JOURNAL OF ENDOUROLOGY
Volume 31, Supplement 1, April 2017

Pp. S-95–S-100
DOI: 10.1089/end.2016.0569

C-SATS: Assessing Surgical Skills Among Urology Residency Applicants

Simone L. Vernez, BA,[1] Victor Huynh, BS,[1] Kathryn Osann, PhD,[2] Zhamshid Okhunov, MD,[1] Jaime Landman, MD,[1] and Ralph V. Clayman, MD[1]

Abstract

Background: We hypothesized that surgical skills assessment could aid in the selection process of medical student applicants to a surgical program. Recently, crowdsourcing has been shown to provide an accurate assessment of surgical skills at all levels of training. We compared expert and crowd assessment of surgical tasks performed by resident applicants during their interview day at the urology program at the University of California, Irvine.

Materials and Methods: Twenty-five resident interviewees performed four tasks: open square knot tying, laparoscopic peg transfer, robotic suturing, and skill task 8 on the LAP Mentor™ (Simbionix Ltd., Lod, Israel). Faculty experts and crowd workers (Crowd-Sourced Assessment of Technical Skills [C-SATS], Seattle, WA) assessed recorded performances using the Objective Structured Assessment of Technical Skills (OSATS), Global Evaluative Assessment of Robotic Skills (GEARS), and the Global Operative Assessment of Laparoscopic Skills (GOALS) validated assessment tools.

Results: Overall, 3938 crowd assessments were obtained for the four tasks in less than 3.5 hours, whereas the average time to receive 150 expert assessments was 22 days. Inter-rater agreement between expert and crowd assessment scores was 0.62 for open knot tying, 0.92 for laparoscopic peg transfer, and 0.86 for robotic suturing. Agreement between applicant rank on skill task 8 on the LAP Mentor assessment and crowd assessment was 0.32. The crowd match rank based solely on skills performance did not compare well with the final faculty match rank list (0.46); however, none of the bottom five crowd-rated applicants appeared in the top five expert-rated applicants and none of the top five crowd-rated applicants appeared in the bottom five expert-rated applicants.

Conclusions: Crowd-source assessment of resident applicant surgical skills has good inter-rater agreement with expert physician raters but not with a computer-based objective motion metrics software assessment. Overall applicant rank was affected to some degree by the crowd performance rating.

Keywords: clinical competence, crowdsourcing, education, interviews, residency, validation studies

Introduction

In the United States, medical students applying to surgical residency programs are selected based on academic achievement and interpersonal skills as determined by performance in medical school, scores in the United States Medical Licensing Examination (USMLE), productivity in research, letters of recommendation, and interview performance. While a clinical skills assessment examination, the USMLE Step 2 crowdsourcing (CS) provides a standardized appraisal of basic patient-centered skills, there is no metric or method for the evaluation of surgical skills such as field perception, manual dexterity, and hand–eye coordination, vital to success in a surgical field.

The development of video-assessment techniques for evaluation of surgical skills among training and experienced surgeons has shown immense promise for the purposes of skills assessment and standardized methods of surgical certification.[1,2] Moreover, assessment of laparoscopic technical performance has been shown to correlate with postoperative patient

[1]Department of Urology, University of California, Irvine, Orange, California.
[2]Hematology-Oncology Division, Department of Medicine, University of California, Irvine, Orange, California.

outcomes, including morbidity, mortality rate, operative times, and readmission rates.[3] However, video assessment by expert surgeons is time-consuming and costly, thereby limiting its wide-spread use.[4]

Recent work in collaboration with Lendvay and colleagues at the University of Washington has demonstrated reliable surgical skills video assessment by nonexpert ''crowd'' workers using validated global assessment scores such as the Global Operative Assessment of Laparoscopic Skills (GOALS) and Global Evaluative Assessment of Robotic Skills (GEARS).[5–12] This body of work has shown that crowd assessments strongly agree with expert assessments of training and experienced surgeons on both laboratory-based and live surgical procedures, showing the ability of the nonexpert crowd to apply modified, validated assessment tools to evaluate the depth perception, bimanual dexterity, and efficiency of surgeons of all skill levels.

