on the type of anastomosis and experience level of a trainee. Thus, it would likely be more difficult to recruit users over AMT for long videos without grossly increasing pay.

Cost is an important factor when choosing to implement a program. Cost in previous crowdsource graded studies ranged from $15.00[19] to $500.00[20] per video with an average around $40/video.[29] However, these costs are for videos between 2 and 10 minutes in length bringing the cost per minute previously reported to between $1.50 and $250. For the duration of this program the average cost per video is $22.27 and cost per minute is $1.03. However, after sending the videos in 2× speed (hire 2)

both the cost per video and cost per minute decreased. This program is therefore less expensive per video, despite significantly longer videos, and per minute than any previously reported program. While $23 per video may be cost prohibitive to some programs, it demonstrates a marked decrease to previously reported numbers. One way to increase the sustainability of the program would be to grade every few videos per trainee instead of every one. Lastly, only 5% of crowdworkers in the program report being incentivized by the money with the rest participating for the experience and knowledge. As the crowdworker population is predominantly undergraduate premedical students, it is probable crowdworkers would participate in a nonpaid program with alternative incentives such as opportunities for surgical observation or physician shadowing.

This study has several limitations. First, after final crowdworker selection, no further quality assurance checks or comparisons to expert graders were conducted. In the future, a quarterly analysis will be incorporated to ensure hired crowdworkers correlate with expert reviewers grading 10% of video. Start-up and hiring is a potential prohibitive aspect of the crowdsourced grading program. Time to post the crowdworker position, wait for enough interested applicants, hold an information session, and analyze training grades takes at least 1 month with significant time investment by a staff member. Therefore, if a program only needs 1 batch of videos graded, a yearly assessment for example, then it would be quicker to use AMT or another similar platform well designed for single uses. A benefit to having a source of validated graders may be that more can potentially be added on an individual basis and analyzed against the qualified pool that has already been chosen. Another limitation is that these crowdworkers only score select inanimate biotissue drills for which they have been trained. Expansion of their capabilities to additional inanimate drills and operative footage is a future aim. Another limitation is that previously validated surgeon graders were used as a "gold standard" when in fact no actual gold standard currently exists to assess surgical technical skills. Lastly, the time from performance of drills to final grade posting is 2-4 weeks which may be too long for surgeons hoping to quickly move to operating robotically. However, providing feedback to residents and fellows over a 1-7 year training program does not require immediate feedback. Under the current system, trainees are still able to receive grades in a timely manner and graduate with a portfolio of their skills developed over time. Lastly, there were small sample sizes during the selection process with only 9 training videos and <20 graders to select from.

Providing consistent feedback to trainees on their performance is critical for improvement. During the first 2 years of the program, expert surgical review with 2 graders was utilized; however, as the program expanded to residents, off-site fellows, and attendings taking courses, expert surgical review was not feasible for the high volume of drills being performed. Crowdsourced assessment therefore filled a crucial role by giving specific feedback to trainees on areas to target whether it be gentleness or instrument handling. Over time, hopefully this results in an increase in skill in inanimate drills.[11]

This is a novel method for statistically selecting a group of nonexpert crowdworkers to grade inanimate surgical video in a weekly fashion to add assessment to an inanimate surgical simulation training program. This is the first study to show sustained use of crowdworkers, over the period of 1 calendar year, rather than 1 time studies. The study additionally determined that inanimate drills are able to be viewed as full videos in double speed but are not ideal to view in segmented form without a decrease in grade accuracy. The same methodology outlined here to establish and sustain a crowdsourcing program can be customized and applied to a variety of inanimate drills and specialties suggesting a wide field of impact.

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. Fecso AB, Bhatti JA, Stotland PK, Quereshy FA, Grantcharov TP. Technical performance as a predictor of clinical outcomes in laparoscopic gastric cancer surgery. *Ann Surg*. 23 Mar 2018. https://doi.org/10.1097/SLA.0000000000002741. [Epub ahead of print].
3. Hogg ME, Zenati M, Novak S, et al. Grading of surgeon technical performance predicts postoperative pancreatic fistula for pancreaticoduodenectomy independent of patient-related variables. *Ann Surg*. 2016;264:482–491.
4. Palter VN, Grantcharov TP. Individualized deliberate practice on a virtual reality simulator improves technical performance of surgical novices in the operating room: a randomized controlled trial. *Ann Surg*. 2014;259:443–448.
5. Bilimoria KY, Chung JW, Hedges LV, et al. National cluster-randomized trial of duty-hour flexibility in surgical training. *N Engl J Med*. 2016;374:713–727.
6. Reznick RK, MacRae H. Teaching surgical skills—changes in the wind. *N Engl J Med*. 2006;355:2664–2669.

