ORIGINAL REPORTS

Crowdsourced Assessment of Surgical Skill Proficiency in Cataract Surgery

Grace L. Paley, MD, PhD,* Rebecca Grove, BA,† Tejas C. Sekhar, BA,* Jack Pruett*, Michael V. Stock, MD,* Tony N. Pira, MD,‡ Steven M. Shields, MD,* Evan L. Waxman, MD, PhD,§ Bradley S. Wilson, MA,* Mae O. Gordon, PhD,* and Susan M. Culican, MD, PhD*†,¶,1

*Department of Ophthalmology and Visual Sciences, Washington University School of Medicine, Saint Louis, Missouri; †Graduate Medical Education, University of Minnesota, Minneapolis, Minnesota; ‡Department of Ophthalmology, Boston University School of Medicine, Boston, Massachusetts; §Department of Ophthalmology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania; and ¶Department of Ophthalmology and Visual Neurosciences, University of Minnesota, Minneapolis, Minnesota

OBJECTIVE: To test whether crowdsourced lay raters can accurately assess cataract surgical skills.

DESIGN: Two-armed study: independent cross-sectional and longitudinal cohorts.

SETTING: Washington University Department of Ophthalmology.

PARTICIPANTS AND METHODS: Sixteen cataract surgeons with varying experience levels submitted cataract surgery videos to be graded by 5 experts and 300+ crowdworkers masked to surgeon experience. *Cross-sectional study*: 50 videos from surgeons ranging from first-year resident to attending physician, pooled by years of training. *Longitudinal study*: 28 videos obtained at regular intervals as residents progressed through 180 cases. Surgical skill was graded using the modified Objective Structured Assessment of Technical Skill (mOSATS). Main outcome measures were overall technical performance, reliability indices, and correlation between expert and crowd mean scores.

RESULTS: Experts demonstrated high interrater reliability and accurately predicted training level, establishing construct validity for the modified OSATS. Crowd scores were correlated with ($r = 0.865$, $p < 0.0001$) but consistently higher than expert scores for first, second, and third-year residents ($p < 0.0001$, paired t-test). Longer surgery duration negatively correlated with training level ($r = -0.855$, $p < 0.0001$) and expert score ($r = -0.927$, $p < 0.0001$). The longitudinal dataset reproduced cross-sectional study findings for crowd and expert comparisons. A regression equation transforming crowd score plus video length into expert score was derived from the cross-sectional dataset ($r^2 = 0.92$) and demonstrated excellent predictive modeling when applied to the independent longitudinal dataset ($r^2 = 0.80$). A group of student raters who had edited the cataract videos also graded them, producing scores that more closely approximated experts than the crowd.

CONCLUSIONS: Crowdsourced rankings correlated with expert scores, but were not equivalent; crowd scores overestimated technical competency, especially for novice surgeons. A novel approach of adjusting crowd scores with surgery duration generated a more accurate predictive model for surgical skill. More studies are needed before crowdsourcing can be reliably used for assessing surgical proficiency. (J Surg Ed 000:1−12.

ABBREVIATIONS: C-SATS, Crowd-Sourced Assessment of Technical Skills; ICC, intraclass correlation coefficient; mOSATS, Modified Objective Structured Assessment of

DOI of original article: http://dx.doi.org/10.1016/j.jsurg.2021.03.001
Professor, Dept of Ophthalmology and Visual Neurosciences, University of Minnesota.
MMC 293, 420 Delaware Street SE, Minneapolis, MN 55455, USA, Phone: 612-625-7905 culican@umn.edu
Meeting Presentation: Presented in part at Educating the Educators − Association of University Professors of Ophthalmology (AUPO) in January 2020 and The Association for Research in Vision and Ophthalmology (ARVO) in April 2019.
Data sharing: Grading data from experts and the crowd will be freely available. Videos will not be available as this data was considered confidential to the operative surgeons. Data sharing beyond this IRB approved study was not included in the consent.
Correspondence: Inquiries to: Susan M. Culican, MD, PhD, Department of Ophthalmology and Visual Neurosciences, University of Minnesota, MMC 293, 420 Delaware Street SE, Minneapolis, MN 55455, USA; e-mail: culican@umn.edu

[1] Present/Permanent address: Susan M. Culican, MD, PhD, Designated Institutional Official
Associate Dean of Graduate Medical Education
Graduate Medical Education, University of Minnesota

https://doi.org/10.1016/j.jsurg.2021.02.004

Technical Skill; OKAP, Ophthalmic Knowledge Assessment Program; PGY, postgraduate year

KEY WORDS: Crowdsourcing, surgical assessment, surgical competence, cataract surgery, phacoemulsification

COMPETENCIES: Patient Care, Medical Knowledge, Practice-Based Learning and Improvement, Systems-Based Practice

INTRODUCTION

Ophthalmology residency training programs carry an enormous responsibility to the profession and to the public when they certify that graduates are competent to perform unsupervised surgical procedures safely. While standardized tests of medical knowledge exist throughout residency and for board certification, there are no analogous technical skill assessments that are broadly adopted. Currently, judgments of surgical competence are made locally after reviewing subjective evaluations of residents by supervising faculty,[1] and these skills assessments vary by training program and are not universally defined. Thus, the current system of assessment lacks objective benchmarks to gauge surgical proficiency and to ensure that all resident physicians across training programs are equally competent to practice autonomously by graduation.[2] This disconnect, between certifying resident competency and the absence of standardized metrics to describe competency, puts programs at risk of misclassifying trainees as competent for independent practice when they may benefit from additional supervised practice and learning.[3] There are several validated grading scales to measure surgical skill[4-6] but these are time- and resource-intensive, subject to in-person rater biases, and largely inefficient for providing timely feedback to trainees.[7-9] Standardized protocols have been proposed with small but promising validation studies utilizing wet lab and surgical simulator technologies,[10-12] but have not been widely implemented for routine use. There is a crucial need to identify an effective surgical assessment tool that is minimally burdensome both in time and cost, that reduces evaluator bias, and that is scalable for use across different programs without requiring local expertise or training.