Motion analysis software has also shown potential for offering an inexpensive, alternative for evaluating basic surgical performance. In particular, McDougall et al.[13] demonstrated that surgical skill task 8 on the LAP Mentor™ (Simbionix Ltd., Lod, Israel) reliably distinguished between high- and low-volume laparoscopic surgeons. The screen task consists of interspersed bands requiring disruption via a right or left hook electrode that is activated by a right or left foot pedal, respectively. Once one band is cut, another band is randomly highlighted. The simulator analyzes both accuracy and efficiency of the trainee.

Skills assessment as part of the interview process for a surgical residency program may possibly identify medical student applicants with better potential for technical success in a surgical career. To this end, we sought to evaluate the agreement between expert and crowd assessment of basic open, laparoscopic, and robotic skills.

Materials and Methods

After obtaining approval from the UC Irvine Institutional Review Board, 25 urology residency interviewees at the University of California, Irvine, were asked to perform four tasks: hand-tied square knots, laparoscopic peg transfer as done in the fundamentals of laparoscopic surgery training module, robotic suturing, and skill task 8 on the LAP Mentor (Simbionix Ltd.). All interviewees had been informed of the nature of the skills tasks 2 weeks before their interview date. The interviewees were given 2 minutes to complete the square knot drill, 5 minutes for the laparoscopic peg transfer, 3 minutes for placing and tying a robotic suture, and 2 minutes to complete skill task 8. The testing was conducted in the Department of Urology's laboratory facility and was overseen by the department's laboratory personnel; at no time were the tasks overseen by any of the physicians with whom the applicants were going to be interviewed later in the day nor were the results of the skills task communicated to any of the faculty before the applicant's interview.

Individual performance videos were recorded and uploaded to the secure web platform www.csats.com and each task performance was assessed in two ways: (1) by using the Crowd-Sourced Assessment of Technical Skills (C-SATS, Seattle, WA) platform and (2) by two Urology faculty experts in the given task. Both crowd and faculty experts were asked to evaluate each of the deidentified tasks using previously validated Objective Structured Assessment of Technical Skills (OSATS), GEARS, and GOALS tools.[14–16] Web platform video training was provided to familiarize crowd workers with the tasks and required evaluation tools. Only those crowd workers with a 95% approval rating on the platform, a minimum of 100 tasks completed, and who completed the video training task instruction module were allowed to evaluate. Crowd workers were paid an average of $0.44 per task; the total cost for the crowd assessment of all four tasks was $85/applicant or $2125 for the 25 interviewees. The faculty of the Department of Urology at the University of California, Irvine, served as the uncompensated expert reviewers with a pair of individuals facile in each given task rating the applicants' performance. Crowd scores were made available to the faculty before the development of the final rank list. Expert scores were completed after the rank match list was established.

Rank on the accuracy and efficiency of motion was recorded for each applicant on the LAP Mentor skills task 8. Crowd workers were asked to evaluate performance on this task using the GOALS score. Overall crowd rank according to the GOALS score was recorded and compared with the overall rank according to the LAP Mentor simulator software.

Statistical analysis

Cronbach's alpha (α) was used to assess agreement of scoring among the crowd evaluators and between the experts for knot-tying, laparoscopic peg transfer, and robotic suturing. A similar analysis was used to contrast the overall crowd ranking and the rank generated by the simulator for skill task 8 based on the average between Efficiency and Accuracy scores. The same statistical analysis was used to determine the relationship between overall manipulative skill rankings to final match ranking. Reliability of agreement was determined based on the following threshold values of Cronbach's α: above 0.9 was considered excellent, between 0.9 and 0.7 was considered good, and below 0.5 was considered poor agreement.[17] Agreement between expert scores and crowd scores was also investigated using linear regression methods with Pearson correlations.

