Research Article ISSN 2689-1093

Surgical Research

# Value of the Crowd-Sourced Assessment of Technical Skills (C-SATS) Platform in Surgical Procedures: A Systematic Review of Evidence

**Tommaselli Giovanni A, Sehat Alvand J, Ricketts Crystal D, Clymer Jeffrey W* and Grange Philippe**

*Ethicon, Inc., Cincinnati OH USA.*

***Correspondence:**

Jeffrey W. Clymer, Ethicon Inc., 4545 Creek Road, Cincinnati, OH 45242, Tel: 513 337-3318.

**Received:** 27 Jun 2022; **Accepted:** 12 Aug 2022; **Published:** 16 Aug 2022

**Citation:** Giovanni TA, Alvand SJ, Crystal RD, et al. Value of the Crowd-Sourced Assessment of Technical Skills (C-SATS) Platform in Surgical Procedures: A Systematic Review of Evidence . Surg Res. 2022; 4(2): 1-12.

## ABSTRACT

***Background:*** *Crowd-Sourced Assessment of Technical Skills (C-SATS) is a surgical data management and learning platform that leverages the knowledge of large expert surgeon and lay groups to assess the technique and technical skills of surgeons in a highly efficient manner. The aim of this systematic review was to summarize published literature on the performance of C-SATS as compared to expert evaluations and assess its use as a training and validation tool in minimally invasive surgery (MIS).*

***Methods:*** *A comprehensive literature search was performed per PRISMA guidelines using the Medline, Embase, and Google Scholar databases on published studies that evaluated the use of C-SATS following MIS, such as laparoscopic or robotic-assisted surgery.*

***Results:*** *A total of 21 reports were included in the review. Twelve studies comparing crowd-sourcing evaluations against expert opinion indicated overall excellent or good correlation with Global Operative Assessment of Laparoscopic Skills (GOALS), Global Evaluative Assessment of Robotic Skills (GEARS), and Robotic Objective Structured Assessment of Technical Skills (R-OSATS) scores, with correlation coefficients (Pearson or Spearman) ranging from 0.69 to 0.95 and reliability index (Cronbach's alpha) from 0.63 to 0.93 across different specialties and surgical approaches. When using C-SATS to assess performance and validation, assessments positively correlated with traditional methods of time and error-based scoring and global rating scale.*

***Conclusions:*** *Based on the current published literature, the C-SATS platform has been shown to efficiently provide crowd-sourced evaluations that correlate favorably with expert evaluation across a range of surgical specialties and approaches. Use of crowdsourcing has uniformly yielded accurate evaluations of surgeons' technical skills in a markedly shorter time than expert reviews. C-SATS may be a cost-effective complement or alternative to traditional models of evaluating surgical proficiency. Future studies are needed to determine whether the use of C-SATS will lead to improved surgical performance and patient outcomes.*

## Keywords

Surgery, C-SATS, Health care systems.

## Introduction

Surgery is a foundational component of health care systems, with 313 million surgeries performed annually all over the world [1]. Although surgery has the potential to improve and prolong lives, often there are complications that can lead to further morbidity and mortality. It has been estimated that at least 4.2 million people worldwide die within 30 days of surgery each year and that the number of postoperative deaths account for 77% of all health-related deaths globally [2], making surgery the third greatest

contributor to deaths [1]. It has also been demonstrated that adverse events (AEs) occur in 14.4% of patients undergoing surgery [3]. A number of AEs are potentially preventable because they can be attributed to technical errors in the surgeon's performance. A recent study indicated that surgical errors were associated with a 27% increased risk of adverse events, a 5% increased rate of prolonged length of stay, and a doubling of incurred costs [4]. Notably, 98% of the patients judged to have had a surgical error suffered an adverse event. Thus, improving the quality of surgical care systems is a pivotal goal to protect the health and lives of patients.

In this era of increased patient awareness around surgical outcomes and fast-paced technology introduction, there is a need to evaluate the performance of surgeons to improve patients' outcomes through consistent training during the early years of their career as well as continued improvement of their surgical skills. Surgical trainees have reported the need to learn evolving surgical techniques in an efficient model [5,6]. Continuing medical education (CME) has been implemented to train new and experienced surgeons, however the CME model was not designed to identify or address individual needs of rapidly changing practices throughout their careers [7]. Although there are a variety of surgical performance evaluation tools, including basic measures (path length, time, economy of motion, mistakes, and errors), structured human assessments, such as Objective Structured Assessment of Technical Skills (OSATS), Global Evaluative Assessment of Robotic Skills (GEARS) scores, and algorithmic assessment using machine learning algorithms [8-12], there is no single standard of surgical skill evaluation.

The expert evaluation approach is the most widely used assessment that leverages the experience and skills of senior surgeons to provide qualitative feed-back. The main issues related to this type of evaluation are the need for significant time and resources from experienced clinicians as well as organizational problems related to the presence of experts in the operating room and management of videos needed for the evaluation.

C-SATS, which stands for Crowd-Sourced Assessment of Technical Skills, is part of the Johnson and Johnson family of companies. C-SATS is a Software as a Service (SaaS), cloud-based, device-agnostic, surgical data management and learning platform. C-SATS was designed to empower surgeons with an AI-driven digital solution to efficiently and consistently gain objective, expert and crowd-sourced feedback and clinical analytics.

The C-SATS platform captures videos of minimally invasive surgery (MIS) procedures and then after the surgical procedure, securely uploads them to the surgeon's cloud-based, Health Insurance Portability and Accountability Act (HIPAA)-compliant and Health Information Trust Alliance Common Security Framework (HITRUST CSF®) certified C-SATS private and personalized video library. Artificial Intelligence (AI) algorithms remove Personal Health Information (PHI) from surgical cases prior to storage in the C-SATS private video library and step segmentation is available for different procedures. These steps ensure the privacy of the patients. Through this secure, cloud-based portal, case videos can be submitted for review by both expert surgeons and crowd-sourced reviewers. Unbiased assessment is received based on the performing surgeon's skill and technique. C-SATS users can monitor personal metrics and trends over time using a private dashboard, while access is also provided to expert coaching and peer-to-peer support. The C-SATS process is depicted in Figure 1.


**Figure 1:** Flow diagram for the C-SATS work process.

Operative time can be plotted for all completed cases and trend lines displayed with a minimum of 5 cases uploaded, helping to identify outlier cases. Trends may also be displayed for selected focus steps. AI with Natural Language Processing (NLP) can aggregate expert feedback by sentiment and streamline qualitative comments.

Qualitative feedback on skill and technique is available from a community of over 350 surgical experts across a dozen specialties with over 17,000 validated case videos, while crowd-source evaluation uses web-based services via Amazon Mechanical Turk [13]. C-SATS maintains a confidential cloud-based platform and also complies with applicable standards, implementation specifications, and requirements of the Health Insurance Portability and Accountability Act of 1996 (HIPAA). Objective performance assessments are generated by crowd-sourced reviewers via scores of GEARS/GOALS domains by case, as well as individual domain scores for each step in the procedure. Global Evaluative Assessment of Robotic Skills (GEARS) includes depth perception, bimanual dexterity, efficiency, force sensitivity, and robotic control. Global Operative Assessment of Laparoscopic Skills (GOALS) includes depth perception, bi-manual dexterity, efficiency, and tissue handling.

The aim of this systematic review is to summarize and evaluate published literature on C-SATS used as a validation tool or to

assess surgical skills in MIS (laparoscopy and robotically-assisted laparoscopy).

