Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills

Shanley B. Deal, M.D.\textsuperscript{a,}\textsuperscript{*}, Thomas S. Lendvay, M.D., F.A.C.S.\textsuperscript{b}, Mohamad I. Haque, M.D., L.T.C.\textsuperscript{c}, Timothy Brand, M.D., L.T.C.\textsuperscript{c}, Bryan Comstock, M.S.\textsuperscript{b}, Justin Warren, M.B.A.\textsuperscript{b}, Adnan Alseidi, M.D., M.Ed.\textsuperscript{a}

\textsuperscript{a}Virginia Mason Medical Center, Department of General Surgery, Mailstop H8-GME, 1100 9th Ave., Seattle, WA, 98101, USA; \textsuperscript{b}Department of Urology, University of Washington, Seattle, WA, USA; and \textsuperscript{c}Madigan Army Medical Center, Department of General Surgery, Tacoma, WA, USA

KEYWORDS:
Surgical skills education; Psychomotor skills; Surgical skills assessment; Crowd sourced data

Abstract

BACKGROUND: Objective, unbiased assessment of surgical skills remains a challenge in surgical education. We sought to evaluate the feasibility and reliability of Crowd-Sourced Assessment of Technical Skills.

METHODS: Seven volunteer general surgery interns were given time for training and then testing, on laparoscopic peg transfer, precision cutting, and intracorporeal knot-tying. Six faculty experts (FEs) and 203 Amazon.com Mechanical Turk crowd workers (CWs) evaluated 21 deidentified video clips using the Global Objective Assessment of Laparoscopic Skills validated rating instrument.

RESULTS: Within 19 hours and 15 minutes we received 662 eligible ratings from 203 CWs and 126 ratings from 6 FEs over 10 days. FE video ratings were of borderline internal consistency (Krippendorff’s alpha = .55). FE ratings were highly correlated with CW ratings (Pearson’s correlation coefficient = .78, \(P < .001\)).

CONCLUSION: We propose the use of Crowd-Sourced Assessment of Technical Skills as a reliable, basic tool to standardize the evaluation of technical skills in general surgery.

© 2015 Elsevier Inc. All rights reserved.
When evaluating laparoscopic skills, an assessment tool must demonstrate reliability, validity, and feasibility to be an acceptable tool to determine competency. In our current assessment environment, most surgical residency graduates are determined to be competent surgeons by virtue of graduating from a residency program and subsequently passing their written and oral board examinations as facilitated by the American Board of Surgery. Both examinations reliably evaluate cognition and judgment but not psychomotor skill. Because of this, our profession relies on an apprenticeship model of skill acquisition without a measure that “every” surgeon has demonstrated acquisition of a safe benchmark of technical skills.

Attempts to determine psychomotor skills competency in the past have largely relied on 3 factors: the evaluation of the number of procedures performed as recorded by the Accredited Council of Graduate Medical Education resident case log system, including case type and number; the speed and relative accuracy of basic laparoscopic skills through successful completion of the Fundamentals of Laparoscopic Skills examination; and subjective evaluation of resident’s operative skills by senior surgical faculty at the resident’s institution. Caseload has limitations as a surrogate for surgical skill as the degree of involvement of the surgical trainee may be variable. Also, relying on an overall number of cases does not account for the innate skill of the surgical trainee nor the amount of surgical training they may acquire outside the surgical suite. Time-based metrics and a proprietary formula can be used for scoring trainees using the Fundamentals of Laparoscopic Surgery program offered by the American College of Surgeons, but this is expensive, labor intensive for grading, limited to specific testing sites, and somewhat less applicable to gynecology and urology. Surgical faculty are frequently used for these surgical evaluations—either real time or less frequently for a taped event. These faculty evaluations can be completed using validated instruments, or using more subjective measures. There are some limitations to this approach. Faculty frequently have competing demands, so there may be a lag time between when the case is finished and when the evaluation is completed. There may also be personal bias that intercedes, to artificially lower or raise the assessment of the trainees performance. These 3 avenues for determination of surgical competency may be inadequate and allow for inconsistency among residents, faculty, and programs across the nation. There has been some hope that virtual reality solutions may be useful mechanisms for resident assessment, but these have mostly fallen short.

