Crowdsourced Assessment of Aesthetic Outcomes after Breast Reconstruction

Sumeet S. Teotia, M.D.
Jake A. Alford, M.D.
Yash Kadakia, B.A.
Nicholas T. Haddock, M.D.

Dallas, Texas

Background: Evaluating the aesthetic success of breast reconstruction can be difficult. Patients, surgeons, and the general population may differ in what constitutes a successful outcome. Recently, crowdsourcing has emerged to accumulate and analyze data on a massive scale. The authors propose that crowdsourcing can be a useful tool to reliably rate aesthetic outcomes of breast reconstruction.

Methods: One hundred one deidentified photographs of patients at various stages of breast reconstruction were gathered. Assessment tools included a five-point Likert scale and the transverse rectus abdominis myocutaneous (TRAM) visual assessment scale. Anonymous crowd workers and a group of expert reconstructive surgeons rated an identical set of photographs on the Likert scale. Crowd workers also rated the set of photographs on the TRAM scale.

Results: The authors obtained 901 anonymous, layperson evaluations on both Likert and TRAM scales. Crowdsourced assessment data collection took 28.6 hours. Expert assessment took 15 months. Expert and crowdsourced scores were equivalent on the Likert scale (overall interrater reliability, $\kappa = 0.99$; 95 percent CI, 0.98 to 0.99). Intrarater reliability among each subcomponent was highly reproducible for the crowd ($r = 0.98$; 95 percent CI, 0.97 to 0.99) and experts ($r = 0.82$; 95 percent CI, 0.77 to 0.87). Breast contour and positioning were most predictive of overall aesthetic result. Skin patch and scar were least predictive of overall aesthetic appearance.

Conclusions: Aesthetic outcomes rated by crowds were reliable and correlated closely with those by expert surgeons. Crowdsourcing can be a rapid, reliable, and valid way to assess aesthetic outcomes in the breast reconstruction patient. (*Plast. Reconstr. Surg.* 147: 570, 2021.)

Components of plastic surgery are based on visual analysis, and assessing visual outcomes is inherently subjective. This core idea suggests that evaluation of any aesthetic result in plastic surgery is imperfect, which is evident in breast reconstruction. Many factors contribute to a successful aesthetic outcome for the reconstructed breast. Although many surgeons and patients are guided by intuition, the degree to which each of these contributes to overall result is not completely known. In addition, discrepancies exist between lay-public perception, surgeon perception, and patient perception of a truly successful aesthetic outcome. One group, or individuals within a group, may value certain qualities over others.

There have been several attempts at quantifying such a subjective aesthetic challenge, each with their own drawbacks. Scoring rubrics and visual scales evaluating aesthetic results have been described.[1–3] These usually require multiple expert raters, which are prone to subjectivity and practical limitations of time and resources demanded of an expert surgeon, with few being validated. The highly validated BREAST-Q scales emphasize quality of life and patient satisfaction after reconstruction, but these are also subjective and do not directly address aesthetic results.[4,5] Many survey-based studies compare patient and surgeon preferences to identify discrepancies. These often have

From the Department of Plastic Surgery, University of Texas, Southwestern Medical Center.
Received for publication September 2, 2019; accepted July 21, 2020.
Presented (ePoster) at the 2020 Annual Meeting of the American Society for Reconstructive Microsurgery, in Fort Lauderdale, Florida, January 10 through 14, 2020.

DOI: 10.1097/PRS.0000000000007637

Disclosure: *None of the authors has a financial interest in any of the products or devices mentioned in this article.*

very small numbers of layperson raters. Virtually all of these suffer from the same flaw: they are all fundamentally reliant on small samples of subjective data. The result is that surgeons and patients are still searching for a reliable, consistent, and meaningful way to evaluate aesthetic outcomes.

Recently, crowdsourcing has emerged as a powerful tool for accumulating and analyzing data on a massive scale. In the broadest sense, crowdsourcing is a method of recruiting a large group of individuals to achieve a single task. Using secure online platforms, crowdsourcing can be used to recruit larger numbers of people more efficiently than ever. In surgery, it has been shown to effectively evaluate technical skill in a variety of surgical fields[6–13] and assess aesthetic outcome of cleft patients.[14]

Crowdsourcing can provide a unique opportunity for plastic surgeons, and with its use, aesthetic assessment can become more objective. We propose that large numbers of evaluations by laypersons can be used to score aesthetic outcomes after breast reconstruction as reliably as expert surgeons. We also propose that crowdsourcing can identify what components of breast reconstruction predict overall successful results.