## Methods

### Search Strategy

A comprehensive literature search was performed according to the 2020 Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) guidelines [14]. To identify published articles reporting studies on the Crowd-Sourced Assessment of Technical Skills (C-SATS) software platform for the evaluation of surgical proficiency during laparoscopic or robotically-assisted laparoscopic procedures, terms including simple text or subject subheadings with the language limited to English and keywords including "crowdsourcing", "crowd sourced", "crowdsource", "crowd source", "crowdsourced", "competitive behavior", "collective intelligence", "collective wisdom", "crowd science", "citizen science" AND "surgery" and "surgical procedure" were used to search Medline, Embase, and Google Scholar databases. A hand search of the bibliographies and citation lists of all relevant reviews and primary studies was performed to identify articles not captured by the electronic searches. No ethical approval was requested because the study is a systematic review. The search included publications until September 7th, 2021.

### Eligibility Criteria

Articles were eligible for inclusion if: 1) they were published on a peer-reviewed journal article representing original health research, 2) methodology and results were included, 3) crowdsourcing or any other form of evaluation of surgical performance in laparoscopy or robotic surgery using C-SATS was used to obtain all or part of the results, 4) involved using validated methods to assess surgical skills. The exclusion criteria were as follows: (1) did not include laparoscopy or robotic surgery, (2) articles not reporting metrics on the performance of C-SATS, (3) reviews, editorials, posters, congress abstracts, and expert opinion articles, and (4) language other than English.

### Data Extraction

Two independent reviewers extracted relevant recommendations from each study (GAT, AJS). Disagreements concerning data extraction were resolved by discussion and consensus. Thereafter, a recommendation matrix was constructed. The following variables were extracted from the articles: list of authors, title of the article, publication date, type of study, type of surgical specialty, type of surgical procedure, assessment method, main parameter recorded, details on participants, main results (including any evaluation score mean or median along with either standard deviation or 95% confidence intervals, any correlation value between crowdsourced and expert evaluation), and author's considerations and conclusions.

### Data Synthesis and Analysis

A preliminary synthesis was performed of the extracted data to categorize and itemize the different studies. The results were then summarized in a narrative synthesis.

## Results

### General Information

The PRISMA flow diagram of the study selection process is shown in Figure 2. Database search yielded 6069 citations which were reduced to 739 after having excluded duplicates. After reviewing the abstracts, we identified 37 potentially relevant studies for which full text was retrieved and all underwent detailed review. Of these studies, 16 were excluded, two because the C-SATS platform was not used, seven because the focus was not laparoscopy or robotic surgery, and seven because they were reviews. Thus, a total of 21 reports were included in the review [15-35].

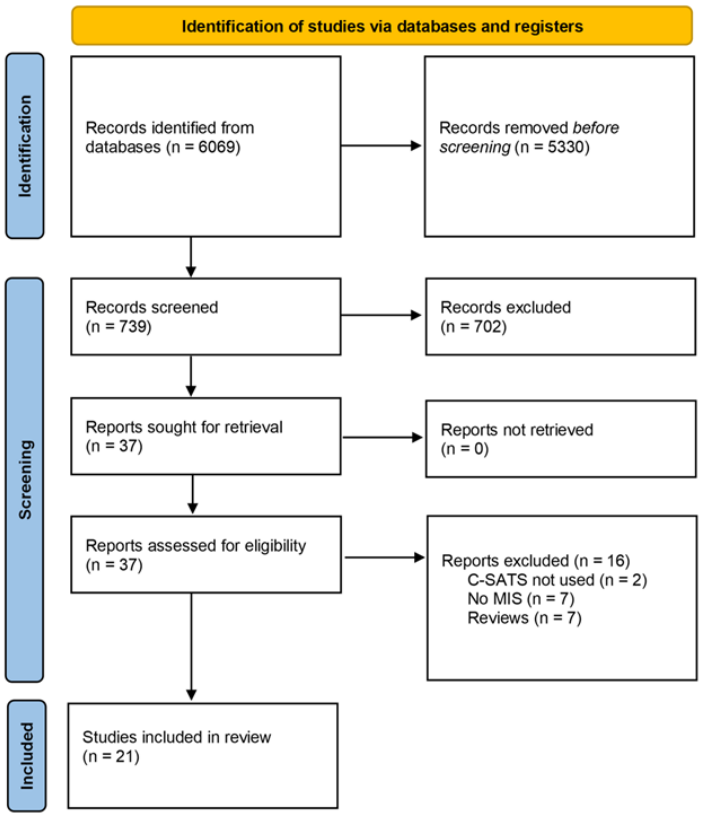


**Figure 2:** PRISMA flow diagram for article selection.

### Description of the studies

Details of the included studies are reported in Table 1.

Regarding country of origin, sixteen articles were from the United States of America [15,17-24,26-28,31,32,34,35], three were from Canada [16,25,30] and two were published jointly by centers from USA and Canada [29,33]. According to study design, three were retrospective studies [20,21,25], two adopted a form of randomization [16,23], while sixteen studies were prospective observational [17,24,30], or comparative studies [15,18,19,22,26-29,31-35].

Thirteen studies focused on robotic approach [15-18,23-27,31-33,35], six studies on laparoscopic approach [19-21,29,30,34], and