One idea gaining popularity is to outsource the procedural assessment to crowdsourced lay raters. Crowdsourcing, which employs large numbers of lay people via the internet to perform a task, represents a novel approach to standardize skill evaluations while minimizing subjective observer biases. While any individual crowd worker score may be inaccurate given their lack of expertise, crowdsourcing averages the scores of many workers to achieve "regression to the mean" of what is ideally an accurate judgment, at a much faster and cheaper scale than obtaining expert reviews. Several other surgical subspecialties including general surgery,[7,8] urology,[13] otolaryngology,[14] and gynecology[15] have published reports promoting crowdsourcing as comparable to expert feedback on technical skill assessments. However, most of these studies show correlation rather than equivalence of crowd and expert ratings and were almost entirely based on simulated surgery or wet lab scenarios. To date, the few study exceptions which examined crowdsourced assessments of actual surgery were limited by poor interrater reliability and reported mixed results regarding whether the crowd could distinguish novices from more experienced surgeons.[16-18]

The purpose of this investigation was to determine whether crowdsourced lay raters can accurately assess ophthalmic surgical skill as compared to experts, utilizing videos of real cataract surgery rather than simulations or wet lab setups. If feasible and accurate, this technology could be used to plot out trainee progression to proficiency and to identify for whom and when educational intervention may be beneficial. Also, as the crowdsourcing approach is not faculty- or institution-specific, crowd assessments of surgery could be used to collect cohort data across training programs, and perhaps even to establish standardized metrics as skill criteria for graduation or board certification.

MATERIALS AND METHODS

Study Design

Single institution, observational prospective cohort study with two arms: cross-sectional and longitudinal.

Setting and Participants

Ophthalmology residents, fellows, and faculty at the home institution were invited to submit cataract phacoemulsification videos for de-identified assessment. Institutional Review Board (IRB) approval and consents were obtained for the video collection and analysis (Washington University IRB ID# 201704153).

Video Collection

The study design included two study arms (Fig S1A). The first arm was a cross-sectional study consisting of 50 cataract videos collected from 15 surgeons at various levels of experience (postgraduate year (PGY)-2, 3, 4, 5 or attending) during the final month of the academic year. Arm 2 was a separate, longitudinal analysis of cataract videos for a cohort of 5 senior residents. Because residents rotate on more or less surgically busy rotations at different times during the year, videos were collected by

case number (every 30 cases: 30, 60, 90, 120, 150, 180) rather than by date. All surgical videos recorded actual surgery cases on live patients where a single surgeon was operating for the duration of interest.

Video Editing

Surgery videos were anonymized regarding patient and surgeon identities, and subsequently edited to include only the phacoemulsification segment, wherein the eye's natural lens is emulsified with an ultrasonic handpiece and aspirated from the eye. Each phaco segment began with entry of the phaco ultrasound probe tip into the eye, and ended when the entire lens was removed or after 10 minutes, whichever came first. As the phacoemulsification step ranged in time length across varying skill levels, longer video segments were truncated to the first 10 minutes to enable uploading to the Amazon Mechanical Turk crowdworker platform (Amazon.com Inc., Seattle, WA). Video editing was performed using iMovie software (Apple Inc., Cupertino, CA) and Windows Movie Editor (Microsoft Inc., Seattle, WA). Video brightness was adjusted to improve visualization of the surgical procedure. All edited videos were approved by the research coordinator as a quality control measure before being submitted for assessment.

Expert Raters

Five attending surgeons were recruited as expert raters for the study. Surgeon raters evaluated all videos independently in random order and did not receive monetary compensation.

Lay Raters

We contracted with the Crowd-Sourced Assessment of Technical Skills company (C-SATS, C-SATS, Inc., Seattle, WA) to obtain lay rater assessments of surgical videos. Crowdsourced lay raters were recruited by C-SATS from Amazon Mechanical Turk, a third-party marketplace that engages and pays lay workers for internet-based tasks. The C-SATS platform included a brief orientation video and one comprehension test question to screen lay raters. The animated orientation video was created by C-SATS, Inc. to familiarize the reviewers with the phacoemulsification step of cataract surgery and the scoring rubric. A correct response by the rater to the comprehension question was required for inclusion in the data analyses. Individual crowdworkers were allowed to rate multiple videos but could rate each video only once.

Outcome Measures

The primary outcome measure for both study arms was overall technical performance. There are several published valid and reliable surgical competency assessments for cataract surgery,[4-6] however many of them contain task-specific domains and ophthalmic technical language that could pose a barrier to crowdsourced lay raters. Thus, the Objective Structured Assessment of Technical Skill (OSATS)[19], the most commonly used technical skills assessment tool in surgery,[20] was modified with the goal being a grading rubric with simplified language easily understandable to a layperson without surgical expertise. The modified OSATS tool (mOSATS) preserves 4 elements of assessment from the OSATS (economy of movement, respect for tissue, flow of operation, instrument handling) and includes a fifth element adapted from the Global Rating Assessment of Skills in Intraocular Surgery[4] assessment tool (microscope centration) (Fig S1B). These domains overlap with content from multiple validated cataract skill assessments,[5,6] and discussion with our expert raters and other local cataract surgeons established face validity for the mOSATS. Each of the 5 categories was graded on a 5-point Likert scale with defined narrative anchors at the low, middle, and high ranges of the scale. To measure overall technical skill performance, a "mean sum score" was derived for each video as the averaged summative score across the 5 categories for both the experts and crowdworkers, with a minimum and maximum score of 5 and 25, respectively.

Statistical Analysis

Data analyses were computed with Statistical Analysis Software V9.4 (SAS Institute Inc., Cary, NC). Interrater reliability was determined by intraclass correlation coefficient (ICC), specifically the Shrout-Fleiss reliability ICC which is a measure of how well groups agree as opposed to how well they correlate (two-way random average measures). ICCs were calculated for single-score or mean k scores for single versus averaged data points, respectively. Mean crowd and expert scores were compared via Pearson correlation (Pearson's r) and matched-pair t-tests. To adjust for score clustering by crowd raters, crowd mean scores were generated with a linear mixed-effects model using an interaction term for surgical experience of trainee and rater type (lay or expert). A repeated measures model was used in the longitudinal study. Consistency was tested using an interaction term for experience as measured by case number from baseline (30, 60, 90, 120, 150, 180) and by rater type (lay or expert). A stepwise regression analysis was performed to explore the strength of crowd score and surgery length (phacoemulsification segment) variables as individual and combined predictors of expert score.

RESULTS

Cross-sectional Study

Fifty phacoemulsification video segments supplied by 15 unique physicians with varying surgical experience were graded by 5 blinded expert surgeons, yielding 250 evaluations. The expert sum scores demonstrated high interrater reliability measured by the intraclass correlation coefficient (ICC) (Fig 1A). The ICC for sum scores from individual experts was 0.891 and the ICC for the mean of expert sum scores was 0.976, suggesting good and excellent reliability,[21] respectively. This finding suggests that any individual expert performed very well but not quite as well as the average of the group. The group expert mean sum score predicted level of surgeon training (Pearson's $r = 0.860$, $p < 0.0001$) establishing construct validity for the mOSATS (Fig 1B). The expert group took 21 days to complete the 50 video evaluations.