The Association for Surgical Education, through a Delphi consensus process, identified determining the best methods/metrics for assessment of technical and nontechnical performance on simulators and in the operating room as one of the top 10 research priorities for 21st century surgical simulation. Objective Structured Assessment of Technical skills is a reliable and valid tool but is time intensive for expert raters, and thus making it less feasible for some centers. Because of this, virtual reality simulators have been proposed as a more efficient assessment tool. Currently, assessment tools are derived from the opinion of one or a few expert surgeons which require that inter-rater reliability be rigorously examined. These evaluation tools are certainly valuable adjuncts to current training; we set out to assess a new crowd-sourced assessment method that generates rapid, reliable, and feasible data to evaluate surgical skill.

Crowd sourcing is a recent phenomenon that utilizes anonymous crowd workers (CWs) to complete tasks. The “crowd” is a group of independent, diverse, anonymous workers that generate data by completing defined tasks. The Amazon Mechanical Turk is an online work marketplace that recruits affordable, readily available nonexperts to complete tasks. A recent study proposed that crowd sourcing may be an alternative, objective method for evaluation of operative performance and when used to evaluate robotic suturing performance found that CWs could rapidly assess skill equivalent to faculty experts (FEs). We sought to evaluate the feasibility of Crowd-Sourced Assessment of Technical Skills (C-SATS).

Methods

Seven general surgery interns were invited to participate in the pilot study on a volunteer basis. The interns were instructed and given ample time for proctored training on peg transfer, precision cutting, and intracorporeal knotting using a standard laparoscopic box trainer (three standard tasks from the Fundamentals of Laparoscopic Surgery). Interns then performed these tasks on the Electronic Data Generator for Evaluation laparoscopic trainer which has video recording capability. Each participant completed the set of tasks, individually, in a test setting away from other participants. Two facilitators were present to orient the intern, review the series of tasks before recording each task, and answer questions. Edited and deidentified videos were uploaded to a private, secure website in a standardized format. These 21 video clips were randomly evaluated by 5 volunteer surgical FEs. A standard evaluation tool was used and will be discussed in detail.

Crowd worker recruitment and selection

Amazon.com Mechanical Turk CWs, generated a minimum of 30 evaluations on each of the same 21 video clips. CWs were included based on rater performance history as well as passing a screening and attention test. Rater performance history was determined by including CWs who had greater than or equal to 95% acceptance rating on historical tasks within the Amazon Mechanical Turk crowd sourcing platform. The screening test required the CW to watch a short side-by-side video of the Fundamentals of Laparoscopic Skills block transfer and identify which video...
showed the surgeon of higher skill to assess the CW’s discriminative ability. In order to further qualify the CW and ensure their complete attention, an attention question was embedded within the survey. The attention question directed the CW to leave the question unanswered. The responses from any CW that failed either the screening or attention questions were excluded. A sample video interface can be viewed demonstrating the use of C-SATS for CW selection (Video 1).

Technical performance measures

Both evaluator groups used the Global Objective Assessment of Laparoscopic Skills (GOALS) validated rating instrument. For both experts and CWs, 5-point Likert rating assessments (1 = worst to 5 = best) were captured on 4 domains. These domains included depth perception, bimanual dexterity, efficiency, and tissue handling as dictated by the GOALS instrument. GOALS includes autonomy as a fifth domain; however, this was excluded from the assessment. All raters were also provided an overall pass or fail assessment for each video. Each rater provided comments to further describe the basis of their rating. These comments were collected and evaluated for common themes and subthemes.

Data collection

Crowd-sourced data collection was adopted from the study and methods described by Chen et al. Eligible CWs were able to access the rating task through Amazon Mechanical Turk. For the task, a custom-built web-based video display and survey tool (Zoho Corp., Pleasanton, CA) was built such that the faculty panel and CWs utilized the same interface for blinded review of the videos. This interface also time-stamped the moment a reviewer’s page loaded and all survey click times for each video review session. Total video review time was derived as the time between page load and last survey click.