**Table 1:** Characteristics of the included studies.

| Author | Year | Country | Type of study | Approach | Specialty | Procedure | Aim | System | Participants | Main parameters collected |
|---|---|---|---|---|---|---|---|---|---|---|
| Addison | 2020 | USA | Prospective observational | Robotic | Bariatrics | Sleeve gastrectomy | Evaluate correlation between operative time and GEARS score | C-SATS | 68 patients included | Operative time<br>Overall and step specific GEARS score (ligation of short gastric vessels; gastric transection, oversewing of staple line) |
| Almarzouq | 2020 | Canada | RCT | Robotic | Urology | Radical prostatectomy | Establish competency cutoffs on the da Vinci Simulator for basic robotic skills. Correlate global scores from the da Vinci Simulator for each task with the GEARS evaluations. Assess transferability of basic robotic skills from the da Vinci Simulator to the OR using the GEARS evaluations. | Da Vinci Surgical Skills Simulator<br>C-SATS | 14 residents randomized to two groups. A - required to practice three sessions on the simulator; B - required to practice until competency was achieved. | Both groups were recorded while performing both on the simulator and during bladder mobilization and UVA during RARP. Recordings were assessed blindly using GEARS tool by C-SATS. |
| Bendre | 2020 | USA | Prospective | Robotic | Urology | Simulation | Face validation of a robotic-assisted pyeloplasty simulation using a 3D-printed of UPJO and objectively assess surgical performance and learning outcomes using C-SATS | 3D printed model of kidney, dilated renal pelvis and ureter, da Vinci Si Robotic system,<br>C-SATS | 11 participants (8 urology residents, PGY-3 to PGY-5 and 3 faculty, fellowship-trained in robotic surgery and with previous pyeloplasty experience) | Time to complete anastomosis<br>Face validity scale<br>GEARS |
| Chen | 2014 | USA | Prospective | Robotic | N/A | Simulation | To demonstrate that crowd-sourced is equivalent to experienced surgeons and to explore a link between the language of the crowd and more accurate ratings of surgical performances. | AMT, Facebook<br>C-SATS | 409 AMT Crowd workers<br>67 Facebook users<br>9 Expert robotic surgeons | Survey derived from GEARS (depth perception, bimanual dexterity, and efficiency).<br>Free text description of the rating |
| Deal | 2016 | USA | Prospective | Lap | General | Simulation | To evaluate the feasibility of Crowd-Sourced Assessment of Technical Skills of general surgery interns on simulators compared to experts | AMT, C-SATS | 203 AMT Crowd workers<br>5 Experts<br>21 video clips | Both evaluator groups used GOALS excluding autonomy as well as rating a pass or fail question |
| Deal | 2017a | USA | Retrospective analysis of public videos | Lap | General | Cholecystect. | To evaluate if CW could assess video-based surgical achievement of the CVS during lap cholecystectomy similar to experts and if this would correlate with surgeon technical performance. | AMT, C-SATS | 617 Crowdsourced workforce - 160 videos<br>5 field expert - 40 videos | GOALS score (excluding autonomy)<br>CVS score (0-6) |
| Deal | 2017b | USA | Retrospective analysis of public videos | Lap | General | Cholecystect. | To evaluate the frequency of completion of the CV and technical performance in videos of cholecystectomy posted on public websites using crowd sourcing. | Youtube, Vevo, SAGES sites<br>Amazon Mechanical Turk<br>C-SATS | 160 videos | CVS rating<br>GOALS (excluding autonomy)<br>Trends between video characteristics (likes, dislikes, views, etc.) and primary performance measures |
| Deal | 2020 | USA | Prospective | Open, Lap | General | Lap: cholecystect., colectomy, inguinal hernia repair<br>Open: ventral and inguinal hernia repair, thyroidectomy | To assess how crowd and intraoperative attending ratings using OSATS or GOALS correlate with SIMPL Zwisch and Performance scales. | SIMPL app<br>C-SATS | 32 surgical procedures recorded by trainee PGY5<br>Raters: attending surgeons and AMT | SIMPL Zwisch and Performance rating scales<br>OSATS (without autonomy)<br>GOALS (without autonomy) |
| Dubin | 2017 | USA | RCT | Robotic | General, urologic, OB/ GYN | Simulation | To evaluate whether robotic VR simulator performance metrics correlate to validated human reviewer GEARS assessments on a basic VR exercise. | dV-Trainer (dVT)<br>da Vinci Skills Simulator (dVSS)<br>C-SATS | 65 participants randomized to use dVT or dVSS first and then cross-over. Warm-up exercise, then Rail and Rod 1 on both simulators. | Simulator metrics<br>GEARS score (without autonomy) |
| Ghani | 2016 | USA | Prospective | Robotic | Urology | Radical prostatectomy | If peer surgeon assessment of technical quality is feasible.<br>If peer and CWs could distinguish differences in technical skills among practicing surgeons. | Video recording<br>C-SATS<br>AMT | 42 urologists provided videos.<br>At least 4 peer reviewers among a pool of 25.<br>30-55 CWs | GEARS score<br>RACE score<br>Summary judgment question for overall skill |
| Goldenberg | 2020 | Canada | Retrospective, case-control | Robotic | Urology | Radical cystectomy | To assess the utility of video review for quality improvement in RARC with UIS as an outcome.<br>To assess if expert or crowd-sourced video review could predict UIS. | Video Recordings<br>C-SATS | 9 cases (10 strictures) 10 controls: contralateral ureter that did not develop a stricture among the same patients (n=8).<br>3 high-volume robotic surgeons; 5 expert raters: 2142 CWs | Five-part questionnaire to evaluate each step of UIS.<br>GEARS score (without autonomy) |
| Holst | 2015a | USA | Prospective | Robotic | Urology | Robotic FLS intracorporeal suturing module | If CWs can discriminate surgical skills as well as expert faculty surgeons | Robotic FLS block transfer task<br>C-SATS | 206 AMT, 3 expert faculty reviewing 5 videos of 3 urology residents and 2 urology faculty | 3 Technical skills domains<br>Time of evaluation crowd vs expert faculty<br>Inter-rater reliability scores |
| Holst | 2015b | USA | Prospective | Robotic | Urology | Live porcine urinary bladder closures | Show that untrained crowdsourcing can discriminate surgical skills on real, living viable tissue | Robotic live bladder closure recorded videos<br>C-SATS | 487 AMT, 7 expert surgery graders | 5 GEARS domains<br>Time of evaluation crowd vs expert faculty<br>Inter-rater reliability scores |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Karani | 2021 | USA | Prospective | Open, Lap, Robotic | Urology | Simulation | Determine if surgical skills could be reliably assessed via crowdsourcing<br>How does surgical skills testing impact resident selection | Surgical skills laboratory<br>C-SATS, GEARS, OSATS, GOALS<br>interview score, USMLE score, JSPE-S, Surveys post-match | 94 urology residency interviewees<br>Crowd workers<br>2 faculty urologists | Correlation between skills testing scores and applicant metrics (interview score, JSPE-S, USMLE score)<br>Agreement between faculty & crowd<br>Predictors of match rank<br>Survey responses |
| Kowalewski | 2016 | USA/ Canada | Prospective | Lap | Urology | Simulation | If crowdsourcing can discriminate surgical skills as well as expert faculty surgeons | Pegboard and suturing skills lab<br>C-SATS, GOALS, OSATS, EDGE | 1438 AMT CW<br>454 recordings of medical students, urology residents, fellows, and faculty from 8 academic urology residents | 5-point rating assessment on 4 domains (depth perception, bimanual dexterity, efficiency, and tissue handling). Pass-fail cutoff scores maximizing sens. vs spec. in a ROC curve. Time of evaluation crowd vs expert faculty.<br>Inter-rater reliability scores |
| Lee | 2016 | Canada | Prospective | Lap | Urology | Simulation | Validate the AUA BLUS skill tasks for assessing basic lap skills of urology trainees, compare traditional and novel technical skill assessment methods, and set pass-fail standards for basic lap skills competency | Pegboard and suturing lab<br>C-SATS, time and error-based scoring (TE), GRS (expert faculty global rating) | 6 attending urologists and 99 trainees | Performance scores (TE, GRS, C-SATS)<br>Pass-fail ratings by 2 methods (norm referenced vs criterion) |
| Martin | 2020 | USA | Prospective observational, pre-post design, multi-institutional rater-blinded trial | Robotic | Urology, GYN, General surgery | Simulation | If robotic surgery novices would improve technical skills after completing FRS training on the RobotiX Mentor, and to compare the effectiveness of FRS across platforms | RobotiX Mentor<br>Pre-test avian model<br>daVinci<br>GEARS<br>C-SATS | 20 residents/novices | Pre & post-test performance inter-rater reliability (C-SATS vs 2 expert GEARS scores); Psychomotor checklist of RobotiX particpants vs psychomotor checklist from residents from Satava study |
| Polin | 2016 | USA | Prospective comparative | Robotic | GYN | Simulation | Determine if crowdsourcing can be used as an alternative to expert evaluators to evaluate robotic surgery skills | R-OSATS<br>C-SATS | 448 CWs<br>105 residents, fellows, and expert robotic surgeons<br>3 expert faculty scorers | Linear mixed effects models, Pearson correlation coefficients between CWs and expert ratings |
| Powers | 2016 | USA/ Canada | Prospective | Robotic | Urology | Partial nephrectomy | How crowdsourcing score performances of live renal hilar dissection compared to expert surgeons | GEARS<br>C-SATS | 3 robotic surgeons (residents & attendings)<br>AMT | Inter-rater reliability of GEARS scores<br>Pearson correlation coefficient |
| Vernez | 2017 | USA | Prospective | Lap | Urology | Simulation | Determine whether crowdsourcing videos of resident applicants could aid in selection process of future residents | LAP Mentor<br>C-SATS, GEARS GOALS, OSATS | 25 resident applicants<br>6 Faculty experts<br>CWs | C-SATS, OSATS, GEARS scores; Efficiency of motion data from LAP mentor vs CWs GOALS score |
| White | 2015 | USA | Prospective | Robotic | Urology<br>Gen surg<br>OB/ GYN | Simulation | Determine if crowdsourcing assessments and surgeon assessments correlated | daVinci robot<br>C-SATS<br>GEARS | 30 crowd workers<br>3 faculty experts<br>49 resident & faculty surgeon videos | Crowd vs Surgeon GEARS scores with correlation coefficients<br>Inter-rater reliability comparing 3 experts<br>Cost of crowd vs surgeon |