A total of 2507 evaluations of the same 50 videos were furnished by 347 distinct, individual crowdworkers. Total active data acquisition time was 7.0 hours. Of these evaluations, 90.3% (n = 2264) were used for analysis after excluding evaluations that failed the screening comprehension criterion. The median number of crowd evaluations per video was 45 (range 41-49). The crowd scores demonstrated poor individual interrater reliability[21] with an ICC for individual crowd workers of 0.015, as compared with 0.891 for individual experts (Fig 1A). Nonetheless, the crowd performed much better as a group than as individuals as seen from the crowd mean ICC of 0.823. The mean crowd sum score had a moderately positive correlation with level of surgeon training ($r = 0.729$, $p < 0.0001$) which suggests that lay raters can reasonably assess surgical skill (Fig 1C), but not as reliably as the experts ($r = 0.860$ versus $r = 0.729$).

The averaged mean sum score across all 50 videos was higher for the crowd as compared with the experts, and the lowest average score given was 16.5 by the crowd versus 5.2 by the experts (Fig 1B-C), indicating grade inflation by the crowd. Notably, while the experts utilized nearly the full range of the grading scale, the crowd scores were compressed into a narrow range that overestimated scores on the lower range and underestimated the highest scoring videos (Fig 1B-C).

Crowd mean and expert mean sum scores were strongly correlated ($r = 0.865$, $p < 0.0001$) (Fig 1D). Despite the high correlation between expert and crowd scores, the absolute values of these scores for individual surgery videos were not in agreement, showing higher discordance for the videos assigned lower scores by the experts (Fig 1E). The ICC for the expert versus crowd mean sum scores for these 50 videos was -0.091, demonstrating statistically that the gap between the scores is large. Examining the mean sum scores for the videos grouped by level of experience showed a similar pattern: crowd mean sum scores were consistently higher than expert mean scores for first, second, and third year residents ($p < 0.0001$, paired t-test; Fig 1F). In addition, crowd versus expert scores for the PGY5 fellows approached but did not reach statistical significance ($p = 0.055$), possibly due to an underpowered sample size (n = 6).

Independently of crowd score, longer surgery duration (as defined by the duration of phacoemulsification) was strongly correlated with both lower training level ($r = -0.855$, $p < 0.001$; Fig 2A) and lower expert mean sum score ($r = -0.927$, $p < 0.0001$; Fig 2B). In a stepwise regression analysis to predict expert mean sum score, surgery length was a better predictor variable than crowd mean sum score. While both crowd score and surgery length were highly correlated with expert score ($r = 0.865$ and -0.927, respectively), neither metric alone predicted expert score very accurately. To determine whether using both metrics together could better predict expert score, a regression equation to convert crowd score plus surgery length into a predicted score was derived from the cross-sectional data. The equation was as follows: Predicted Expert Mean = -11.18 + -0.018*video_length (in seconds) + 1.643*crowd_score. This equation generated a predicted score that more closely approximated the real expert score when utilizing both crowd score and surgery length ($r^2 = 0.92$) (Fig 2C) as compared to using crowd score alone ($r^2 = 0.75$) or surgery length alone ($r^2 = 0.86$). The final prediction model fit the actual data well, with differences between real (mean expert) and predicted values varying within +/-3 points on the 25-point scale (Fig 2C). The predicted values and real expert mean scores displayed excellent correlation ($r = 0.959$, $p < 0.0001$) (Fig 2D), outperforming the accuracy of crowd scores alone (Fig 1D-E).

Longitudinal Study

Five resident surgeons provided phacoemulsification videos at 6 distinct time points in training (30th cataract case, 60th case, 90th case, 120th case, 150th case, and 180th case). Two surgeons were missing videos at a single time point due to technical problems, yielding 28 of 30 intended videos. Each video was graded by the same 5 blinded expert surgeons yielding 140 evaluations, and a total of 1,725 evaluations were furnished by 366 distinct, individual crowdworkers. The expert group took 8.5 days to complete the 28 video evaluations. For the crowdsourced evaluations, 76.2% (n = 1,314) satisfied the comprehension criterion for inclusion and the crowd evaluations were completed in 9.0 hours. The

