The American Journal of Surgery (2016) 211, 398-404

The American Journal of Surgery®

Association for Surgical Education

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

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

[a]*Virginia Mason Medical Center, Department of General Surgery, Mailstop H8-GME, 1100 9th Ave., Seattle, WA, 98101, USA;* [b]*Department of Urology, University of Washington, Seattle, WA, USA; and* [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.

Objective and unbiased assessment of surgical skills remains an ongoing challenge in surgical education. Rigorous psychomotor assessment has been neglected, while assessment of knowledge and judgment is commonplace using well validated, high stakes assessment tools. Competency-based assessment of surgical graduates is the future of our educational environment. To produce safe, accurate surgeons we must be able to assess their operative performance using a timely, cost-effective, reliable, and valid tool.

C-SATS served as a study sponsor and donated the use of their product for the completion of this study. In addition, three of the contributing authors: Thomas Lendvay MD FACS, Bryan Comstock MS and Justin Warren MBA are C-SATS employees who conducted the C-SATS preparation of videos and analysis of data. Interpretation of data and writing of the manuscript was a collaborative endeavor.

* Corresponding author. Tel.: +1-253-948-6920; fax: +1-206-341-0048.

E-mail address: Shanley.Deal@virginiamason.org

Manuscript received April 11, 2015; revised manuscript August 31, 2015

0002-9610/$ - see front matter © 2016 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amjsurg.2015.09.005

When evaluating laparoscopic skills, an assessment tool must demonstrate reliability, validity, and feasibility to be an acceptable tool to determine competency.[1] 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.[2] 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.[3] 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.[4] 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.[5]

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.[6] 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.[7] 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).[8–11] 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 knot-tying 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