Results

A total of 25 medical students selected from more than 200 applicants for a residency interview participated in the study. Students were divided into five groups of four individuals and one group of five interviewees. Each group completed each of the four tasks in random succession, moving from task-to-task at 5-minute intervals. With instruction and task rotation time, all 25 students completed the skills tasks in just under 2 hours.

In total, 150 expert ratings (two per task video clip) completed by six experts were obtained within an average of 22 days (6–34 days). For the open knot tie, a total of 1606 crowd assessments for 50 videos (two separate clips per applicant) were obtained in 3 hours and 4 minutes. For laparoscopic peg transfer, 749 assessments were provided within 3 hours and 3 minutes. For robot-assisted suturing, 767 assessments were obtained in 3 hours and 26 minutes. Finally, for LAP Mentor skill task 8, 816 assessments were obtained in 3 hours and 27 minutes.

Table 1. Agreement Between Expert and Crowd Scores According to Skills Task Assessment

	Minimum	*Maximum*	*Mean*	*SD*
		Task 1: open knot tie (OSATS)		
Expert	6.36	14.5	11.24	2.29
Crowd	14.45	17.43	16.38	0.85
Agreement		Cronbach's $\alpha = 0.623$		
		Task 2: laparoscopic peg transfer (GOALS)		
Expert	3.5	13.5	8.83	2.83
Crowd	4.81	11.75	7.5	2.01
Agreement		Cronbach's $\alpha = 0.916$		
		Task 3: robotic suture (GEARS)		
Expert	4.5	14	8.15	2.80
Crowd	10.9	19.84	15.04	2.09
Agreement		Cronbach's $\alpha = 0.864$		

GOALS = Global Operative Assessment of Laparoscopic Skills; OSATS = Objective Structured Assessment of Technical Skills; SD = standard deviation.

For open knot tie, the mean score awarded by experts was 11.24 compared with 16.38 by crowds. Agreement between expert and crowd assessment scores for the open knot tie was 0.62. Mean score for laparoscopic peg transfer was 8.83 and 7.5 for experts and crowds, respectively; the agreement was 0.92. For robotic suture, mean scores were 8.15 for experts and 15.04 for the crowd with an agreement of 0.86 between experts and crowd (Table 1). Notably, for all three tasks, no applicants identified as among the five bottom performers by experts appeared among the top five performers designated by the crowd and vice versa. Inter-rater agreement between experts for open knot tie, laparoscopic peg transfer, and robotic suture was 0.72, 0.76, and 0.72, respectively; inter-rater agreement among the crowd was not available.

Figures 1–4 display a scatterplot linear regression analysis of mean expert scores relative to crowd scores. Each displays a line of best fit and Pearson correlation coefficient. For open knot tie, for laparoscopic peg transfer, and for robotic suture tasks, expert and crowd scores were positively correlated with $r = 0.69$, 0.89, and 0.79, respectively. This indicates that there is good agreement between experts and crowd despite the fact that for open knot tie and robotic suture cases, the crowd scores tend to be higher for applicants compared with expert scores. For peg transfer, however, expert and crowd scores are in close agreement.

Applicant rank according to the Simulation Task 8 software assessment and crowd assessment showed poor agreement ($\alpha = 0.32$). When subdivided into rank according to simulator efficiency score and accuracy score separately, agreement remained low ($\alpha = 0.27$ *vs* $\alpha = 0.13$).