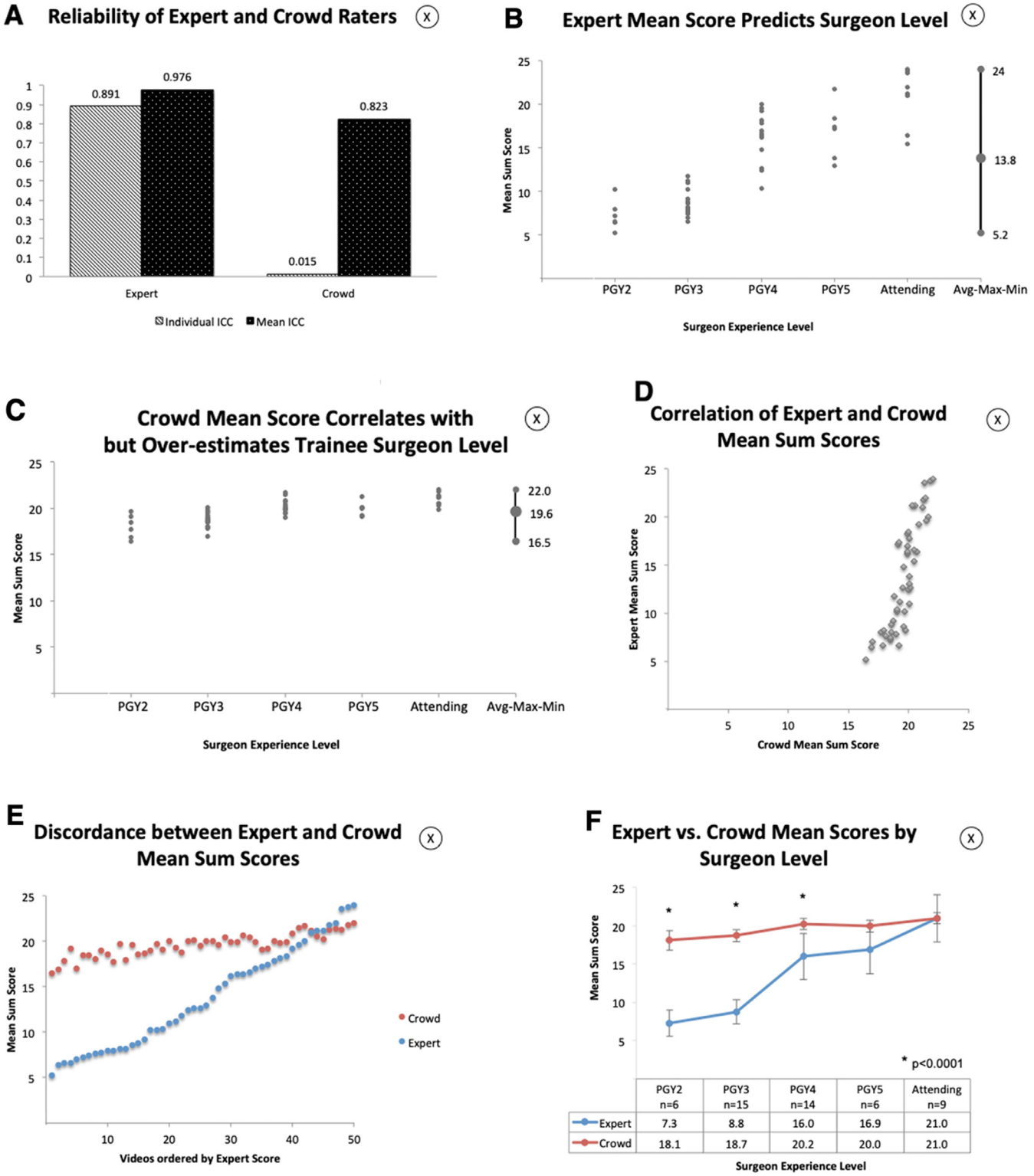

FIGURE 1. Crowd rater mean scores correlate with but do not agree with expert scores, and the crowd significantly overestimates surgeon ability for all 3 years of residency training. For the cross-sectional study arm ⊗ (n = 50): A: Reliability of blinded expert raters and crowd raters: Intraclass correlation coefficient (ICC), a measure of rater reliability by agreement as opposed to correlation, is shown for the sum scores of the experts and crowdworkers from the cross-sectional study. When scores are averaged across each group (mean ICC), the crowd performs nearly as well as experts, but the individual crowdworkers perform poorly in comparison with individual experts (individual ICC). B: Expert mean sum scores accurately predicted surgeon level of training, establishing construct validity for the modified OSATS grading rubric (Pearson's r = 0.860, $p < 0.0001$). C: Crowd mean sum score also correlates with surgeon level (r = 0.729, $p < 0.0001$) but not as well as the experts. The crowd used a narrow range of the grading scale weighted toward superior scores for all surgeons with a correspondingly elevated group mean, as compared to the experts (B) who utilized the full scoring range with a group mean close to the mid-range (Average-Maximum-Minimum plots). D: Crowd and expert mean scores were highly correlated (r = 0.865, $p<0.0001$), but failed to show good absolute

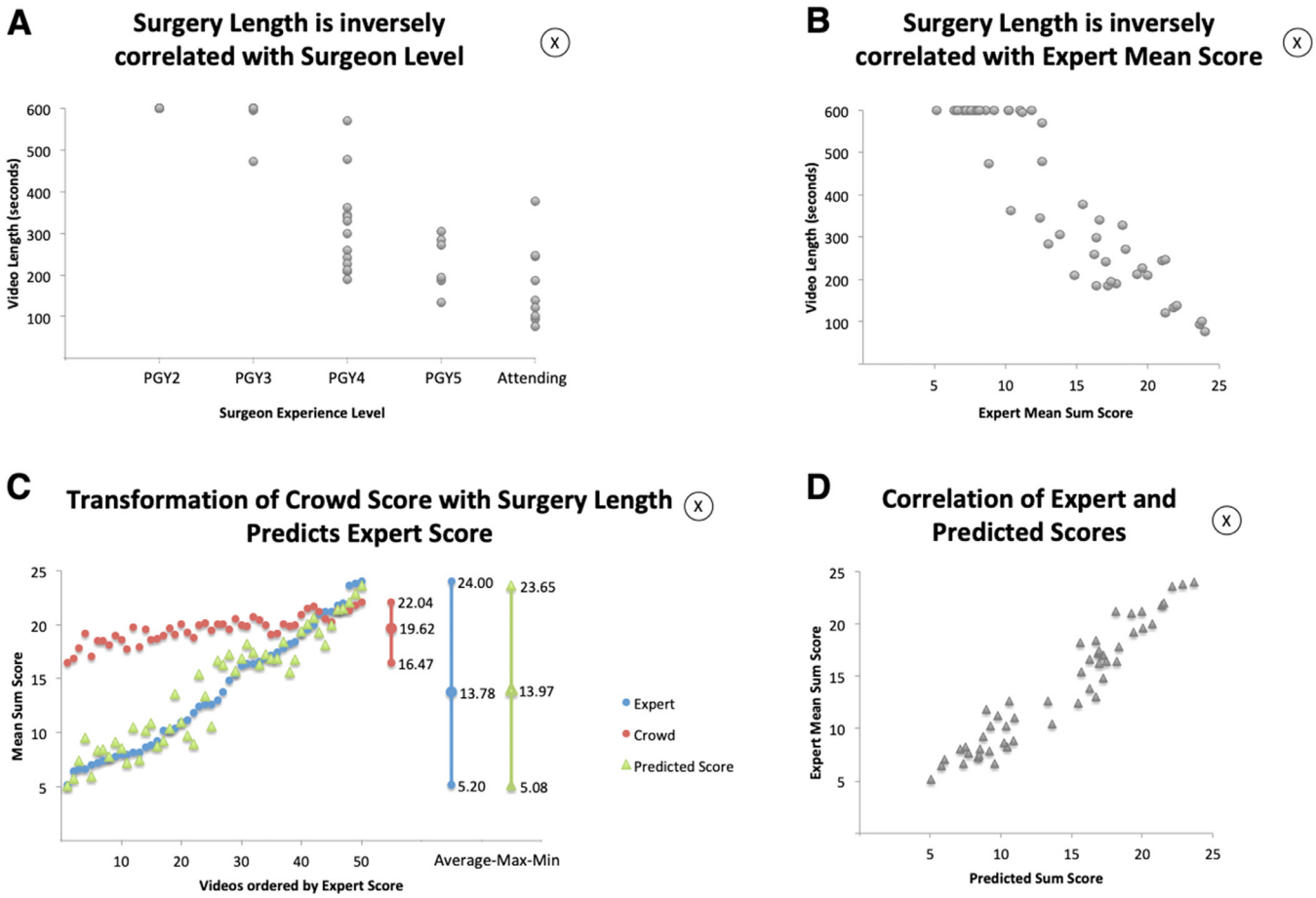

FIGURE 2. Cross-sectional cohort demonstrates that surgery duration correlates with surgeon training level and expert score, and surgery duration improves upon the predictive accuracy of crowd score to approximate expert score. For the cross-sectional study arm ⊗ (n = 50): A: Longer surgery length (as defined by phacoemulsification duration) was strongly correlated with lower training level (r = -0.855, $p < 0.001$). B: Longer surgery length was strongly correlated with lower expert mean score (r = -0.927, $p<0.0001$). C-D: A regression equation to convert crowd score plus surgery length into predicted expert score was derived from the cross-sectional data (Predicted Expert Mean = -11.18 + -0.018*video_length_in_seconds + 1.643*crowdscore). This equation generated a predicted score (green markers) that more closely approximated actual expert score (blue markers) ($r^2 = 0.92$) than did crowd score alone (red markers) as illustrated by (C) absolute values and (D) correlation plots (n = 50, r = 0.959, $p < 0.0001$).

median number of crowd evaluations per video was 47 (range 40-55).