Ratings from random CWs naturally vary as they do with any set of expert raters. The CWs were from a variety of countries, but most were from North America, and most had scored surgeon videos previously. For this project, we included only performance ratings from CWs who passed a calibration test and an attention test. A calibration test was used to ensure that CWs correctly chose one of the two options, where the correct choice is obvious to an expert surgeon. In each rating questionnaire, we additionally embedded a single-item attention test that begins by asking the CW “not” to answer the following question: “How well do you pay attention to details?” Ratings of CWs who responded to this question were excluded from all analyses. Missing data were otherwise not allowed for individual GOALS items in the survey and CWs were not allowed to submit or be paid for incomplete tasks.

Analytic approach

The 4 GOALS items from experts and eligible CW ratings were summed into a single numeric global performance score ranging from 4 (worst) to 20 (best). We evaluated inter-rater
reliability of FE GOALS ratings through the use of Krippendorff’s alpha statistic. Krippendorff’s alpha statistic is a measure for assessing internal reliability between raters or between items on a questionnaire when there are variable numbers of rater’s rating tasks. Because there are no known guidelines for Krippendorff’s alpha statistic, we assessed the reliability of outcomes of FE using frequently cited thresholds for Cronbach’s alpha statistic:\[ \alpha \geq .9 \text{ (excellent, high stakes testing)}, \ .9 > \alpha \geq .7 \text{ (good, low stakes testing)}, \ .7 > \alpha \geq .6 \text{ (acceptable)}, \ .6 > \alpha \geq .5 \text{ (poor)}, \text{ and } \alpha < .5 \text{ (unacceptable)}. \] Because multiple ratings per expert and CW were allowed across video performances (only one rating per video), we determined the mean crowd-based GOALS ratings and 95% confidence interval (CI) for each video/task using a linear mixed-effects model clustering on rater ID. A minimum of \( n = 31 \) eligible CW ratings were obtained to obtain a CW rating of 95% CI width of ±1 point on the GOALS rating scale. Linear mixed-effects models

Figure 2  FE GOALS ratings by video ID showing the high degree of variability within expert surgeon raters.

Figure 3  CW GOALS ratings by video ID showing the variability within CW; box plots demonstrate the central tendency of CW ratings.
derived an average CW rating and FE rating for each video clip and Pearson correlation coefficients were used to measure the strength of association between CW and FE ratings. Finally, we also assessed the duration that each CW and FE were on the web-linked survey.

**Results**

Responses were collected rapidly from participating CWs. Over a period of 19 hours and 15 minutes 803 ratings were received, of which 662 were eligible, from 203 unique Amazon Mechanical Turk CWs. In contrast, 126 ratings from 6 FEs were completed over the course of 10 days from the initial notification.

Expert GOALS ratings were highly correlated (Pearson’s correlation coefficient = .78) with CW ratings (P < .001) (Fig. 1). FE GOALS ratings were found to have borderline internal consistency across videos (Krippendorff’s alpha = .55). Fig. 2 demonstrates variation in FE ratings. Fig. 3 demonstrates the variability seen within CW ratings for each video. Fig. 4 displays the percentage of FE and CW raters who indicated that the video was endorsed with a “passing” rating (Pearson correlation coefficient = .80).

Expert and crowd comments were reviewed for similarity from a qualitative perspective using theme and sub-theme analysis. Comments were remarkably consistent between FEs and CWs. Both raters focused on feedback regarding depth perception, poorly coordinated bimanual dexterity, rough use of instruments, missing the target, inefficiency, and slow speed. From our review, the comments from either FE or CW would yield similar feedback to the learner on their final report.

**Comments**

In the rapidly changing educational environment, surgical educators have new pressures to assess new surgeons in a fair and efficient manner. This study describes the experience with a new assessment tool from a feasibility perspective. The encouraging evidence supports the conclusion that CW assessment of technical performance ratings may be a reliable surgical skills assessment tool. This initial analysis demonstrates that a crowd-sourced method for evaluation is efficient, reliable, and practical. Furthermore, it addresses the inherent problem of inter-rater and interinstitution reliability error.