Agreement between experts and crowd on overall student rank as determined by performance on all three surgical tasks

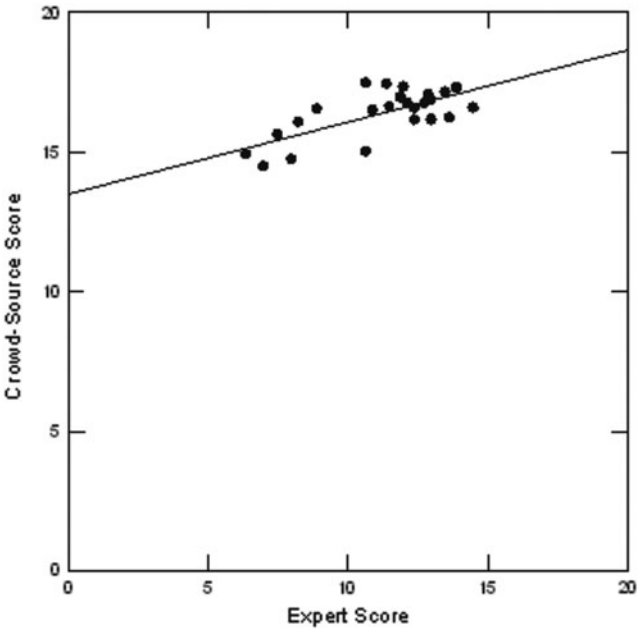

FIG. 1. Linear regression open knot tie, $Y = 13.48 + 0.258X$, $r = 0.69$.

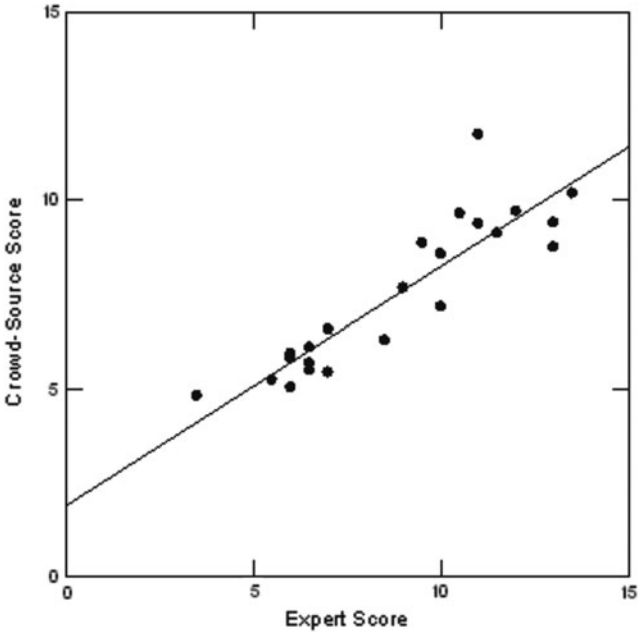

FIG. 2. Linear regression laparoscopic peg transfer, $Y = 1.87 + 0.64X$. $r = 0.89$.

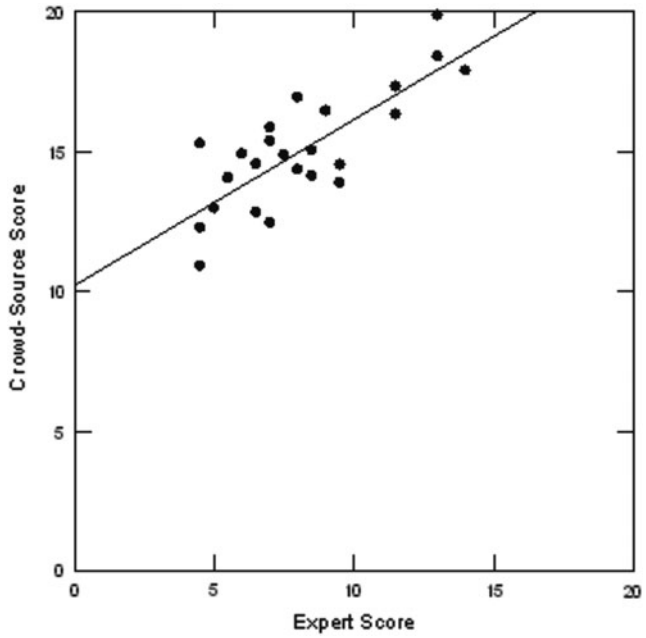

FIG. 3. Linear regression robotic suture (GEARS), $Y = 10.2 + 0.59X$. $r = 0.79$. GEARS = Global Evaluative Assessment of Robotic Skills.