PATIENTS AND METHODS

Photograph Sample Set

The study was approved and monitored by the University of Texas Southwestern Medical Center Institutional Review Board. We collected 101 patient photographs of the senior authors' (N.T.H. and S.S.T.) patients at various stages of breast reconstruction. Based on previous reports, we determined that a set of 101 images would provide an adequate range of outcomes and variations in appearance to assess reliability. All patients underwent breast reconstruction at the authors' institution. The number of previous operations or plan for future operations did not exclude participants. All patients consented to participation in the study by way of approved and monitored institutional photographic and research consent. Digital photographs were taken by the authors' professional office in a standardized frontal view as part of routine surgical follow-up. Images were cropped in a uniform fashion to standardize the positioning and lighting for the assessment platform. All photographs were deidentified before being uploaded to any assessment tool. Any patients with identifying marks, tattoos, clothing, piercings, or other identifying characteristics were excluded from participation. Photographs were handled on secure, Health Insurance Portability and Accountability Act of 1996–compliant platforms at all times from initial patient recruitment to final data analysis. Sixty-two photographs of patients undergoing autologous breast reconstruction were collected; 33 photographs of patients undergoing implant-based reconstruction were collected; and six photographs of patients undergoing oncoplastic reconstruction were collected.

Raters

Crowdsourcing

Using the Amazon Mechanical Turk Platform (www.mturk.com) through C-SATS (C-SATS, Inc., Seattle, Wash.), anonymous crowd workers were recruited to provide assessments and collect responses on a secure platform. Crowd workers were aged at least 18 years. All crowd workers were informed and gave consent by way of the online module. Nine hundred one responses were gathered from the crowd workers. Crowd participants were not required to participate in multiple surveys, and involvement in more than one survey was not excluded. The full set of 101 images (including autologous, implant-based, and oncoplastic photographs) was evaluated by means of crowdsourcing.

Expert Surgeons

Five plastic surgeons specializing in breast reconstruction were recruited to participate. Expert reconstructive surgeons evaluated the full 101-image set using the same anonymous, secure online platform. To minimize potential bias, the surgeons who operated on the patients being evaluated were not asked to rate those images.

Assessment of Aesthetic Outcome

Crowd workers and experts rated each photograph on a five-point Likert scale. Crowd workers were asked to rate the same images on a separate transverse rectus abdominis myocutaneous (TRAM) visual analogue scale to minimize the impact of a given scale's quality on the overall comparison between experts and laypersons. The five-point Likert scale attempted to capture a more discrete and formalized system of assessment, whereas the TRAM visual scale was included to capture more subjective variability between participants.

Likert Scale

Each group of reviewers assessed the images using a five-point Likert scale. Participants were shown an image and prompted to rate several parameters of the aesthetic appearance of the breast on a scale from 1 to 5, with descriptions corresponding to each discrete score. Visual guides

were not given to either group. Breast fullness; nipple-areola complex, shape, and contour; scar appearance, size, and fullness; and overall breast appearance were assessed in this manner.

TRAM Visual Analogue Scale

Crowd workers also assessed the images using the TRAM visual analogue scale. Participants were shown the same images at a separate time and prompted to rate additional parameters of the reconstruction. As opposed to the discrete Likert scale, participants were not given explicit scoring rubrics but rather were asked to rate various subcomponents of the aesthetic appearance of the breast on a continuous scale from 0 to 100. Breast fullness, nipple-areola complex, contour, scar appearance and placement, lower pole, positioning on the chest wall, skin appearance, and overall breast appearance were assessed in this manner.