7. Teitelbaum EN, Soper NJ, Santos BF, et al. A simulator-based resident curriculum for laparoscopic common bile duct exploration. *Surgery*. 2014;156:880–887. 890-883.

8. Hogg ME, Tam V, Zenati M, et al. Mastery-based virtual reality robotic simulation curriculum: the first step toward operative robotic proficiency. *J Surg Educ*. 2017;74:477–485.

9. Sheth SS, Fader AN, Tergas AI, Kushnir CL, Green IC. Virtual reality robotic surgical simulation: an analysis of gynecology trainees. *J Surg Educ*. 2014;71:125–132.

10. Tergas AI, Sheth SB, Green IC, Giuntoli R.L. 2nd, Winder AD, Fader AN. A pilot study of surgical training using a virtual robotic surgery simulator. *JSLS*. 2013;17:219–226.

11. Tam V, Zenati M, Novak S, et al. Robotic pancreatoduodenectomy biotissue curriculum has validity and improves technical performance for surgical oncology fellows. *J Surg Educ*. 2017;74:1057–1065.

12. Hung AJ, Jayaratna IS, Teruya K, Desai MM, Gill IS, Goh AC. Comparative assessment of three standardized robotic surgery training methods. *BJU Int*. 2013;112:864–871.

13. Fried GM, Feldman LS, Vassiliou MC, et al. Proving the value of simulation in laparoscopic surgery. *Ann Surg*. 2004;240:518–528.

14. Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg*. 2002;236:458–464.

15. Wazny K. "Crowdsourcing" ten years in: a review. *J Glob Health*. 2017;7:020602.

16. 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;29:187–203.

17. Chunara R, Chhaya V, Bane S, et al. Online reporting for malaria surveillance using micro-monetary incentives, in urban India 2010-2011. *Malar J*. 2012;11:43.

18. Cooper S, Khatib F, Treuille A, et al. Predicting protein structures with a multiplayer online game. *Nature*. 2010;466:756–760.

19. Aghdasi N, Bly R, White LW, Hannaford B, Moe K, Lendvay TS. Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. *J Surg Res*. 2015;196:302–306.

20. 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.

21. White LW, Kowalewski TM, Dockter RL, Comstock B, Hannaford B, Lendvay TS. Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. *J Endourol*. 2015;29:1295–1301.

22. Polin MR, Siddiqui NY, Comstock BA, et al. Crowdsourcing: a valid alternative to expert evaluation of robotic surgery skills. *Am J Obstet Gynecol*. 2016;215:644. e641-644 e647.

23. Lee JY, Andonian S, Pace KT, Grober E. Basic laparoscopic skills assessment study: validation and standard setting among canadian urology trainees. *J Urol*. 2017;197:1539–1544.

24. 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;29:1183–1188.

25. Powers MK, Boonjindasup A, Pinsky M, et al. Crowdsourcing assessment of surgeon dissection of renal artery and vein during robotic partial nephrectomy: a novel approach for quantitative assessment of surgical performance. *J Endourol*. 2016;30:447–452.

26. Ghani KR, Miller DC, Linsell S, et al. Measuring to improve: peer and crowd-sourced assessments of technical skill with robot-assisted radical prostatectomy. *Eur Urol*. 2016;69:547–550.

27. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.

28. Aggarwal R. Intraoperative surgical performance measurement and outcomes: choose your tools carefully. *JAMA Surgery*. 2017;152:995–996.

29. Katz AJ. The role of crowdsourcing in assessing surgical skills. *Surg Laparosc Endosc Percutan Tech*. 2016;26:271–277.

SUPPLEMENTARY INFORMATION

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jsurg.2018.10.007.

APPENDIX A

.

