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

Reference:

Author Institutions:
1University of Washington School of Medicine, Seattle, Washington, USA
2Department of Mechanical Engineering, University of Minnesota, Minneapolis, Minnesota, USA
3Stanford University School of Medicine, Palo Alto, California, USA
4Department of Urology, Madigan Army Medical Center, Tacoma, USA
5Department of Urology, University of Washington, Seattle, Washington, USA
6Florida Hospital Nicholson Center, Orlando, Florida, USA

Tags
Clinical domain
Medical Expert
Educational domain
Assessment
(Post)graduate
(Residency training)

Background
Lack of direct observation by faculty has been a long-standing problem in graduate medical education. Formative-focused assessment of surgical skills is critical to learner development, but current tools, such as OSATS, require time on the part of faculty. This study investigated whether a novel technique, crowd-sourced assessments of technical skills (CSATS), may supplement faculty observation and assessment of technical skills. Crowd sourcing uses a unique tool. Amazon’s mechanical turk, to allow individuals meeting minimal qualifications, to judge performance using a rating scale via the internet.

Purpose
The purpose of this study was to compare the ratings obtained from crowd-sourcing with those of expert faculty, with specific focus on reliability and correlation of crowd-sourced scores with those of expert faculty.
Type of paper

Research: psychometric design

Key Points on the Methods

1) The technical skill evaluated was suturing porcine bladder tissue using 12 videos from residents performing the procedure in a lab

2) The selection criteria for the crowd sourced group are well described, and the criteria to be included are surprisingly minimal, but require the crowd-sourcing group to correctly identify better performance on a block and box task

3) Because of mechanical turk rules, the identity of the crowd sourced raters and their characteristics is unknown

4) 50 crowd sourced rating using the GEARS tool (Global Evaluative Assessment of Robotic Skills) were compared with the ratings of 7 “expert” faculty of the skills of residents suturing a living porcine bladder tissue

5) The authors used a validated instrument to judge technical skills quality (GEARS)

6) Only reliability and correlations on a scale score were provided

Key Conclusions

The reliability of the crowd sourced GEARS rating, using Cronbach alpha, was high at 0.91. This is partially a function of the large number of raters (50) per video. There was also a strong correlation between faculty rating and the crowd sourced ratings. Based on these results, the authors claim that C-SATS could be an efficient way to supplement faculty ratings and provide residents with greater quantity and timely feedback. While intriguing, there are several caveats not mentioned by the authors. One, crowd-sourced GEARS means were higher on all 12 tapes compared to the faculty ratings, strongly suggesting leniency effects among crowd-sourced ratings. Two, crowd sourcing only provides ratings, not narrative and specific feedback. In fact, it appears such narrative feedback would be inappropriate for C-SATS given the anonymity of the raters of the Mechanical Turk. Thus it appears C-SATS may be most appropriate for identifying outliers on the low end of the skill spectrum. Low GEARS ratings from crowd-sourcing, based on this small study, suggests such low ratings could serve as a red flag for programs. However, as the authors point out, these are short videos and much work needs to be done.
Spare Keys – other take home points for clinician educators

C-SATS is an innovative, out-of-box approach to assessing technical skills, but the best use of C-SATS is yet to be determined and much research is needed before C-SATS should be included as part of a program of assessment.

Shout Out

To Holst for “think outside the box” in the assessment of technical skills.