Statistical Analysis

For each image, we collected scores from each group of raters (expert and crowd) using the assessment tools described above. Review time and costs incurred for each reviewer pool were also measured. For purposes of this study, we aggregated responses to consider crowd workers and expert surgeons as individual raters, to allow for high-powered intrarater and interrater reliability analysis. Intrarater reliability therefore refers to the reliability of individual crowd scores compared to other individual crowd scores for a given photograph. Interrater reliability refers to the aggregated crowdsourced score compared to the aggregated surgeon score for a given photograph. Thus, ranges and standard deviations for individual scorers were not necessary for analysis in this study, which simply attempts to compare massively aggregated layperson assessment to surgeon assessment. Intrarater reliability and confidence intervals were determined by calculating Pearson correlation coefficient (r) of total TRAM and Likert scores, and the subcomponent scores of each scale (e.g., fullness, nipple-areola complex, scar, shape, positioning). Interrater reliability and confidence intervals between expert assessment and crowd assessment were calculated using Cohen's kappa coefficient (κ) for total Likert score and for each subcomponent of the scale.

RESULTS

Raters

Nine hundred one crowd-worker responses were generated for each survey. Crowdsourced responses were gathered in discrete intervals over the course of 6 days. Total data collection time for crowdsourced responses was 28.6 hours. Incomplete responses were excluded from final analysis; however, no restriction was placed on the time taken to complete the required responses. Crowd workers were primarily located in the United States (88 percent) and India (11 percent). Respondents participated from nine different countries. Given limited time and resources of the participating expert scorers, expert reconstructive surgeon data took much longer. Responses from five surgeons were gathered over 13 months, with multiple reminders and prompts. Overall scores are shown in Table 1.

Outcome Assessment

Overall, crowdsourced aesthetic assessment data correlated highly with expert surgeon data for all components measured in this study. Interrater reliability between the crowd and experts was greater than 0.95 for all subcomponents of the rating scale, including an overall correlation of 0.99. Intrarater reliability was highly reproducible for both groups, with intrarater reliability of the crowd assessment significantly higher than the expert ratings.

Intrarater Reliability

The five-point Likert scale aggregated scores were consistent within and across groups. Pearson correlation coefficients measured intrarater reliability (r) of the crowd workers at 0.98 (95 percent CI, 0.97 to 0.99), with greater than 0.96 intrarater reliability for each subset. This demonstrates high precision and reproducibility of crowd worker assessment (Table 1). The Pearson correlation coefficient measuring intrarater reliability of the smaller cohort of surgeons was 0.82 (95 percent CI, 77 to 0.87), with subcomponent reliability ranging from 0.58 to 0.78 (Table 2). There was no difference in reliability among autologous, implant-based, or oncoplastic reconstruction subsets. This indicates strong correlation but wider variation in scores than the crowd workers.

Table 1. Intrarater Reliability of Crowd Assessment

Component	Intrarater Reliability Pearson (r)	95% CI
Overall	0.98	0.97–0.99
Breast fullness	0.96	0.94–0.97
Nipple and areola	0.98	0.97–0.99
Overall shape/contour	0.95	0.92–0.97
Scar	0.97	0.95–0.98
Size and placement	0.97	0.95–0.98

Table 2. Intrarater Reliability of Expert Assessment

Component	Intrarater Reliability Pearson (*r*)	95% CI
Overall	0.82	0.77–0.87
Breast fullness	0.58	0.49–0.66
Nipple and areola	0.78	0.72–0.83
Overall shape/contour	0.64	0.56–0.72
Scar	0.77	0.71–0.82
Size and placement	0.72	0.65–0.78

Table 3. Correlation between Crowd and Expert Assessment of Likert Scale Items

Likert Scale Component	Interrater Reliability (κ)	95% CI
Overall	0.99	0.98–0.99
Breast fullness	0.95	0.93–0.97
Nipple and areola	0.98	0.97–0.99
Overall shape/contour	0.96	0.94–0.97
Scar	0.98	0.97–0.98
Size and placement	0.98	0.96–0.99

Interrater Reliability

Cohen's kappa (κ) measuring interrater reliability between groups was measured at 0.99 (95 percent CI, 0.98 to 0.99), showing extremely high correlation between expert and layperson scores on the Likert scale. Each subcomponent of the scale also showed high correlation, with each subcomponent measuring interrater reliability at 0.95 or greater. There was no difference in reliability between autologous, implant-based, or oncoplastic reconstruction subsets. The data are summarized in Table 3 and Figure 1.