The group expert mean sum score predicted level of surgeon training ($r = 0.833$, $p < 0.0001$) again showing construct validity for the mOSATS. Importantly, the expert mean score increased for each individual resident as the residents progressed through the 6 time points in cataract training (Fig 3E: blue markers), indicating discriminative construct validity for the mOSATS to track skill acquisition over time for any given trainee.

As with the cross-sectional study, the crowd used a narrower range on the grading scale (crowd 19.1-21.9 versus expert 9.0-20.8) and had a higher mean sum score averaged across all 28 videos (crowd 20.2 versus expert 15.3). Likewise, the crowd mean and expert mean sum scores for the longitudinal data set were again well correlated ($r = 0.792$, $p < 0.0001$; Fig 3A). The crowd mean score only moderately correlated with resident level of experience ($r = 0.662$, $p = 0.0001$).

Despite the correlation between expert and crowd scores, crowd scores were consistently higher than

value agreement (perfect agreement would plot linearly along y = x). E: Crowd and expert mean scores for individual surgery videos show discordance especially for videos given lower scores by the experts. F: Crowd mean scores were higher than expert scores for first, second, and third year residents ($p<0.0001$, paired t-test) and approached borderline significance for the PGY5 fellows ($p = 0.055$, paired t-test). PGY: postgraduate year. Ophthalmology residency begins in PGY2 after a year of general internship training. Table lists group means for each level of surgeon experience. Error bars indicate standard deviation.

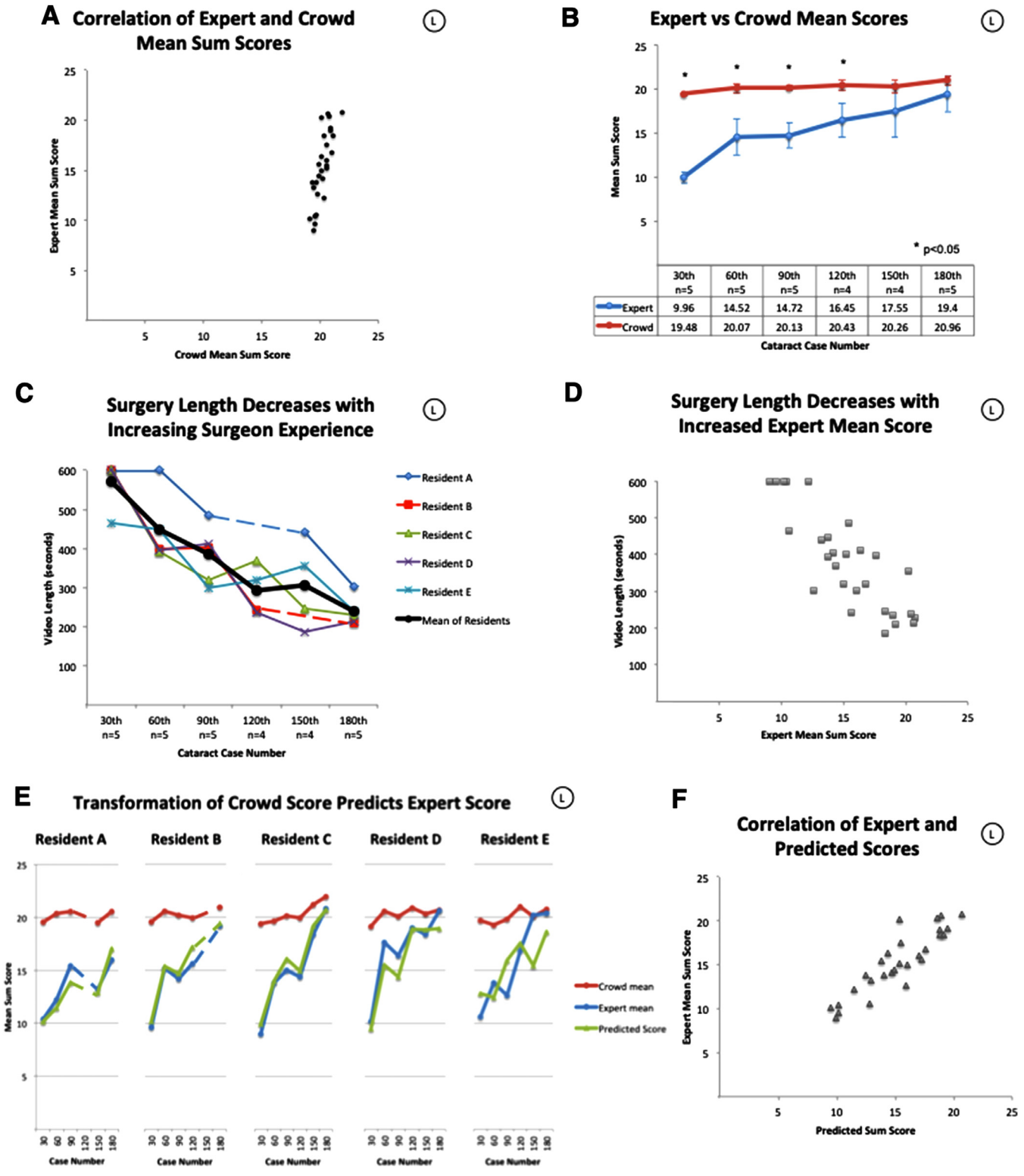

FIGURE 3. Longitudinal cohort reproduces correlation of expert scores with crowd scores and with surgery duration, and validates predictive model for estimating expert score. For the longitudinal study arm Ⓛ (n = 28): A: Crowd mean scores correlated with expert scores ($r = 0.792$, $p < 0.0001$). B: Resident physicians gain surgical experience with increasing case number, as measured by the blinded expert assessments. Similar to the cross-sectional study, crowd scores do not agree with expert scores and demonstrate significant over-estimation of skill for beginner-intermediate surgeons as compared to expert scores (30^{th}-120^{th} cases: $p < 0.05$, paired t-test) leading to higher averaged mean scores and a constricted grading range as compared to expert scores. Table lists group means for each level of surgeon experience. Error bars indicate standard deviation. C: Surgery length (as defined by phacoemulsification duration) inversely correlates with resident experience ($r = -0.827$, $p < 0.0001$). Hashed lines connect time points separated by missing data points to show overall trends. D: Longer surgery duration (mean of residents) was strongly correlated with lower expert mean sum score ($r = -0.845$, $p < 0.0001$). E-F: When applied to the independent longitudinal data set, the regression equation derived from the cross-sectional data set yields a predicted score that closely approximates the expert score ($r^2 = 0.80$) as illustrated by (E) absolute values and (F) correlation plots (n = 28, $r = 0.896$, $p < 0.0001$). Hashed lines connect time points separated by missing data points to show overall trends.

expert scores for the 30th, 60th, 90th, and 120th cases ($p < 0.05$, paired t-test; Fig 3B), and were not significantly different for the 150th and 180th cases. The ICC for the expert versus crowd mean sum scores for these 28 videos was low at -0.33, indicating a very large numerical score gap. Resident surgical time also decreased with increasing surgical experience, demonstrating an inverse correlation with level of training, i.e., more proficient residents performed cataract surgery faster ($r = -0.827$, $p < 0.0001$; Fig 3C). Similarly, longer surgery duration (as averaged across residents for each time point) was again strongly correlated with lower expert mean sum score ($r = -0.845$, $p < 0.0001$; Fig 3D).