Review of the expert and crowd comments revealed similar themes between both groups. The most common comments regarding efficiency were in regards to speed but also on instruments missing the target “overshooting” in both FE and CW pooled comments. Tissue handling had similarly agreeable commentary with FE and CW focusing on how “rough” the surgeon was and frequently bumping into or hitting objects in the frame. Depth perception was equally congruent in feedback from both FEs and CWs with most videos receiving comments about perceived difficulty with depth perfection and poor coordination such as dropping the needle or suture. Finally, bimanual dexterity generated comments from both FE and CW regarding an idle left hand, awkward transfers, and poor
use of both hands. Overall, comments for each task and video were highly correlated based on theme analysis.

All 5 of the expert reviewers were experienced laparoscopists and surgeons who have been training and evaluating residents for many years. The inter-rater variability between these expert surgeons highlights the need for a more consistent mechanism to provide this type of evaluation and feedback regarding trainees’ performance of commonly performed laparoscopic tasks. In addition, it emphasizes the importance of training faculty raters appropriately. We acknowledge that the expert faculty may have been potentially underprepared for their evaluative role in this setting. However, their orientation and training was identical to that of the CWs and each of the 5 expert faculty have had experience with using GOALS to evaluate resident performance at their respective institutions.

In contrast to the importance of reliability among a small group of expert raters, we feel that inter-rater reliability of CW is less vital in this study. One of the central tenets of crowd sourcing is that one has a large enough sample size to be able to accurately measure the “signal,” or a gage as to how one technical performance is relative to another. It is therefore of lesser importance what the overall distribution of CW ratings is, so long as one has enough CW ratings to measure the average with accuracy. In this study, the number of CW ratings was obtained such that for any given surgical performance the CW rating of 95% CI was within \( \pm 1 \) point on the GOALS scale.

The rationale for presenting the distribution of expert scores, and for inter-rater reliability analyses in general, is that while one can obtain a limited number of expert ratings, these expert opinions can widely vary. As such, in the setting where ratings will be obtained from multiple expert raters it is imperative that inter-rater reliability is high to establish an accurate ground truth. It is duly noted that experts in this study had less than ideal inter-rater reliability. However, these experts were given basic training in the judgment of technical assessment and had prior experience with the GOALS assessment tool. We, therefore, believe that the comparison of crowd-based assessment with expert-based assessment remains accurate in this study. Crowd sourcing offers an alternative methodology for obtaining a larger number of ratings per surgical performance, where the law of large numbers ensures that the rating will be close to the underlying truth and allows for higher confidence in the overall average rating even if variability in crowd ratings is greater than that of experts.

There are limitations to this study including a small sample size and the use of novice learners only. Certainly, future investigations should determine the ability of the crowd to stratify learners based on experience from novice to expert. This tool also requires continued vigorous evaluation of additional parameters to include reliability, ease of use, cost, ease of interpretation by faculty and learners, and impact on patient safety and surgical outcomes.

Swing et al.\(^6\) proposed a set of standards by which assessment tools should be evaluated and implemented. Van Nortwick et al.\(^7\) reviewed and made recommendations for validation study reporting in the context of a longitudinal research program approach. Both validation methodologies provide the framework for additional research to prove C-SATS is acceptable and applicable for surgical evaluation.

**Conclusions**

C-SATS has broad potential applications. Proposed applications include annual or biannual assessment of surgical resident technical skill to ensure progression of skill and provide feedback. C-SATS also allows for national benchmarking. Objective measures obtained map to Accredited Council of Graduate Medical Education Milestones for the Patient Care domains focused on open, endoscopic, laparoscopic, and robotic surgery. Additionally, C-SATS may provide a tool to ensure that graduating residents achieve acceptable psychomotor performance standards as a supplement to oral and written boards. In conclusion, this pilot study demonstrates the feasibility of Crowd-Sourced Assessment of Technical Skills (C-SATS) and its potential to standardize the evaluation of technical skills in general surgery residency training.

**Supplementary data**

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.amjsurg.2015.09.005.

**References**