Predictive Value of Aesthetic Subcomponents

The predictive value of each subcomponent on the overall aesthetic score was measured using crowdsourced TRAM and Likert scores. Each subcomponent was compared to overall aesthetic scores to determine item-item correlation. Scar appearance and skin patch were observed to have the least correlation to overall aesthetic appearance. Scar appearance had the lowest correlation on both Likert (r = 0.67; 95 percent CI, 0.55 to 0.76) and TRAM (r = 0.69; 95 percent CI, 0.57 to 0.78) scales. Skin patch was measured on only the TRAM scale (r = 0.65; 95 percent CI, 0.51 to 0.74) but showed a low comparative correlation to overall aesthetic result.

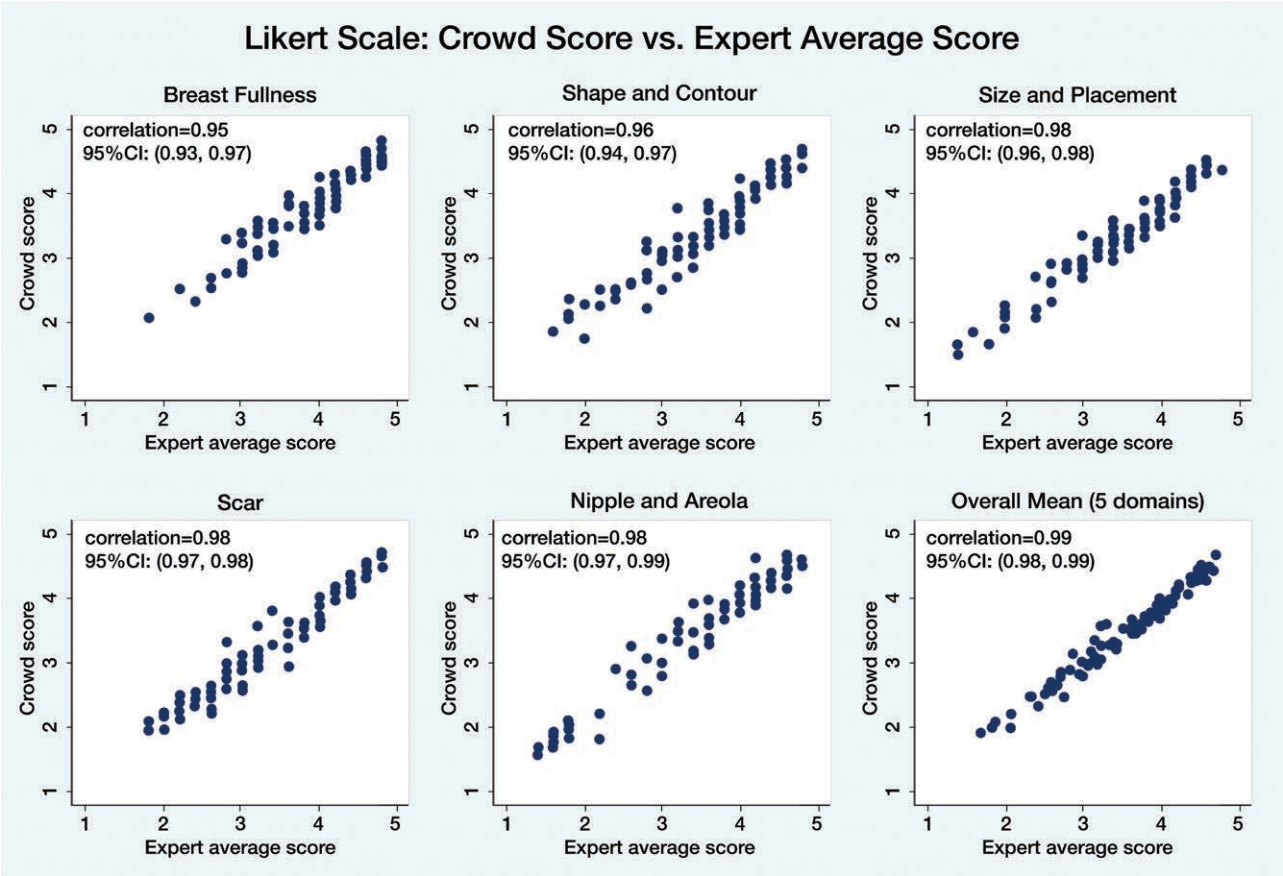

Fig. 1. Correlation between crowd and expert assessment of Likert scale items.