GEARS: Global Evaluative Assessment of Robotic Skills, C-SATS: Crowd-Sourced Assessment of Technical Skills, UVA: Urethro-vesical anastomosis, RARP: Robot-assisted radical prostatectomy, UPJO: Uretero-pelvic junction obstruction, PGY: Post-Graduate Year, AMT: Amazon Mechanical Turk, GOALS: Global Objective Assessment of Laparoscopic Skills, CW: crowd-worker, CVS: critical view of safety, OSATS: Objective Structured Assessment of Technical Skills, SIMPL: System for Improving Procedural Learning, VR: Virtual Reality, RACE to Robotic Anastomosis and Competency Evaluation, RARC: Robot-assisted radical cystectomy, UIS: uretero-ileal stricture, FLS: Fundamentals of Laparoscopic Surgery, SVD: seminal vesicle dissection, AA: anterior vesico-urethral anastomosis, LAD: lymph node dissection, USMLE: United States Medical Licensing Examination, JSPE-S: Jefferson Scale of Physician Empathy for Students, ROC: Receiver Operating Characteristic, FRS: Fundamentals of Robotic Surgery.

**Table 2:** Outcomes of studies comparing experts versus crowd-sourced workers skill assessments using C-SATS.

| Reference | Raters | Videos | No. ratings/time | Mean scores (SD) | Correlation experts/CWs |
|---|---|---|---|---|---|
| Chen | 9 Experts<br>406 CWs<br>67 Facebook users | NR | Experts: NR/24 days<br>CWs: NR/5 days<br>Facebook users: 24 days | GEARS score<br>Experts: 12.11 (1.45)<br>CWs: 12.21 (2.35)<br>Facebook users: 12.06 (2.01) | Expert: 95%CI 11.00-13.22<br>CWs: 95%CI 11.98-12.44<br>Facebook: 95%CI 11.56-12.55 |
| Deal 2016 | 6 Experts<br>203 CWs | 21 | Experts: 126/10 days<br>CWs: 662/19 hrs 15 mins | GOALS score<br>NR | PCC: 0.78 ($p<0.001$) |
| Deal 2017a | 5 Experts<br>617 CWs | 40/160 | Experts: 200/5 days<br>CWs: 8462/24 hrs | CVS score<br>Data reported for each video score group | SCC: 0.8 ($p<0.001$) |
| Ghani | 4 Experts<br>30-55 CWs | 12 | Experts: 318 GEARS/15 days<br>33 RACE/15 days<br>CWs: 2531 GEARS/21 hrs<br>459 RACE/38 hrs | GEARS score range<br>Experts: 15.8-21.7<br>CWs: 20.9-19.1 | PCC<br>GEARS: 0.78 ($p<0.001$)<br>RACE: 0.74 ($p<0.001$) |

| | | | | | |
|---|---|---|---|---|---|
| Holst 2015a | 2 Experts<br>208 CWs | 5 | Experts: 50/26 hrs<br>CWs: 50/2hrs 50 mins/video | GEARS score<br>Experts:<br>PGY-2: 7.0 (1.0)<br>PGY-4: 11.0 (1.7)<br>PGY-5: 8.33 (2.1)<br>Attending surg #1: 10.3 (2.5)<br>Attending surg #2: 14.7 (0.6) | Cronbach's alpha: 0.91<br>SCC: 0.93 (excellent agreement) |
| Holst 2015b | 7 Experts<br>487 CWs | 12 | Experts: 12/14 days<br>CWs: 12/4 hrs 28 mins | GEARS<br>Data reported for each video | Cronbach's alpha: 0.93<br>SCC: 0.91 (excellent agreement) |
| Karani | 2 Experts<br>32 CWs | 282 | Experts: 564/19.5 days<br>CWs: 9024/60 minutes | GEARS score<br>Experts: 11.36 (3.78)<br>CWs: 10.8 (2.9)<br>GOALS score<br>Experts: 9.26 (3.82)<br>CWs: 12.4 (1.68)<br>OSATS score<br>Experts: 12.31 (3.82)<br>CWs: 16.73 (1.0)<br>Average of three scores<br>Experts: 10.97 (2.89)<br>CWs: 13.32 (1.38) | Cronbach's alpha<br>GEARS score: 0.88 (good)<br>GOALS score: 0.66 (fair)<br>OSATS score: 0.32 (poor)<br>Average of three scores: 0.80 (good) |
| Kowalewski | NR | 24 | Experts: 120/10 days<br>CWs: 1438/48 hrs | GOALS score Peg transfer<br>Experts: 13.07<br>CWs: 11.60<br>GOALS score Suturing<br>Experts: 14.52<br>CWs: 12.37 | Overall inter-rater reliability<br>Cronbach's alpha: 0.83<br>Peg transfer<br>Cronbach's alpha: 0.79<br>PCC: 0.95 (p<0.001)<br>Suturing<br>Cronbach's alpha: 0.92<br>PCC: 0.70 (p=0.01) |
| Polin | 3 Experts<br>448 CWs | 60 | Experts: NR<br>CWs: 2119/16 hours | R-OSATS score<br>NR | PCC<br>Tower transfer. 0.75 (p=0.005)<br>Roller coaster: 0.91 (p<0.001)<br>Big Dipper: 0.86 (p<0.001)<br>Train tracks: 0.76 (p=0.004)<br>Figure of 8: 0.87 (p<0.001) |
| Powers | 3 Experts<br>30 CWs | 14 | Experts: 14/13 days<br>CWs: 548/11 hrs 33 mins | GEARS score<br>Renal artery dissection score<br>NR | PCC<br>GEARS score<br>Video level: 0.82 (p<0.001)<br>Surgeon level: 0.84 (p<0.001)<br>Renal artery dissection: 0.83 (p<0.001) |
| Vernez | 6 Experts | NS | Experts: 150/22 days<br>CWs<br>Knot tie: 1606/3 hrs 4 mins<br>Peg transfer:749/3 hrs 3 mins<br>Suturing: 767/3 hrs 26 mins<br>Mentor lap: 8816/3hrs 27 mins | Open knot tie (OSATS)<br>Experts: 12.24 (2.29)<br>CWs: 16.38 (0.85)<br>Lap peg transfer (GOALS)<br>Experts: 8.83 (2.83)<br>CWs: 7.5 (2.01)<br>Robotic suture (GEARS)<br>Experts: 8.15 (2.80)<br>CWs: 15.04 (2.09) | Open knot tie (OSATS)<br>Cronbach's alpha: 0.623<br>PCC: 0.69<br>Lap peg transfer (GOLAS)<br>Cronbach's alpha: 0.916<br>PCC: 0.89<br>Robotic suture (GEARS)<br>Cronbach's alpha: 0.864<br>PCC: 0.79 |
| White | 3 Experts | 98 | Experts: NR<br>CWs<br>Pegboard: 1433/108 hrs 48 mins<br>Suturing: 1498/8 hrs 52 mins | GEARS score<br>NR | Pegboard<br>Cronbach's alpha: 0.84<br>CC: 0.79<br>Suture<br>Cronbach's alpha: 0.92<br>CC: 0.86 |