The regression equation that was derived from the cross-sectional data to convert crowd score plus video length into a predicted expert score (Fig 2C) was then applied to the crowd scores and video lengths from the longitudinal data. The resulting predicted scores again more accurately approximated the real expert mean score ($r^2 = 0.80$) as compared with crowd score alone ($r^2 = 0.63$) (Fig 3E). The predicted values and real expert mean sum scores demonstrated excellent correlation ($r = 0.896$, $p < 0.0001$) and agreement (Fig 3F), surpassing the accuracy of the crowd scores alone (Fig 3A-B). This predictive accuracy from using the regression equation derived from the first data set on the independent second data set suggests robustness of our model. Combining the 2 datasets of videos and plotting the residuals (observed minus predicted expert sum score, as predicted from crowd sum score) for the combined dataset revealed the crowd systematically inflates scores for low performers (Fig S2); this biased, non-random residual pattern (heteroscedasticity) demonstrates how looking only at correlation (r and r^2) can be misleading and highlights the importance of first checking other statistical measures to validate the model.

Student Rater Analysis

We further hypothesized that non-surgeon lay raters with more exposure or training could outperform lay raters with no or minimal training. To test this hypothesis, the group of student workers with no surgical experience, who were exposed to dozens of cataract videos during the video editing process ($n = 3$ raters), later graded those same videos ($n = 78$ videos) in a randomized fashion. The mean composite scores by these student raters incorporated the full grading spectrum similar to the expert scores, in contrast to the crowd scores which included only a narrow scale range (Fig S3A). The student rater mean sum scores more closely approximated expert mean scores than did crowd mean scores for the cross-sectional data (Fig S3A) and the longitudinal data (Fig S3B). For the pooled data (n=78), the correlation between student rater mean sum scores and expert mean sum scores ($r = 0.913$, $p < 0.0001$) (Fig S3C) was slightly better than that of between the crowd and experts ($r = 0.852$, $p < 0.0001$) (Fig S3D). The ICC for the expert versus student rater mean sum scores was high at 0.82, indicating good score agreement and a much better model fit than the negative ICC indices seen in the crowd versus expert analyses.

DISCUSSION

While expert review of surgical video is considered the gold standard for grading surgical skill, it is time consuming and expensive and therefore not feasible as a routine assessment in surgical residency programs. Our study examined the hypothesis that crowdsourced lay rater evaluations of cataract surgery would be equivalent to expert evaluations, using real, non-simulated surgery videos sampled across levels of training and time spent in residency training.

We found that our masked experts consistently discriminated between levels of surgeon experience, providing construct validity for the modified OSATS assessment. In addition, our crowd and expert mean scores were highly correlated, similar to other crowdsourced assessment studies.[8,13,15,18] However, measures of correlation alone ignore the difference between absolute expert and crowd mean scores. Correlation does not describe the actual agreement between data, and is frequently misused to proclaim equivalence.[22,23] In particular, the crowd tended to give inflated scores to novice and intermediate surgeons. The crowd and expert scores were thus significantly different at lower levels of surgical experience, then tended to converge as experience increased, suggesting that the crowd was not able to discern beginner surgeons from accomplished ones. As a result, this crowdsourced platform did not demonstrate construct validity in the assessment of videotaped cataract surgery. In contrast to the crowd, the "student raters" in our study more closely matched expert scores and use of the entire grading scale range, with a correspondingly high ICC showing agreement. This may suggest that the subject matter (phaco technical skill) is too complex or nuanced for lay raters to assess with sufficient sensitivity, rather than a failure of crowdsourced assessment.

Study Strengths

Our expert panel demonstrated excellent interrater reliability (ICC) while utilizing the full grading scale. In order to avoid institutional and subjective biases, the experts were attending ophthalmologists from 3 different academic universities and 2 private practices, none of whom were involved in making the cataract

videos. Blinded expert review of videos reduced the potential of cognitive biases that may affect in-person evaluation, e.g., halo effect, by attendings who are familiar with and are simultaneously supervising the resident surgeon. In comparison with other studies of crowdsourced skill assessments, our study reports the univariate data and compares absolute values and spread (range), rather than focusing on correlation. We argue that the validity of the metric should be based on equivalence rather than correlation and presenting the data this way reveals large discrepancies between crowd and expert scores, especially for trainee surgeons for whom this technology was intended. We report the outcomes of two separate studies, cross-sectional and longitudinal performance, in which the second data set independently reproduced the score patterns seen in the first data set. The expert scores from the longitudinal study traced the learning curve for individual trainees. Despite a limited dataset with 5-6 data points per trainee, our data mirrored prior seminal work suggesting an inflection point for increased phaco efficiency and decreased vitreous loss complications after a resident's first 80 cases.[24] While our particular group did not have a surgically challenged resident, this model may potentially detect an individual whose trajectory lags behind peers and thus may benefit from customized additional training.

While crowd scores did not replicate expert scores, the close correlation prompted us to look for other variables that could improve upon crowd scores. Regression analysis suggested that time spent in surgery was a reliable indicator of skill level, and that the combination of crowd score and surgery length (as defined by the duration of phacoemulsification) was a better estimator of actual expert score than crowd score alone. Employing multiple variables to approximate expert evaluation proved to be a robust model. Applying the regression equation derived from the first data set to the second, independent set of surgery videos predicted the actual expert score with reasonable accuracy. This novel "conversion factor" will need to be tested in future studies to ascertain its reproducibility and generalizability to other surgical skill sets. Also, we would caution that faster surgery is not necessarily better, as it takes skill to be simultaneously fast and good at surgery. Our results likely reflect that less skilled learners moved more hesitantly and took time to consider next steps in comparison to more experienced surgeons.