Table 4. Likert Scale Item-Item Correlation for Crowdsourced Data

Pearson Correlation Coefficients (*r*)	Correlation with Overall Appearance	95% CI
Breast fullness	0.70	0.58–0.79
Nipple and areola	0.75	0.64–0.82
Overall shape/contour	0.87	0.81–0.91
Scar	0.67	0.55–0.76
Size and placement	0.92	0.88–0.95

Table 5. TRAM Item-Item Correlation for Crowdsourced Data

Pearson Correlation Coefficients (*r*)	Correlation with Overall Appearance	95% CI
Breast fullness	0.79	0.70–0.85
Nipple and areola	0.73	0.62–0.81
Smooth/natural outline	0.89	0.84–0.92
Scar	0.69	0.57–0.78
Size and placement	0.95	0.93–0.97
Lower edge	0.73	0.64–0.79
Positioning	0.76	0.66–0.83
Skin patch	0.65	0.51–0.74

Breast contour and position of the breast were most highly correlated with overall aesthetic appearance. Breast contour had high correlation on both Likert (r = 0.87; 95 percent CI, 0.81 to 0.91) and TRAM (r = 0.89; 95 percent CI, 0.84 to 0.92) scales. Breast position and placement on the chest wall was most highly correlated with overall aesthetic score on both Likert (r = 0.92; 95 percent CI, 0.88 to 0.95) and TRAM (r = 0.95; 95 percent CI, 0.93 to 0.97) scales. Interitem correlations are shown in Tables 4 and 5, and representative graphs are demonstrated in Figures 2 through 5.

DISCUSSION

Crowdsourcing presents an opportunity to study public preferences of aesthetic outcomes on a larger scale than has been possible ever before. It is a powerful approach to gathering massive amounts of data. Using an online platform, it was possible to gather responses from thousands of raters from around the world rapidly. By sampling responses from close to 1000 respondents, we were able to address two major questions.

The first is whether or not a group of crowdsourced laypersons could rate the outcomes of breast reconstruction as accurately as expert trained breast reconstructive surgeons. Our evidence demonstrates that this was indeed true (κ = 0.99; 95 percent CI, 0.98 to 0.99). Although it is reasonable to assume that a single trained surgeon could evaluate outcomes better than a single or small group of the general public, there is a "strength in numbers" phenomenon. The massive amount of responses allowed for reliable and valid assessment equal to that of the experts. In fact, we also found that the crowdsourced responses had a greater intrarater reliability (r = 0.98; 95 percent CI, 0.97 to 0.99) and demonstrated less variability in responses than the smaller group of expert raters (r = 0.82; 95 percent CI, 0.77 to 0.87). This finding is likely attributable to the small group of experts recruited but is still a valuable insight. Although the individual assessment of a surgeon and patient is always invaluable, large-scale aggregated ratings, even when performed by laypersons, negates an individual's rating that is prone to variability. In addition, the data collection time was drastically different in the two groups, with expert data collected over many months and crowdsourced assessment completed in a matter of hours. This suggests that crowdsourcing is not only an effective and reliable assessment tool but also a practical and efficient one.

The second question addressed which factors the layperson values most in breast reconstruction. Our data suggest that breast position on the chest wall (r = 0.92 and r = 0.95) and contour (r = 0.87 and r = 0.89) were mostly closely correlated with overall aesthetic outcome. Scar appearance (r = 0.67 and r = 0.69) and skin patch (r = 0.65) were least predictive of a positive outcome. Previous studies have investigated this topic, and almost all involve collecting surveys or ratings from a small group of individuals and comparing those to surgeons' responses.[1,15,16] However, it is unwise to make generalizable conclusions about public perception of outcomes based on such small surveys. There is almost certainly sampling bias in these groups based on cultural, geographic, demographic, and other individual differences. Although our study does not claim to make any definitive statements about public preferences, to our knowledge this is the largest amount of data collected on patient preferences. We can say much more confidently that breast contour and positioning appear to be most valuable to patients and the general public. Of course, further investigation is needed on this topic, especially with more focus on specific cultural and geographic preferences.