SD: Standard Deviation, CWs: Crowd-sourced Workers, NR: Not Reported, CI: Confidence Interval, PCC: Pearson's Correlation Coefficient, CVS: Critical View of Safety, SCC: Spearman's Correlation Coefficient, RACE: Robotic Anastomosis and Competency Evaluation, PGY: Post-graduate Gear, NS: Not Specified, CC: Unspecified Correlation Coefficient

two on a mix of approaches: open and laparoscopic [22], open, laparoscopic and robotic [28].

Eleven studies focused on urological surgery [16,17,24-30,33,34], four studies on general surgery [19-22], one on bariatric surgery [15] and one gynecology [32]. Three studies reported on a mix of specialties, including urology, general surgery and gynecology [23,31,35], while Chen [18] did not report any specific specialty.

Regarding the specific procedures evaluated, twelve studies evaluated surgical skills using simulators [17-19,23,26,28-32,34,35], one used a porcine model for bladder repair [27], while nine studies evaluated C-SATS on surgical procedures on patients during cholecystectomy [20,21], sleeve gastrectomy [15], partial nephrectomy [33], radical prostatectomy [16,24], radical cystectomy [25] and a mix of open (ventral and inguinal hernia repair, thyroidectomy) and laparoscopic (cholecystectomy, colectomy, inguinal hernia repair) procedures [22].

### Summary of Evidence – Validation of crowd-sourcing against expert opinion

Thirteen studies used C-SATS to determine if crowdsourcing assessment and surgeon assessment correlated and if crowd sourced workers were non-inferior to experts in rating surgical skill using videos [18-20,24-29,32-35]. Results from these studies are reported in Table 2.

Twelve of the aforementioned studies indicated that crowd-sourced workers' and experts' ratings using the C-SATS platform had an overall excellent or good correlation between them in GOALS, GEARS, and R-OSATS scores (validated scores to assess laparoscopic and robotic surgical skills), and different specialties, with either Pearson's (evaluating the linear relationship between two continuous variables) or Spearman's (evaluating the monotonic relationship) correlation coefficients ranging from 0.69 to 0.95 and Cronbach's alpha (measure of internal consistency) from 0.63 to 0.93 (see Table 2). Only one study evaluated the correlation between crowd-sourced workers and experts in rating the OSATS score (evaluating skill in performing surgical tasks in open procedures) of a suturing task using a simulator, showing poor correlation, with a Cronbach's alpha of 0.32 [28]. The same group [34] on the other hand previously found a fair correlation for the OSATS score of square knot tying simulation task between experts and crowd-sourced workers. Studies comparing crowd-sourced workers and experts showed that crowd-sourced workers assessed technical skill equivalent to experts more rapidly, which led to significantly more ratings, and reduced variability.

A single study comparing the operative technique of uretero-ileal anastomoses resulting in clinically significant uretero-ileal stricture (UIS) with contralateral anastomosis found that crowd-sourced assessment was not predictive of UIS (p=0.62) [25]. The authors examined whether surgeon-perceived risk of UIS or crowd-sourced assessment of robotic skill are associated with the development of UIS. De-identified videos were analyzed by five high-volume surgeons and crowd workers to determine GEARS score. Also, no association between the expert mode response and UIS (OR 0.42; 95% confidence interval [CI] 0.05–3.45; p=0.91) was identified.

Several studies provided further information on the performance of a crowd-sourced work force. Some studies also evaluated qualitative feed-back from raters. Chen et al.[18] were able to isolate meaningfully different ratings by using writing style cues in the justification for the examiner's grading. This was evidenced by a significant difference between "predicted-better" and "predicted-worse" sets, based on the presence of words such as "but" and related negation words, which were found to occur much more frequently in the better set of responses. Deal et al.[19] using theme and sub-theme analysis, showed that comments were consistent between experts and crowd-sourced workers, yielding similar feed-back to the learner in the final report.

One study [26] evaluated the difference in scores of a single urology resident performing an intracorporeal robotic simulation task without any warm-up (cold) and then after 10 minutes of practice with faculty-guided feed-back (warm). The crowd-sourced workers indicated a 14% improvement in performance between "cold" and "warm". Karani et al.[28] found that none of the laparoscopic, robotic or open simulation tasks performed by urology residents evaluated by the crowd correlated with other metrics, such as department match rank, interview score, USMLE Step 1 score, or JSPE-S score, or were predictive of match rank, not adding value to the applicant selection process. Polin et al.[32] also evaluated the minimum number of crowd-sourced workers scores of 5 robotic dry lab drills performed by gynecologists and general surgeons needed to maintain a high correlation with expert evaluation, determining that obtaining 15 crowd-sourced worker assessments per trainee is sufficient. Finally, White et al.[35] noted that the crowd disagreed with experts for the 90$^{th}$ percentile rocking peg-board performance, scoring nearly two points more critically, meaning the expert rewarded the top 90$^{th}$ percentile performances higher. They also demonstrated that using C-SATS is cost-effective; the cost to grade an individual performance for different robotic simulation tasks using crowd-sourced workers was approximately $16.50, while using three experts the cost ranged from $54.00 to $108.00.

### Summary of Evidence – Use of C-SATS to assess performances

Two studies used C-SATS to evaluate surgeons' skills [16,30]. Almarzouq et al.,[16] assessed urology residents transferability of basic robotic skills from the da Vinci Surgical Skills Simulator to the OR using GEARS evaluations of 50 C-SATS crowd-sourced reviewers. The authors randomized the participants into 2 groups: group A, which was required to practice three sessions on the simulator, while group B could practice the same exercises until proficiency was reached. Both groups were recorded during simulation and then when performing robotic-assisted radical prostatectomy in the OR. Total GEARS scores for "ring and rail 2" and "suture sponge" correlated with the total GEARS scores during

urethro-vesical anastomosis (rho=0.86, p=0.007; and rho=0.90, p=0.002, respectively), as well as GEARS sub-scores for "energy and dissection", "ring and rail 2", and "dots and needles" exercises correlated with bladder mobilization. The authors stated that the results showed basic robotic skills could be transferred from the simulator to the OR.

Lee et al.[30] initiated a national skills assessment study focusing on laparoscopic skills. All performances of four standardized tasks from the American Urological Association (AUA) Basic Laparoscopic Urological Surgery (BLUS) curriculum were video recorded and assessed using three methods, including time and error-based scoring, expert and C-SATS global rating scores, both using GOALS scores. The authors demonstrated that the C-SATS method of assessment correlated with the traditional methods of time and error-based scoring and the global rating scale ($p<0.01$).

### Summary of Evidence – Use of C-SATS as a validation tool

Six articles used the C-SATS platform and crowd-sourced workers as a validation tool [15,17,21-23,31].