	Errors			OSATS					
Video Name	Damaged Material	Broken Suture	Air Knot	Gentleness	Time and motion	Instrument Handling	Flow of Operation	Tissue exposure	Summary Score
GI #1	0	0	0	5	4	3	2	5	4
GI #2	1	0	1	3	2	3	1	4	3
GI #3	0	0	0	5	4	5	4	5	4
HJ #1	0	0	0	5	3	3	3	4	4
HJ #2	0	0	0	5	3	4	3	5	4
HJ #3	0	0	0	5	5	4	5	5	5
IHJ #1	0	0	0	5	3	4	4	4	4
IHJ #2	0	0	3	3	3	5	2	3	3
IHJ #3	0	0	2	3	3	5	3	3	3
IHJ #4	0	0	0	5	4	4	4	3	4
PJ #1	0	0	1	5	2	2	1	3	2
PJ #2	0	0	0	5	3	4	2	5	4
PJ #3	0	1	1	4	4	4	3	3	4
PJ #4	0	0	3	4	2	4	2	4	3
PJ #5	0	0	0	5	4	5	4	5	5

FIGURE A1. Example returned grading sheet from a crowdworker.

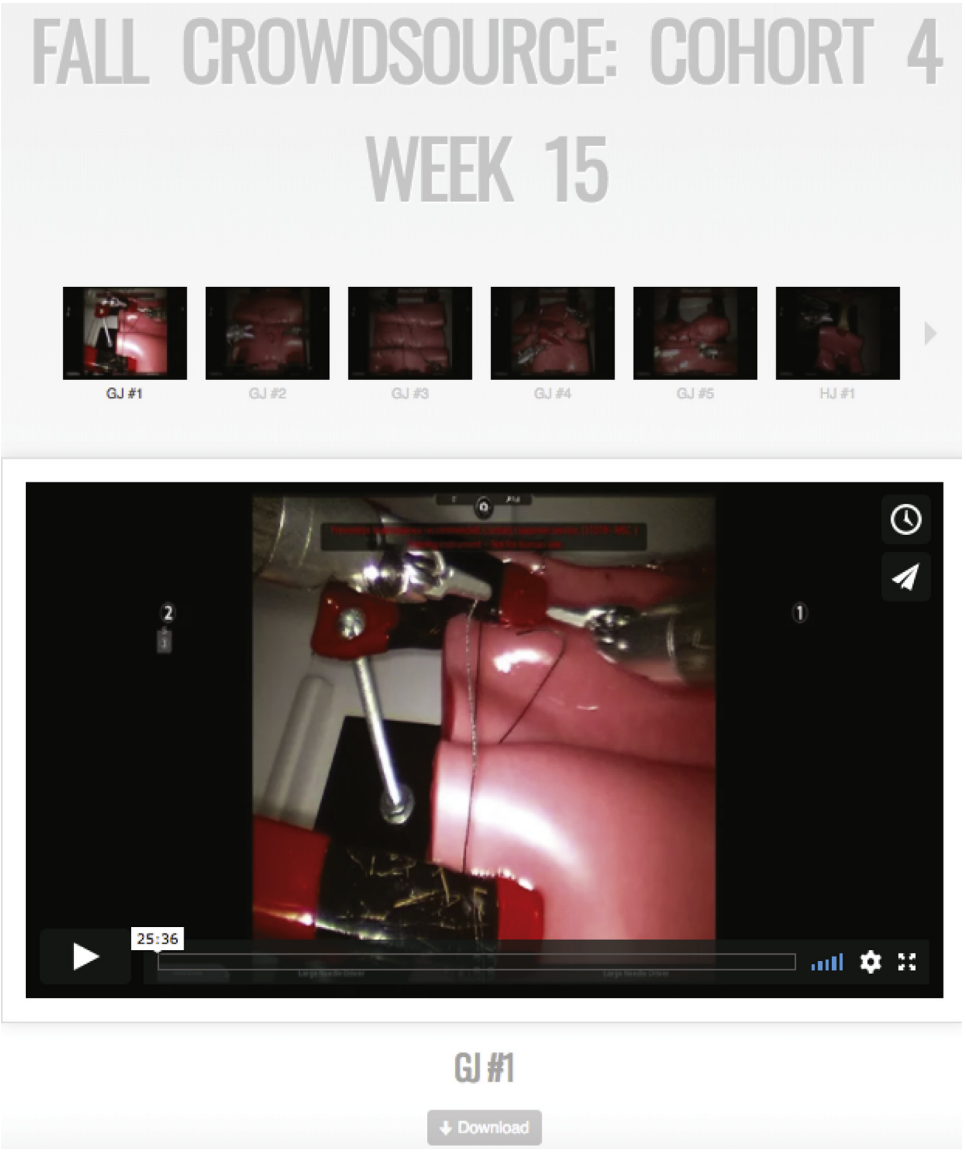

FIGURE A2. Example Vimeo site sent to a cohort of crowdworkers.