Although there is a clear utility to crowdsourcing, limitations should be considered. First, aesthetic assessments inherently have limitations when determining a successful surgical result. As powerful and objective as an aesthetic rating tool may be, whether by crowd or by surgeon,

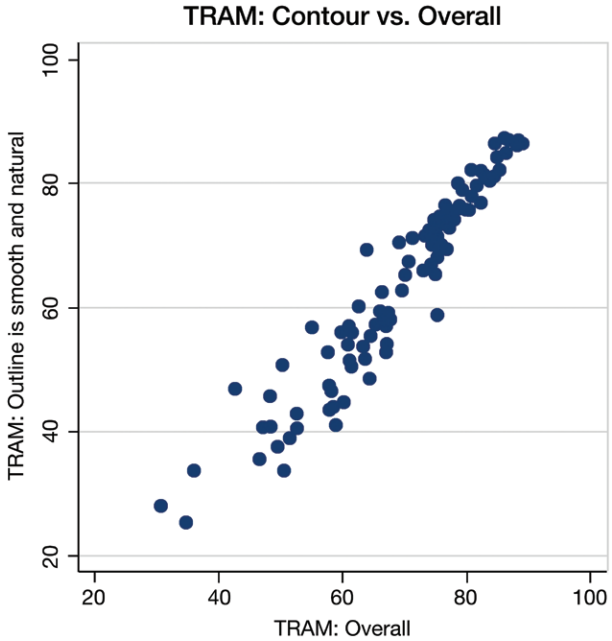

Fig. 2. Correlation between breast contour and overall aesthetic outcome as determined by TRAM scale crowd evaluations.

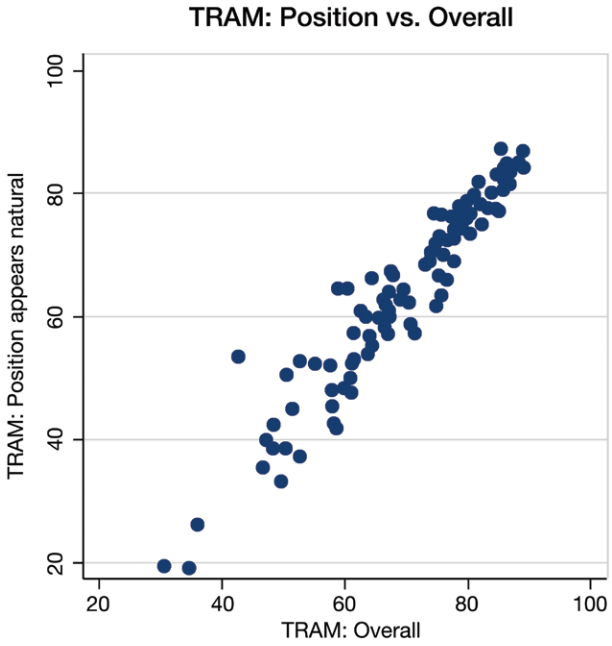

Fig. 3. Correlation between breast positioning and overall aesthetic outcome as determined by TRAM scale crowd evaluations.

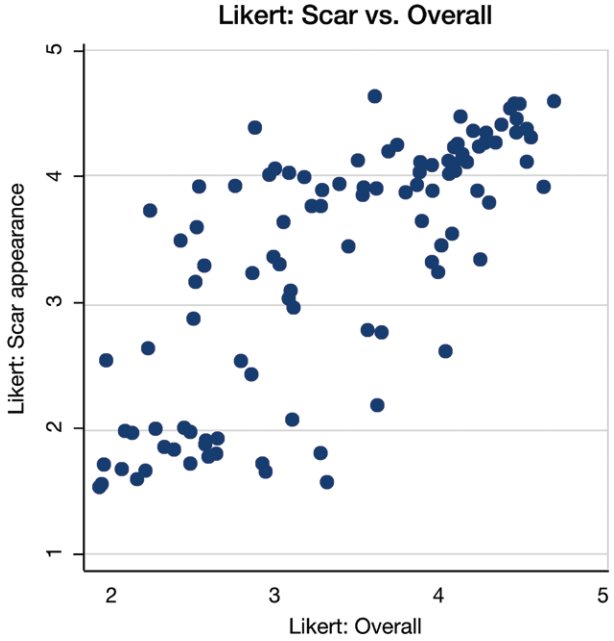

Fig. 4. Correlation between scar appearance and overall aesthetic outcome as determined by Likert scale crowd evaluations.