Addison et al. [15] demonstrated that operative time, a widely used metrics to evaluate surgical skill, was not the ideal parameter to assess robotic bariatric surgical skills. They found no correlation between operative time and overall and step-specific GEARS scores provided by crowd-sourced workers using C-SATS with the exception of gastric transection, which showed a weak correlation. The authors speculated that operative time and GEARS score reflect different dimensions of surgical skill.

Bendre et al. [17] described the development and face validation of a robotic pyeloplasty simulation using a 3D-printed silicone-based model of ureteropelvic junction obstruction for surgical training. GEARS scoring was used to objectively assess performance and learning outcomes. The authors reported that while using C-SATS to track performance, there was a trend toward an increase in overall GEARS score, although it was not statistically significant. However, when the GEARS score was divided by category, there was a mean improvement in each category, with depth perception reaching statistical significance.

Deal et al. [21] used C-SATS to assess the relationship between operative quality as determined by the correct visualization of the critical view of safety (CVS) during laparoscopic cholecystectomy and viewing characteristics of online laparoscopic videos. It was reported that only one video of 160 (0.06%) achieved a passing CVS score of ≥5. Average CVS ratings were highly correlated with the probability of assigning a pass or fail rating for completing the CVS ($r^2=0.95$; $p < 0.001$), as well as with GOALS scores ($r^2=0.79$; $p<0.001$). YouTube videos (n = 139) with higher views, likes, or subscribers did not correlate with better quality. The average CVS and GOALS scores were no different for videos with >20,000 views (22%) compared with those with <20,000 (78%).

The same authors [22] assessed the correlation of crowd OSATS or GOALS rating with the system for improving procedural learning (SIMPL) Zwisch and Performances scales. SIMPL is a smartphone based mobile application for the evaluation of operative performance and autonomy capturing three metrics, an autonomy metric (Zwisch scale), a difficulty scale and a performance metric [36]. Correlations between crowd-sourced ratings using GOALS and OATS and SIMPL global operative performance ratings tools were weak (GOALS/Zwisch r=-0.40; OSATS/Zwish r=0.11; GOALS/performance r=-0.06; OSATS/performance r=0.22). On the other hand, attending surgeons' GOALS and OSATS ratings did correlate with SIMPL metrics (GOALS/Zwisch r=0.77; OSATS/Zwish r=0.65; GOALS/performance 0.93; OSATS/performance r=0.59) suggesting that crowd sourcing may be more suitable for technical assessment, while attending assessment may be needed for evaluation of global performance.

Dubin et al. [23] used two robotic simulators (dV-Trainer and dVSS) metrics and C-SATS GEARS scoring of the task "ring and rail 1" to determine if there is a difference between robotic virtual reality simulator performance assessment and validated human reviewers. Results were contrasting, with a strong correlation between the GEARS score and some of the simulator metrics (time to complete versus efficiency, time to complete versus total score economy of motion versus depth of perception, and overall score versus total score; rho≥0.70, $p<0.0001$), but not others (bimanual dexterity versus economy of motion, efficiency versus master workspace range, bimanual dexterity versus workspace range, and robotic control versus instrument collision showed only a weak correlation; rho≥30, p=NS).

Martin et al. [31] evaluated whether robotic surgery novices would show improved technical skill performances after completing Fundamentals of Robotic Surgery (FRS) proficiency-based psychomotor skills training using the RobotiX Mentor VR simulator platform. The results were then compared to determine its effectiveness versus other previously published FRS training platforms. The FRS curriculum is a proficiency-based progression curriculum consisting of didactic modules and a simulation-based skill curriculum for the acquisition of basic robotic skills [37,38]. Two experts provided ratings for videos managed by the C-SATS platform. Participants demonstrated improved performance across all GEARS domains as well as for time and errors as measured by psychometric checklist. Improvements in novices' skill after FRS training on the RobotiX Mentor using GEARS scores was not inferior to improvement reported after FRS training on previously published platforms.

## Discussion

Technical skills of surgeons may be fundamental in particularly demanding surgical steps, such as bleeding control, tissue handling, and lengthy operations, including a positive effect on complications such as surgical site infection and venous thromboembolism [39-44]. Such technical skills have been positively correlated with outcomes showing the importance of technique in complex procedures [45].

Evaluation of technical skills or the validity of a tool, such as a physical or virtual simulator used to train surgeons, by experts can be time consuming and expensive. This is due to the need for the expert to either be present during the procedure to evaluate or to spend time examining a full video following the procedure. The C-SATS platform has a number of features which may be particularly useful to execute a fast, reliable, blinded, and cost-effective evaluation of technical skills. It allows a seamless upload of surgical videos to a secure cloud-based platform, where the videos are anonymized so that patient's privacy is preserved, and a blinded evaluation of the surgeon can be accomplished. Moreover, the artificial intelligence algorithm allows the system to efficiently extract short clips of the critical steps of the procedure to submit for feedback. This resource can then be easily used to recruit both experts and crowd-sourced workers, according to specific needs and goals, such as rating trainees, giving feed-back for a surgeon's personal development or testing new solutions or research hypotheses.

Studies included in this review [18-20,24-29,32-35] demonstrated a significant positive correlation between experts and crowd-sourced workers across different specialties, suggesting that ratings of technical skill provided by these two groups are similar. Agreement between expert and crowd assessment scores were found when surgical tasks, surgical approaches (open, laparoscopic and robotic), and varying levels of surgical skills were evaluated (Table 3). These findings are in agreement with a recent systematic review of crowd-sourcing platforms which found moderate to very strong correlation between crowd-sourced workers and experts [46].

**Table 3.** C-SATS use in different specialties and surgical approaches

| Specialty/Approach | Correlation between Experts & CWs | Assessment of Skills/ Training | Validation Tool |
|---|---|---|---|
| Urology | 7P, 1N | 2§ | 1 |
| General surgery | 2P | - | 2 |
| Bariatric surgery | - | - | 1 |
| Gynecology | 1P | - | - |
| Misc (urology, general, gynecology) | 1P | - | 2 |
| Laparoscopy | 4P | 1§ | 1 |
| Robotically-assisted | 7P, 1N | 1§ | 4 |
| Misc (open, robotic, laparoscopic) | 1P | - | 1 |

P = Significant correlation between experts and crowd-sourced workers (CWs); N = no significant correlation between experts and CWs; § = positive outcomes.

Outcomes of these correlation studies also demonstrated that the use of C-SATS and involving a crowd allows a faster (near-immediate, with significantly higher number of evaluations in a shorter time compared to experts) and more cost-effective way of blindly evaluating surgical skills. Another important factor when using multiple examiners (such as crowd raters) is inter-rater reliability. While in some studies inter-rater reliability among experts has been shown to be limited [19,33], the importance of this variable is lower when using crowd-sourced workers, since their involvement permits having a large enough sample size to accurately measure relative technical performance. The high number of ratings which are yielded by a crowd-based population of raters allows a higher confidence in the overall average rating due to narrow confidence intervals, even if the variability in crowd ratings is greater than that of experts. Moreover, it is possible, using artificial intelligence and Bayesian-like approaches [47] to identify common key terms in qualitative feed-back to determine ratings and providing a more detailed report even using crowd workers [18,19,32].

These studies have listed different limitations. Some studies submitted only a limited number of videos or performances to raters [18,26,29] or left the crowd-workers free to evaluate any number of videos [32], while most of them were performed in a dry-lab setting or in controlled wet-lab environment, limiting the applicability of the results to actual OR cases [26,27,29,32,35]. Other studies report a limited sample size and the use of novice learners only [19] or a limited use of experts [33]. Nevertheless, this body of evidence, taken as a whole, provide convincing evidence that the use of videos and technical skill tools used by crowd-sourced workers yield positive correlation with evaluation from surgical experts.