was 0.83. Neither expert nor crowd rank based on skills testing performance showed good agreement with the overall match rank list, which also took into account academic record and interview performance ($\alpha = 0.46$) (Table 2). However, of those applicants ranked within the bottom five on the match list, three out of five were identified as poor performers by the crowd (i.e., students ranked 23, 24, and 24 on the match list were 25, 22, and 24, respectively, according to the crowd [Table 3]). Interesting, although after the fact, just 2 out of 5

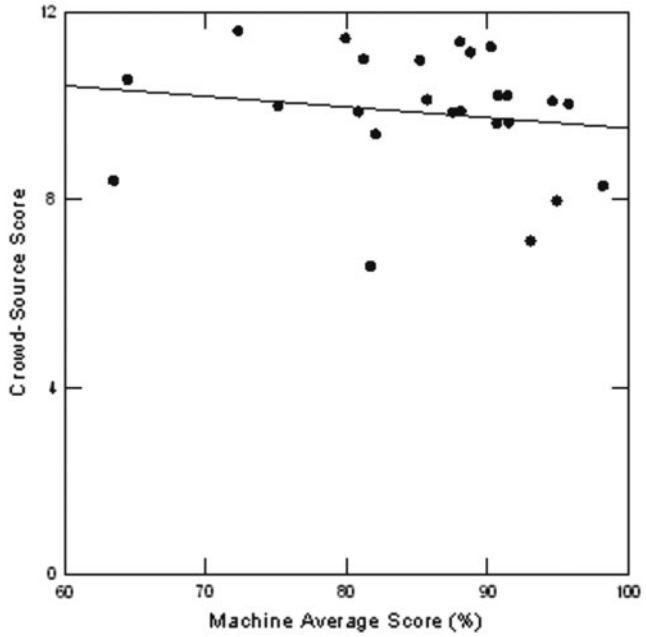

FIG. 4. Linear regression simulation task (GOALS), $Y = 11.8 - 2.28X$. $r = -0.16$. GOALS = Global Operative Assessment of Laparoscopic Skills.

TABLE 2. RANK ACCORDING TO SKILLS TASK PERFORMANCE

	Cronbach's α
Crowd *vs* expert	0.830
Expert *vs* ultimate match rank	0.485
Crowd *vs* ultimate match rank	0.464

of the bottom 25 were ranked among the bottom-most performers by the faculty experts.

Discussion

Currently, surgical residents are selected on the basis of USMLE scores, letters of recommendation, as well as completing an away rotation at the program in question.[18] Manipulative skills are not assessed and only rarely are commented upon in letters of recommendation.

Innovative assessments of surgical skill such as video assessment show promise for the creation of a new paradigm for evaluation at all levels of surgical experience.[4] However, a major impediment is that current modalities lack standardization and are both costly and time-consuming. CS, a term coined by Jeff Howe in 2006,[19] is a method of leveraging the power of many, attentive, although nonexpert, online community members, to a variety of challenges in an inexpensive and efficient manner. CS has been utilized within surgical disciplines to evaluate residents as well as established, practicing surgeons.[5–12] Herein we tested whether the crowd could assess skills similar to expert surgeons among a group of applicants to a surgical program; we further hypothesized that such information might be of value in the subsequent faculty-generated match rank list.

The crowd rapidly provided surgical skills assessment on three basic dry laboratory surgical skills tasks with fair to good agreement when compared with expert reviewer scores. Previous studies have indicated that CS could be used to reliably differentiate skill levels among training and practicing surgeons, consistently identifying the lowest performers, with application to even more dynamic environments, namely, live surgical settings.[10–12] Indeed, our results showed that the crowd consistently identified top and bottom performers in open, laparoscopic, and robotic settings, suggesting that surgical skills assessment of medical student applicants via CS provides assessments comparable to that of faculty members.