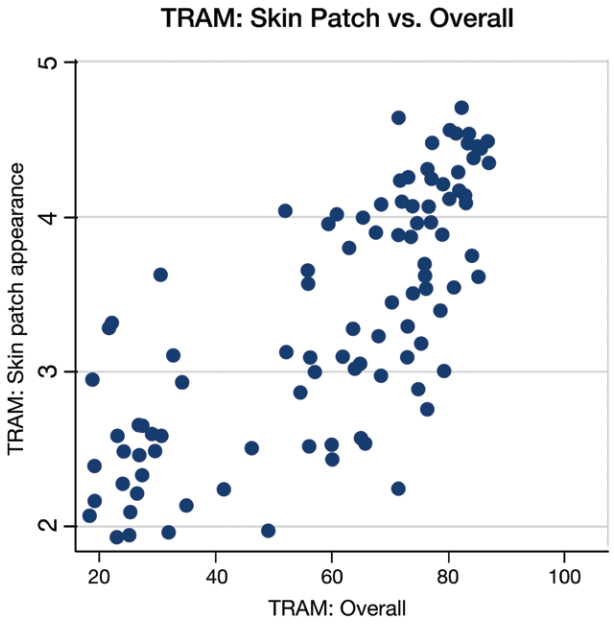

Fig. 5. Correlation between skin patch appearance and overall aesthetic outcome as determined by TRAM scale crowd evaluations.

the question remains of whether aesthetic assessment alone determines an overall positive result. Validated patient questionnaires may be equally important. Patient satisfaction arguably takes priority over any objective result. However, any quantitative assessment will always have utility as a tool in improving care.

The anonymity of the crowd workers allows for rapid and large-scale evaluation with minimal bias, but it limits the ability to investigate subsets of the crowd. We were unable to study differences in ratings by sex, race, ethnicity, age, or other demographic characteristics of the crowd workers themselves. It seems very likely that, had this

information been available, we would have found differences in many of these subsets. In addition, because of data collection constraints of the platform, we viewed the crowd-sourced workers as a single reviewer. This limited our ability to focus on specific factors that may have affected crowd workers' scores compared with the surgeons' scores. We do not have data on how long each individual reviewer took to score an image or a set of images. We do not know an individual crowd worker's range of scores or individual intrarater reliability. However, the power of crowd sourcing lies in the vast number of raters, making individual biases and confounding factors as statistically insignificant as possible. This is seen in the intrarater reliability of the crowd as a whole.

In addition, although demographic data of the patients were not collected, the aim of the study was not to use crowdsourced ratings to assess differences in various subsets of patients. The aim was primarily proof of concept. Given any subset of patient photographs, crowd workers are shown here to be able to assess aesthetic ratings reliably and as accurately as expert surgeons. Future studies should use this new validated assessment tool to further investigate aesthetic outcomes in various populations, looking at body mass index, number of operations, comorbidities, radiation therapy, age, race, ethnicity, and other potential confounders.

Our deidentified photographs were frontal images of only the breasts. A more robust truncal analysis would include oblique and lateral views, abdominal donor sites from deep inferior epigastric perforator reconstruction, and videos for assessment of animation deformity. In addition, stronger conclusions could have been made with a larger group of experts from multiple centers. We compared crowd workers to surgeons as a whole, but surgeons clearly do not always have the same gold standards for aesthetic outcomes. Individual preferences vary, and perceptions of our group of recruited surgeons may conflict with other centers. Many surgeons took months between assessments, so this lag time in scoring could have certainly affected the consistency of data collected from surgeons. Likewise, crowdsourced workers who rated images rapidly over a matter of hours were provided no visual guide for assessments. It is reasonable to expect that many workers adjusted their ratings as they viewed more images. However, the vast number of ratings by the crowd and the random order of the images to each rater were designed to minimize this potential bias. The major benefit of the crowd workers is the statistical power that thousands of assessments provide.