Expert rating should still be considered the gold standard for assessment of video-based surgical skill, since there are factors, such as indications for surgery, variations in anatomy and patient factors that preclude the possibility of proper evaluation by lay persons. Moreover, some surgeons might not want to be evaluated by a reviewer lacking a surgical background. Considering also that crowds demonstrated a high concordance with experts in identifying the extremes of the spectrum of skills, it is likely that C-SATS might be valuable in identifying trainees with deficiencies and allowing experts to target training resources to those deficiencies rather than to administer the same curriculum to all trainees. Another approach would be to gather a world-wide "crowd" of experts, who can quickly and efficiently evaluate videos using their expertise, also providing high-level feed-back to learners. Moreover, the possibility of using C-SATS to track personal development of surgical skill may be useful for surgeons who wants to improve their surgical dexterity.

Several studies, built upon the demonstrated correlation between experts and crowd-sourced workers, used C-SATS in the evaluation of technical skills. Almarzouq et al.[16] assessed the transferability of basic robotic skills at the simulator to the OR and found that both total and individual domains GEARS scores at the simulation tasks ring and rail 2 positively correlated with those found in the OR. On the other hand, Lee, et al.[30] used C-SATS to evaluate the GOALS scores of four American Urological Association Basic Laparoscopic Urological Surgery curriculum tasks among urology trainees. The authors stated that C-SATS is a reliable and valid tool for the assessment of basic laparoscopic skills, as opposed to time- and resource-consuming traditional methods of technical skills assessment, including global rating scores by experts and

time plus error-based scoring methods. It was determined that C-SATS provided a timely blinded valuation, and did not require further expert assessment. These characteristics may be valuable in an ongoing training (or improvement) program, where evaluations are frequent and iterative.

C-SATS has been used to determine the effectiveness of simulators, both physical [17] and virtual [23,31]. Although two of these studies demonstrated that several simulator metrics were well-matched, and in some instances significantly matched, with scores assigned by crowd-workers, one was not (e.g. GEARS robotic control to simulator instrument collision) [23]. This lack of correlation for virtual robotic simulator metrics may be due to differences in how the tasks are evaluated by the software and by human evaluators. Bendre et al. [17] while only finding a trend toward an increase in overall GEARS scores after the use of a physical simulator in performing robotic pyeloplasty, also found that there was a mean improvement in each GEARS category and depth perception showing a significant improvement. This allowed the author to demonstrate the face-validity of the simulator.

Deal at al. [22] used C-SATS to validate smart-phone based mobile application (SIMPL) metrics using both expert and crowd-sourcing OSATS and GOALS scores. They demonstrated that both OSATS and GOALS scores correlated well with SIMPL metrics when using experts. Crowd workers performed significantly worse than experts when assessing the correlation between global performance (SIMPL metrics) to technical performance (OSATS and GOALS scores). This may be due to the fact that, in this particular study, crowd-sourced workers may need additional training or that experts may have been biased, over-correlating technical performance with global performance. Nevertheless, the authors acknowledge that the study demonstrated that technical assessment feedback from crowd-sourced workers using C-SATS was reliable and timely. This is in contrast with the feedback provided by the expert which had a response rate of only 81%, and only after repeated invitations to complete the task.

C-SATS has also been shown to be useful to evaluate the quality of surgical videos posted on online resources [21]. The authors showed a low frequency of critical view of safety (only 1 over 139 videos) and average GOALS score (all below average) during laparoscopic cholecystectomy in frequently used online surgical videos. Using crowd-sourced workers permitted the quick evaluation of a high number of videos, allowing the determination that trainees should be cautious when using public domain websites for surgical learning.

## Conclusions

Technical skills are primarily acquired during medical school, residency and the first years into practice. Most of the training is based upon mentorship and feed-back from expert surgeons, often in a qualitative, uncontrolled, and possibly biased fashion. Another challenge is the progression of surgeons in building up their technical skills and optimizing their performances, since generally there is no formal program, besides continuous medical education, allowing for ongoing evaluation.

The acquisition and maintenance of technical skill is one of the essential professional attributes that a surgeon should demonstrate to ensure optimal patient outcomes. Surgeons require an environment of lifetime learning, often hard to achieve in the environ of a busy practice. Mentorship, expert feed-back and tool validation for physical or virtual simulation can be time and resource consuming.

The C-SATS platform efficiently provides crowd-sourced evaluations that correlate favorably with expert review across a range of surgical specialties (general, urology, gynecology) and approaches (robotic, laparoscopic). Use of crowdsourcing has uniformly yielded accurate evaluations of surgical skills in a markedly shorter time period than for expert reviews. C-SATS appears to be a cost-effective complement or alternative to traditional models to acquire surgical proficiency. It may also represent a first step allowing the objective link between surgical skill improvements and reduction of patient complications. Nevertheless, future studies are needed to determine whether the use of C-SATS will lead to this goal.

## References


1. John G Meara, Andrew J M Leather, Lars Hagander, et al. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. The lancet. 2015; 386: 569-624.
2. Naghavi M, Abajobir AA, Abbafati C, et al. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980-2016: a systematic analysis for the Global Burden of Disease Study 2016. The Lancet. 2017; 390: 1151-1210.
3. Oliver Anderson, Rachel Davis, George B Hanna, et al. Surgical adverse events: a systematic review. The American Journal of Surgery. 2013; 206: 253-262.
4. Florence E Turrentine, Worthington G Schenk, Timothy L McMurry, et al. Surgical errors and the relationships of disease, risks, and adverse events. The American Journal of Surgery. 2020; 220: 1572-1578.
5. Britt LD, Ajit K. Sachdeva, Gerald B. Healy, et al. Resident duty hours in surgery for ensuring patient safety, providing optimum resident education and training, and promoting resident well-being: a response from the American College of Surgeons to the Report of the Institute of Medicine,"Resident Duty Hours: Enhancing Sleep, Supervision, and Safety". Surgery. 2009; 146: 398-409.
6. Temple J. Time for training: a review of the impact of the European Working Time Directive on the quality of training. Medical Education England. 2010. 33.
7. Sachdeva AK. The new paradigm of continuing education in surgery. Archives of Surgery. 2005; 140: 264-269.
8. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Journal of British Surgery. 1997; 84: 273-278.