Study Limitations

This study utilized a grading rubric with a global performance score based on 5 domains addressing general procedural elements, none of which were cataract specific. The objective of this study was to explore the feasibility and validity of crowdsourced assessments for cataract surgery, so in order to facilitate lay rater comprehension, the grading scale was intentionally less technically specialized than other more granular, cataract-specific assessment tools. Locally-developed granular assessments tailored to each training program will still play an important role in guiding resident education, whereas the implementation of an accessible, standardized skill evaluation tool like this one may better fulfill universal accreditation requirements and our profession's social contract with the public. A follow-up study is ongoing to examine the relative contribution of each of the 5 mOSATS domains to the crowd and expert scores, as well as analysis of the free-text expert rater comments using natural language processing techniques to determine what specific surgical steps residents are excelling at or need further practice.

Our study evaluated only the phaco segment as proof of concept for crowdsourced assessment. While phacoemulsification is a critical skill for ophthalmic surgeons, there are many more technical elements, as well as surgical judgment, clinical acumen, and patient counseling components of cataract and other ocular surgery, which are equally important yet outside the scope of this study. Encouragingly, work done by other researchers indicates that skill assessments for specific tasks such as phaco or capsulorrhexis are highly predictive of overall technical skill.[25] This suggests that trainee global surgical progression could be monitored using select rubric items.

Other limitations of our longitudinal study included a relatively small sample size of 5 surgeons and 2 missing video time points. We did not standardize cases by nuclear density or dictate a specific phaco technique. But these limitations also signal one of this study's strengths: that we examined real-world, actual surgery cases while leveraging existing workflows (routine video recording when available). In surgical training programs, more complex cases are typically assigned to more experienced trainees, decreasing the likelihood that more experienced surgeons received higher scores due to "easier" cataracts. As ophthalmology residency programs are relatively small compared to residencies like general surgery, our class size of 5 residents per year (a medium-sized program) provides compelling pilot data to inspire a future multisite collaborative study. Another challenge of this study was the need to manually segment the videos, although promising investigations on machine learning to discern phases of cataract surgery[26] imply this may soon become a historical problem. Finally, an important limitation of this study was the use of a private company's platform (C-SATS) to obtain the crowd evaluations, which may be financially prohibitive for training programs.[8,16] The cost of directly contracting with crowdworkers via Amazon Mechanical Turk marketplace,

which would not benefit from the C-SATS platform ease of use, was estimated at $46 to $96 per video for ~46 crowd evaluations (derived from an estimated Turker rate of $0.10 to $0.25 per minute ($6 to $15 per hour), plus an additional 20% markup by Amazon).[27] Nevertheless, when considered on a per resident cost basis, even $100 per graduating resident for one standardized cataract surgery assessment appears reasonable next to the Ophthalmic Knowledge Assessment Program (OKAP in-service exam: $395 per resident per year), the written qualifying exam (WQE: $1950), and the oral board exam ($1950).

Future Work

How can the crowd method be improved to better match the experts? Our student rater example suggests that simply educating crowdworkers, via repeated exposure to surgery videos and/or structured tutorials describing operative technique, may improve identification of lower performers as compared to naïve crowd raters. While there will probably still be some compromise in accuracy, crowdsourcing has the benefit of faster turnaround; in this study and earlier crowdsourcing assessment studies,[7,8,13,18] expert panels took days to weeks to return the evaluations, whereas the crowd took a matter of hours. Since the majority of ophthalmology resident surgical experience usually occurs in the final year of training, accurate assessments with swift feedback are essential to maximize the limited time remaining for remediation before the expected graduation date.[3,28] Another reasonable strategy might be to develop or utilize existing reading centers for surgical evaluations. A recent study showed that reading center graders were nearly as reliable and accurate as experts for assessment of oculoplastic morphological outcomes, whereas crowdsourced lay observers were less accurate and consistent despite a strong overall correlation between the three groups.[29] Reading centers may be more time- and cost-effective than experts while more reliable than naïve crowdworkers, and could represent a feasible middle ground.

Alternatively, artificial intelligence may play an increasing role in assessment. In theory, deep learning algorithms can be trained to detect distinct signatures of operative technique that experts intuit, such as speed or confidence of phaco tip movements.[30] Objective skill assessments for specific tasks are already available on cataract simulators with built-in modules, some of which have been shown to discriminate novice from expert surgical proficiency[12] and, once mastered in the simulated setting, may lead to improved live surgical performance by trainees.[31]

A major motivating factor in designing this study was to derive a faster, cheaper, more feasible standardized surgical assessment than expert review. While crowdsourcing at face value clearly has limitations as described in this study, strategies to refine the output, such as training the crowd or combining crowd scores with other surgery metrics, may enable this technology to still play a role in standardizing surgical assessment for resident education. If every resident across the country submitted their de-identified 100th cataract case to a national registry for standardized assessment, it could generate a large cohort data set via which real competency benchmarks may be developed. This could ensure uniformity with an accepted standard of surgical competence for graduating residents that is currently lacking. It could be the surgical skills equivalent of the OKAP in-service exam that tests medical knowledge at specific time points in residency training and compares trainee performance against their peers. It may even allow for personalized training plans, as individual residents may reach the competence benchmark sooner or later than the 86 cataract cases assigned by the Accreditation Council for Graduate Medical Education (ACGME).

Important future work for this and similar studies will entail linking the validated skill assessments to patient-related outcomes. A landmark study examining blinded expert review of videos by practicing bariatric surgeons found that lower peer ratings of skill were significantly associated with longer operations and higher complication rates.[32] Recently, patient outcomes in laparoscopic sleeve gastrectomy were shown to be associated with variations in surgical techniques as measured by blinded peer review of intraoperative videos.[33] Additionally, there is evidence for skill in one surgical procedure predicting skill in other surgical domains, suggesting that there is generalizability to this type of assessment.[25] Similar work could be done in ophthalmology and other procedural fields, however, scalability will depend on fostering the optimal assessment infrastructure that integrates into existing workflows while capitalizing on new technologies, such as machine learning.

Implications

Our study builds upon and refines conclusions from prior crowdsourcing assessment studies and suggests exciting avenues for further research, including bolstering crowd score accuracy via other variables such as duration of surgery, training the crowd for better accuracy, and a possible role for reading centers as an economic yet reliable alternative. With appropriate optimization, crowdsourcing surgical skill evaluations have the potential to become a standardized assessment that would permit collection of surgical skills data on a national scale that could inform standards for resident surgical competence.

CONCLUSIONS

We report the first study to examine the feasibility and validity of crowdsourcing evaluations of cataract

surgery videos. The data presented here show that while crowdsourced scores of surgical skill correlated well with the gold standard of expert scores, these methods are not equivalent as the crowd consistently overestimated technical competency. This trend was observed across two independent data sets of surgical videos, and the crowd-expert score discrepancy was particularly pronounced in the group of trainee surgeons whom we as educators are most interested in accurately assessing proficiency. Interestingly, a regression model that adjusted crowd scores based on surgery duration produced a surprisingly accurate predictive model for surgical skill as compared to expert scores. We believe this topic deserves further study as a means of facilitating crowdsourced assessment of surgical proficiency for reliable documentation of trainee progression and standardization of surgical assessment.