CONCLUSIONS

Crowdsourcing is a useful tool for assessing aesthetic outcomes after breast reconstruction. Crowdsourced assessment was equivalent to that made by expert surgeons, with higher intrarater reliability. Breast contour and positioning are most strongly correlated with successful aesthetic results. Crowdsourcing can be a rapid, reliable, and valid way to assess aesthetic outcomes in the breast reconstruction patient, and is likely beneficial in many other patients.

Nicholas T. Haddock, M.D.
1801 Inwood Road
Dallas, Texas 75390
nicholas.haddock@utsouthwestern.edu
Instagram: haddockmd
Twitter: haddockmd
Professional: drnicholashaddock

Sumeet S. Teotia, M.D.
1801 Inwood Road
Dallas, Texas 75390
sumeet.teotia@utsouthwestern.edu

ACKNOWLEDGMENTS

This study was funded by the University of Texas Southwestern Medical Center Department of Plastic Surgery. This work was supported by CSATS, Inc. (Seattle, Wash.), which was used as a vendor for this study. C-SATS provided assistance regarding rater recruitment, data collection, processing, and analysis from the online ranking modules.

REFERENCES

1. Song AY, Fernstrom MH, Scott JA, Ren DX, Rubin JP, Shestak KC. Assessment of TRAM aesthetics: The importance of subunit integration. *Plast Reconstr Surg.* 2006;117:15–24.
2. Maass SW, Bagher S, Hofer SO, Baxter NN, Zhong T. Systematic review: Aesthetic assessment of breast reconstruction outcomes by healthcare professionals. *Ann Surg Oncol.* 2015;22:4305–4316.
3. Visser NJ, Damen TH, Timman R, Hofer SO, Mureau MA. Surgical results, aesthetic outcome, and patient satisfaction after microsurgical autologous breast reconstruction following failed implant reconstruction. *Plast Reconstr Surg.* 2010;126:26–36.
4. Pusic AL, Klassen AF, Cano SJ, et al. Use of the BREAST-Q in clinical outcomes research. *Plast Reconstr Surg.* 2012;129:166–167.
5. Patel KM, Hannan CM, Gatti ME, Nahabedian MY. A head-to-head comparison of quality of life and aesthetic outcomes following immediate, staged-immediate, and delayed oncoplastic reduction mammaplasty. *Plast Reconstr Surg.* 2011;127:2167–2175.
6. 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.
7. Aghdasi N, Bly R, White LW, Hannaford B, Moe K, Lendvay TS. Crowd-sourced assessment of surgical skills in cricothyrotomy procedure. *J Surg Res.* 2015;196:302–306.

8. Deal SB, Lendvay TS, Haque MI, et al. Crowd-sourced assessment of technical skills: An opportunity for improvement in the assessment of laparoscopic surgical skills. *Am J Surg.* 2016;211:398–404.
9. White LW, Kowalewski TM, Dockter RL, Comstock B, Hannaford B, Lendvay TS. Crowd-sourced assessment of technical skill: A valid method for discriminating basic robotic surgery skills. *J Endourol.* 2015;29:1295–1301.
10. Polin MR, Siddiqui NY, Comstock BA, et al. Crowdsourcing: A valid alternative to expert evaluation of robotic surgery skills. *Am J Obstet Gynecol.* 2016;215:644.e1–644.e7.
11. Kowalewski TM, Comstock B, Sweet R, et al. Crowd-sourced assessment of technical skills for validation of basic laparoscopic urologic skills tasks. *J Urol.* 2016;195:1859–1865.
12. Katlic MR, Coleman J, Russell MM. Assessing the performance of aging surgeons. *JAMA* 2019;321:449–450.
13. Dimick JB, Scott JW. A video is worth a thousand operative notes. *JAMA Surg.* 2019;154:389–390.
14. Lendvay TS, White L, Kowalewski T. Crowdsourcing to assess surgical skill. *JAMA Surg.* 2015;150:1086–1087.
15. Tse RW, Oh E, Gruss JS, Hopper RA, Birgfeld CB. Crowdsourcing as a novel method to evaluate aesthetic outcomes of treatment for unilateral cleft lip. *Plast Reconstr Surg.* 2016;138:864–874.
16. Dikmans REG, Nene LEH, Bouman MB, et al. The aesthetic items scale: A tool for the evaluation of aesthetic outcome after breast reconstruction. *Plast Reconstr Surg Glob Open* 2017;5:e1254.