9. Alvin C Goh, David W Goldfarb, James C Sander, et al. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. The Journal of urology. 2012; 187: 247-252.
10. Carol E Reiley, Henry C Lin, David D Yuh, et al., Review of methods for objective surgical skill evaluation. Surgical Endoscopy. 2011; 25: 356-366.
11. Jacob Rosen, Jeffrey D Brown, Lily Chang, et al. Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete Markov model. IEEE Transactions on Biomedical engineering. 2006; 53: 399-413.
12. Julian J H Leong, Marios Nicolaou, Louis Atallah, et al. HMM assessment of quality of movement trajectory in laparoscopic surgery. Computer Aided Surgery. 2007; 12: 335-346.
13. Paolacci G, Chandler J Ipeirotis PG. Running experiments on amazon mechanical Turk. Judgment and Decision Making. 2010; 5: 411-419.
14. Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021; 372: n71.
15. Poppy Addison, Andrew Yoo, Jacqueline Duarte-Ramos, et al. Correlation between operative time and crowd-sourced skills assessment for robotic bariatric surgery. Surgical endoscopy. 2021; 35: 5303-5309.
16. Ahmad Almarzouq, Jason Hu, Yasser A Noureldin, et al. Are basic robotic surgical skills transferable from the simulator to the operating room? A randomized, prospective, educational study. Canadian Urological Association Journal. 2020; 14: 416-422.
17. Hersh H Bendre, Archana Rajender, Philip V Barbosa, et al. Robotic dismembered pyeloplasty surgical simulation using a 3D-printed silicone-based model: development, face validation and crowdsourced learning outcomes assessment. Journal of robotic surgery. 2020; 14: 897-902.
18. Carolyn Chen, Lee White, Timothy Kowalewski, et al. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. Journal of surgical research. 2014; 187: 65-71.
19. Shanley B Deal, Thomas S Lendvay, Mohamad I Haque, et al. Crowd-sourced assessment of technical skills: an opportunity for improvement in the assessment of laparoscopic surgical skills. The American Journal of Surgery. 2016; 211: 398-404.
20. Shanley B Deal, Dimitrios Stefanidis, Dana Telem, et al. Evaluation of crowd-sourced assessment of the critical view of safety in laparoscopic cholecystectomy. Surgical endoscopy. 2017; 31: 5094-5100.
21. Shanley B Deal, Adnan A Alseidi. Concerns of quality and safety in public domain surgical education videos: an assessment of the critical view of safety in frequently used laparoscopic cholecystectomy videos. Journal of the American College of Surgeons. 2017; 225: 725-730.
22. Shanley B Deal, Rebecca E Scully, Gregory Wnuk, et al. Crowd-Sourced and Attending Assessment of General Surgery Resident Operative Performance Using Global Ratings Scales. Journal of Surgical Education. 2020; 77: e214-e219.
23. Ariel K Dubin, Roger Smith, Danielle Julian, et al. A comparison of robotic simulation performance on basic virtual reality skills: simulator subjective versus objective assessment tools. Journal of minimally invasive gynecology. 2017; 24: 1184-1189.
24. Khurshid R. Ghani, Bryan Comstock, David C. Miller, et al. PNFBA-02 technical skill assessment of surgeons performing robot-assisted radical prostatectomy: relationship between crowdsourced review and patient outcomes. The Journal of Urology. 2017; 197: e609-e609.
25. Mitchell Goldenberg, Michael Ordon, John R D'A Honey, et al. Objective Assessment and Standard Setting for Basic Flexible Ureterorenoscopy Skills Among Urology Trainees Using Simulation-Based Methods. Journal of endourology. 2020; 34: 495-501.
26. Daniel Holst, Timothy M Kowalewski, Lee W White, et al. Crowd-sourced assessment of technical skills: an adjunct to urology resident surgical simulation training. Journal of endourology. 2015; 29: 604-609.
27. Daniel Holst, Timothy M Kowalewski, Lee W White, et al. Crowd-sourced assessment of technical skills: differentiating animate surgical skill through the wisdom of crowds. Journal of endourology. 2015; 29: 1183-1188.
28. Rajiv Karani, Shlomi Tapiero, Francis A Jefferson, et al. Crowd-Sourced Assessment of Surgical Skills of Urology Resident Applicants: Four-Year Experience. Journal of Surgical Education. 2021; 78: 2030-2037.
29. Timothy M Kowalewski, Bryan Comstock, Robert Sweet, et al. Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. The Journal of urology. 2016; 195: 1859-1865.
30. Jason Y Lee, Sero Andonian, Kenneth T Pace, et al. Basic laparoscopic skills assessment study: validation and standard setting among Canadian urology trainees. The Journal of urology. 2017; 197: 1539-1544.
31. John Rhodes Martin, Dimitrios Stefanidis, Ryan P Dorin, et al. Demonstrating the effectiveness of the fundamentals of robotic surgery (FRS) curriculum on the RobotiX Mentor Virtual Reality Simulation Platform. Journal of robotic surgery. 2021; 15: 187-193.
32. Michael R Polin, Nazema Y Siddiqui, Bryan A Comstock, et al. Crowdsourcing: a valid alternative to expert evaluation of robotic surgery skills. American journal of obstetrics and gynecology. 2016; 215: 644.e1-644.e7.
33. Mary K Powers, Aaron Boonjindasup, Michael Pinsky, 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. Journal of endourology. 2016; 30: 447-452.

34. Simone L Vernez, Victor Huynh, Kathryn Osann. et al. C-SATS: assessing surgical skills among urology residency applicants. Journal of endourology. 2017; 31: S95-S100.
35. Lee W White, Timothy M Kowalewski, Rodney Lee Dockter. et al. Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. Journal of endourology. 2015; 29: 1295-1301.
36. Brian C George, Ezra N Teitelbaum, Shari L Meyerson, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. Journal of surgical education. 2014; 71: e90-e96.
37. Boris Zevin, Jeffrey S Levy, Richard M Satava, et al. A consensus-based framework for design, validation, and implementation of simulation-based training curricula in surgery. Journal of the American College of Surgeons. 2012; 215: 580-586.e3.
38. Richard M Satava, Dimitrios Stefanidis, Jeffrey S Levy, et al. Proving the effectiveness of the fundamentals of robotic surgery (FRS) skills curriculum: a single-blinded, multispecialty, multi-institutional randomized control trial. Annals of surgery. 2020; 272: 384-392.
39. Darrell A Campbell Jr, William G Henderson, Michael J Englesbe, et al. Surgical site infection prevention: The importance of operative duration and blood transfusion—results of the first American College of Surgeons–National Surgical Quality Improvement Program Best Practices Initiative. Journal of the American College of Surgeons. 2008; 207: 810-820.
40. Chan MM, Hamza N, Ammori BJ. Duration of surgery independently influences risk of venous thromboembolism after laparoscopic bariatric surgery. Surgery for Obesity and Related Diseases. 2013; 9: 88-93.
41. Stefan Kessler, Stefan Kinkel, Wolfram Käfer, et al. Influence of operation duration on perioperative morbidity in revision total hip arthroplasty. Acta orthopaedica belgica. 2003; 69: 328-333.
42. Leong G, Wilson J, Charlett A. Duration of operation as a risk factor for surgical site infection: comparison of English and US data. Journal of Hospital Infection. 2006; 63: 255-262.
43. Levi D Procter, Daniel L Davenport, Andrew C Bernard, et al. General surgical operative duration is associated with increased risk-adjusted infectious complication rates and length of hospital stay. Journal of the American College of Surgeons. 2010; 210: 60-65.e2.
44. Tze-Woei Tan, Jeffrey A Kalish, Naomi M Hamburg, et al. Shorter duration of femoral-popliteal bypass is associated with decreased surgical site infection and shorter hospital length of stay. Journal of the American College of Surgeons. 2012; 215: 512-518.
45. John D. Birkmeyer, Jonathan F. Finks, Amanda O'Reilly, et al. Surgical skill and complication rates after bariatric surgery. New England Journal of Medicine. 2013; 369: 1434-1442.
46. Rikke G.Olsena, Malthe F.Genétb, LarsKonge, et al. Crowdsourced assessment of surgical skills: A systematic review. The American Journal of Surgery. 2022.
47. Dimitri P. Bertsekas, John N. Tsitsiklis. Introduction to probability. 2000: Athena Scientinis.