TABLE 3. BOTTOM FIVE ON THE MATCH RANK COMPARED WITH CORRESPONDING OVERALL RANK ON SKILLS PERFORMANCE (CROWD AND EXPERT)

Match rank	*Corresponding crowd rank*	*Corresponding expert rank*
21	4	2
22	8	6
23	25	25
24[a]	22	20
24[a]	24	24

[a]Only 23 out of 25 applicants were ultimately ranked for the match. Those ranked 24 in this analysis are the applicants who were not offered a rank position to match into a resident position, corresponding to the lowest ranking on the overall match rank.

Previously, McDougall et al.[13] demonstrated that motion assessment scores on skill task 8 on the LAP Mentor simulator differentiate between high- and low-volume laparoscopic surgeons. A larger, subsequent study showed that skill task 8 effectively distinguished between nonlaparoscopic surgeons, novice, and experienced laparoscopic physicians.[4] Our findings do not indicate that the simulator software and crowd assessment show acceptable agreement among medical students. These findings also are in contradistinction to a recent study conducted by the Basic Laparoscopic Urologic Skills consortium, which found good correlation between the EDGE simulator objective motion assessment software (EDGE device; Simulab Corp., Seattle, WA) and crowd assessment.[10]

A potential explanation for the lack of agreement between simulator and crowd assessments of medical student applicants is the fact that each applies a different assessment tool, focused on different task elements. In our study, ranks determined by the simulator software were made according to efficiency of movement and accuracy of task completion. Crowd assessments were based on the parameters outlined in GOALS, including tissue handling and bimanual dexterity, neither of which is specifically taken into consideration in objective motion analysis. Furthermore, GOALS has not been validated in the evaluation of skills in a computer-simulated virtual reality environment.

Interestingly, according to our findings, it is clear that the process currently used to select resident applicants is discordant with overall rank by surgical skills as measured by the basic tasks described in this study. Indeed, the final rank list is based on many factors, including the academic track record, USMLE scores, letters of recommendation, personal background and, importantly, overall ''fit'' of the student with the department. However, three out of five of the bottom five performers identified by CS fell into the bottom 20% on the match list. As faculty had access to crowd scores when deciding rank list, it follows that performance scores may have impacted the decision process for the match, especially when determining the lowest ranking applicants. The present study indicates that CS can feasibly be used to assess the basic open, laparoscopic, and robotic surgical skills of resident applicants in a dry laboratory setting. This evaluation may be easily incorporated into the resident selection process, allowing timely, relatively low-cost appraisals of applicants' technical ability. In the future, by establishing selection criteria based on hands-on, practical skills, applicants who are best suited to a surgical career may be selected. Moreover, those without surgical aptitude may be identified early and encouraged to pursue alternate career paths.

Urology is a unique surgical subspecialty that combines both office-based and surgical skill. It has been suggested that urology residency may benefit from further subdivision between office-based and surgical urology to more effectively meet the clinical needs of the urologic patient population.[20,21] While thought-provoking and, as shown in this study, feasible, it still remains to be seen whether an applicant's manipulative skills as a medical student will be predictive of their subsequent performance during their residency and thereafter. This hypothesis will require long-term follow-up to test whether these early results correlate with the effective completion of residency and, among the graduates, predict which ones self-select for a career doing major urologic surgery *vs* a primarily office-based practice.

There are several weaknesses in our study. Namely, this represents a primary feasibility study with only 25 participants. Indeed, it will take several years to accrue a sufficient number of participants to create a truly robust statistical analysis. Second, we were not able to obtain questionnaires from a sufficient number of participants to judge how much surgical exposure they had or whether they had practiced any or all of the tasks before their interview. Thus, it was not possible to infer how the level of skill may have related to experience as opposed to inherent skill. In the future, this information will be obtained at the time of the interview process. Third, educational characteristics that may be relevant to a crowd worker's ability to learn and score tasks could not be identified. Chen et al.[5] previously found that that the use of specific language cues in free-text analysis of surgical tasks correlated with a higher likelihood for the crowd worker to agree with expert reviews. Additional screening measures may improve the reliability of crowd scores in relation to faculty review. Fourth, the true value of this type of screening is dependent on the accrual of long-term follow-up data with regard to discerning any correlation between the performance on the interview skills tasks and subsequent outcome of residency training and career path. Indeed, an individual's ''grit'' in training may overcome innate technical deficiencies and may result in greater long-term technical success.