FUNDING

This work was supported in part by an unrestricted grant from Research to Prevent Blindness, Inc. to the Department of Ophthalmology and Visual Sciences at Washington University; Vision Core Grant P30 EY 02687 from the National Institutes of Health; the Elizabeth Ann Brandom Charitable Lead Trust; and Experiment.com Crowdfunding Platform. The sponsors or funding organizations had no role in the design or conduct of this research.

FINANCIAL DISCLOSURES

Declarations of interest: none relevant to this study.

SMC and ELW have received honoraria for invited lectures including the AUPO Excellence in Medical Education (ELW) and Straatsma (SMC) Awards. MOG and BSW have grant funding from NIH (NCATS UG1 EY025182, NEI UG1 EY025181, NEI R01 EY026199, NEI R01 EY026641, NEI R21 EY030524 and NEI R21 EY031125, NEI UG1 EY025183, NEI UG1 EY025182, NEI R01 EY026199, respectively). No financial disclosures for any other author.

ACKNOWLEDGMENTS

The authors thank Dr. Jenny Chen (UC Davis) for her assistance in resident physician recruitment and surgical education for this study; Amy Jones (Washington University) for her help coordinating consents; and the residents, fellows, and attending faculty who provided video recordings of their surgeries for this study.

REFERENCES

1. O'Day DM. Assessing surgical competence in ophthalmology training programs. *Arch Ophthalmol*. 2007;125:395–396.
2. Pellegrini VD Jr., Ferguson PC, Cruess R, et al. Sufficient competence to enter the unsupervised practice of orthopaedics: what is it, when does it occur, and do we know it when we see it? AOA critical issues. *J Bone Joint Surg Am*. 2015;97:1459–1464.
3. Gedde SJ, Volpe NJ, Feuer WJ, Binenbaum G. Ophthalmology resident surgical competence: a survey of program directors. *Ophthalmology*. 2020;127:1123–1125. Epub Feb 20.
4. Cremers SL, Lora AN, Ferrufino-Ponce ZK. Global Rating Assessment of Skills in Intraocular Surgery (GRASIS). *Ophthalmology*. 2005;112:1655–1660.
5. Saleh GM, Gauba V, Mitra A, et al. Objective structured assessment of cataract surgical skill. *Arch Ophthalmol*. 2007;125:363–366.
6. Golnik KC, Beaver H, Gauba V, et al. Cataract surgical skill assessment. *Ophthalmology*. 2011;118:427. e1-5.
7. 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.
8. Deal SB, Stefanidis D, Telem D, et al. Evaluation of crowd-sourced assessment of the critical view of safety in laparoscopic cholecystectomy. *Surg Endosc*. 2017;31:5094–5100.
9. Paley GL, Shute TS, Davis GK, Culican SM. Assessing Progression of resident proficiency during ophthalmology residency training: utility of serial clinical skill evaluations. *J Med Educ Train*. 2017;1:018.
10. Fisher JB, Binenbaum G, Tapino P, Volpe NJ. Development and face and content validity of an eye surgical skills assessment test for ophthalmology residents. *Ophthalmology*. 2006;113:2364–2370.
11. Taylor JB, Binenbaum G, Tapino P, Volpe NJ. Microsurgical lab testing is a reliable method for assessing ophthalmology residents' surgical skills. *Br J Ophthalmol*. 2007;91:1691–1694.
12. Thomsen AS, Kiilgaard JF, Kjaerbo H, et al. Simulation-based certification for cataract surgery. *Acta Ophthalmol*. 2015;93:416–421.
13. 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.

14. Aghdasi N, Bly R, White LW, et al. Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. *J Surg Res*. 2015;196:302–306.
15. Polin MR, Siddiqui NY, Comstock BA, et al. Crowd-sourcing: a valid alternative to expert evaluation of robotic surgery skills. *Am J Obstet Gynecol*. 2016;215:644. e1-e7.
16. Conti SL, Brubaker W, Chung BI, et al. Crowd-sourced assessment of ureteroscopy with laser lithotripsy video feed does not correlate with trainee experience. *J Endourol*. 2019;33:42–49.
17. 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;30:447–452.
18. 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.
19. Martin JA, Regehr G, Reznick R, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg*. 1997;84:273–278.
20. Vaidya A, Aydin A, Ridgley J, et al. Current status of technical skills assessment tools in surgery: a systematic review. *J Surg Res*. 2020;246:342–378.
21. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1:307–310.
23. Porter AM. Misuse of correlation and regression in three medical journals. *J R Soc Med*. 1999;92:123–128.
24. Randleman JB, Wolfe JD, Woodward M, et al. The resident surgeon phacoemulsification learning curve. *Arch Ophthalmol*. 2007;125:1215–1219.
25. Mishra K, Zafar S, Vedula SS, Sikder S. Can we efficiently use structured rating scales to objectively assess global technical skill in cataract surgery? *J Cataract Refract Surg*. 2019;45:1682–1683.
26. Yu F, Silva Croso G, Kim TS, et al. Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA Netw Open*. 2019;2:e191860.
27. Samuel A. Amazon's Mechanical Turk has Reinvented Research. JSTOR Daily; 2018. https://daily.jstor.org/amazons-mechanical-turk-has-reinvented-research/. Accessed Aug 22, 2020.
28. Bartley GB. Verifying surgical competence: our fiduciary responsibility. *Ophthalmology*. 2020;127:997–999. Epub Apr 9.
29. Rootman DB, Bokman CL, Katsev B, et al. Crowd-sourcing morphology assessments in oculoplastic surgery: reliability and validity of lay people relative to professional image analysts and experts. *Ophthalmic Plast Reconstr Surg*. 2020;36:178–181.
30. Kim TS, O'Brien M, Zafar S, et al. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *Int J Comput Assist Radiol Surg*. 2019;14:1097–1105.
31. Thomsen AS, Bach-Holm D, Kjaerbo H, et al. Operating room performance improves after proficiency-based virtual reality cataract surgery training. *Ophthalmology*. 2017;124:524–531.
32. Birkmeyer JD, Finks JF, O'Reilly A, et al. Surgical skill and complication rates after bariatric surgery. *N Engl J Med*. 2013;369:1434–1442.
33. Chhabra KR, Thumma JR, Varban OA, Dimick JB. Associations between video evaluations of surgical technique and outcomes of laparoscopic sleeve gastrectomy. *JAMA Surgery*. 2020:e205532. Dec 16.

SUPPLEMENTARY INFORMATION

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.jsurg.2021.02.004.