Conclusions

CS may be reliably used to generate an assessment of surgical skills among medical student applicants to a surgical residency program. Overall applicant rank was, in part, determined by performance on the basic surgical skills tasks; however, given other factors that were taken into account, it in and of itself did not correlate well with ultimate match rank.

Acknowledgments

This study was aided by the laboratory endeavors of Renai Yoon, Christina Hwang, and Tyler Valdez.

Author Disclosure Statement

No competing financial interests exist.

References

1. Scott DJ, Vlantine RJ, Bergen PC, et al. Evaluating surgical competency with the American Board of Surgery In-Training Examination, skill testing, and intraoperative assessment. Surgery 2000;128:613–622.
2. Matsuda T, Ono Y, Terachi T, et al. The endoscopic surgical skill qualification system in urological laparoscopy: A novel system in Japan. J Urol 2006;176:2168–2172.
3. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, et al. Surgical skill and complication rates after bariatric surgery. N Engl J Med 2013;369:1434–1442.
4. Matsuda T, McDougall EM, Ono Y, et al. Positive correlation between motion analysis data on the LapMentor virtual reality laparoscopic surgical simulator and the results from videotape assessment of real laparoscopic surgeries. J Endourol 2012;26:1506–1511.

5. 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.
6. 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.
7. 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.
8. Holst D, Kowalewski TM, White LW, et al. Crowd-sourced assessment of technical skills: An adjunct to urology resident surgical simulation training. J Endourol 2015;29:604–610.
9. Polin MR, Siddiqui NY, Comstock BA, Hesham H, Brown C, Lendvay TS, Martino MA. Crowdsourcing: A valid alternative to expert evaluation of robotic surgery skills. Am J Obstet Gynecol 2016 [Epub ahead of print]; DOI: 10.1016/j.ajog.2016.06.033.
10. Kowalewski TM, Comstock B, Sweet R, et al. Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. J Urol 2016;195:1859–1865.
11. 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.
12. Powers MK, Boonjindasup A, Pinsky M, et al. Crowd-sourcing 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.
13. McDougall EM, Corica FA, Boker JR, et al. Construct validity testing of a laparoscopic surgical simulator. J Am Coll Surg 2006;202:779–787.
14. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, et al. Objective assessment of technical surgical skills. Br J Surg 2010;97:972–987.
15. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. Am J Surg 2005;190:107–113.
16. Goh AC, Goldfarb DW, Sander JC, Miles BJ, Dunkin BJ. Global evaluative assessment of robotic skills: Validation of a clinical assessment tool to measure robotic surgical skills. J Urol 2012;187:247–252.
17. Cronbach LJ, Shvelson RJ. My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas 2004;64:391–418.
18. Weissbart SJ, Stock JA, Wien AJ. Program directors' criteria for selection into urology residency. Urology 2015;85:731–736.
19. Wired 14.06: The rise of crowdsourcing. Available at www.wired.com 2009-01-04. (Last accessed on 2015-07-02).
20. Lange PH. Are we training enough surgeon scientists to secure the future of urology? AUA News 2008;13:1.
21. Steers WD, Schaeffer AJ. Is it time to change the training of urology residents in the United States? J Urol 2005; 173:1451.

Address correspondence to:
Ralph V. Clayman, MD
Department of Urology
Dean-Emeritus
University of California, Irvine
333 City Boulevard West, Suite 2100
Orange, CA 92868

E-mail: rclayman@uci.edu

Abbreviations Used

CS = crowdsourcing
CSATS = 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
SD = standard deviation
USMLE = United States Medical Licensing Examination