Surgical Endoscopy (2022) 36:8458–8462
https://doi.org/10.1007/s00464-022-09142-9

2021 SAGES ORAL

Blinded intraoperative skill evaluations avoid gender-based bias

Poppy Addison[1,2] · Daniel Bitner[2] · Paul Chung[2,3] · Saratu Kutana[2] · Samuel Dechario[4] · Gregg Husk[5] · Mark Jarrett[5,6] · Julio Teixeira[1,6] · Anthony Antonacci[1,5] · Filippo Filicori[1,2,6]

Received: 7 September 2021 / Accepted: 14 February 2022 / Published online: 23 February 2022

Abstract

Introduction Gender bias has been identified consistently in written performance evaluations. Qualitative tools may provide a standardized way to evaluate surgical skill and minimize gender bias. We hypothesized that there is no difference in operative time or GEARS scores in robotic hysterectomy for men vs women surgeons.

Methods Patients undergoing robotic hysterectomies performed between June 2019 and March 2020 at 8 hospitals within the same hospital system were captured into a prospective database. GEARS scores were assigned by crowd-sourced evaluators by a third party blinded to any surgeon- or patient-identifying information. One-way ANOVA was used to compare the mean operative time and GEARS scores for each group, and significant variables were included in a one-way ANCOVA to control for confounders. Two-tailed p-value < 0.05 was considered significant.

Results Seventeen women and 13 men performed a total of 188 hysterectomies; women performed 34 (18%) and men performed 153 (81%). Women surgeons had a higher mean operative time (133 ± 58 vs 86.3 ± 46 min, $p = 0.024$); after adjustment, there were no significant differences in operative time ($p = 0.607$). There was no significant difference between the genders in total GEARS score (20.0 ± 0.77 vs 20.2 ± 0.70, $p = 0.415$) or GEARS subcomponent scores: bimanual dexterity (3.98 ± 0.03 vs 4.00 ± 0.03, $p = 0.705$); depth perception (4.04 ± 0.04 vs 4.05 ± 0.02, $p = 0.799$); efficiency (3.79 ± 0.02 vs 3.82 ± 0.02, $p = 0.437$); force sensitivity (4.01 ± 0.04 vs 4.05 ± 0.05, $p = 0.533$); or robotic control (4.16 ± 0.03 vs 4.26 ± 0.01, $p = 0.079$).

Conclusion There was no difference in GEARS score between men vs women surgeons performing robotic hysterectomies. Video-based blinded assessment of skills may minimize gender biases when evaluating surgical skill for competency evaluation and credentialing.

Keywords Video-based assessment · Gender · Disparities · Education

Despite achieving gender parity in medical school enrollment, women make up only 22% of active general surgeons in 2020 [1]. Men are more likely to be preferentially treated in leadership opportunities, promotion, and salary [2]. While this represents a complex, pervasive social issue not limited to medicine, there has been some work over the last 20 years to determine the cause and consequences of gender disparities in surgery in hopes of achieving gender equality. Given that ideal physicians are often described with more traditionally masculine traits than feminine ones, gender role expectations contribute to the burden of advancement by women physicians [3]. These implicit biases are difficult to ignore when evaluating the competency of trainees [3–6].

Subjective, qualitative assessments have been an invaluable tool in surgical education, whether formalized through ACGME-mandated assessments or informal feedback given directly from the evaluator to the trainee. However, analyses of these assessments have found that men are more likely

✉ Poppy Addison
paddison@northwell.edu

[1] Department of General Surgery, Lenox Hill Hospital, Northwell Health, New York, NY, US

[2] Intraoperative Performance Analytics Laboratory (IPAL), Department of General Surgery, Lenox Hill Hospital, Northwell Health, New York, NY, US

[3] Department of General Surgery, Long Island Jewish Medical Center, Northwell Heath, Queens, NY, US

[4] Institute for Spine and Scoliosis, Lawrenceville, NJ, US

[5] Northwell Health, Manhasset, NY, US

[6] Donald and Barbara Zucker School of Medicine at Hofstra/Northwell, Hempstead, NY, US

to receive positive feedback [3, 5]. One analysis of subjective feedback across nine surgical specialties found that men received more superlative words (“terrific”, “a star”), and women received more phrases about personality (“always smiling”, “never seems to get upset”)[5]. This gender bias is internalized, with women residents under-rating their own technical skills despite blinded reviewers identifying no difference in skill between women and men [7, 8]. These findings are consistent across other fields outside of general surgery, suggesting that this disparity is systemic and requires careful consideration [7, 9].

Given that direct assessment and feedback provides critical guidance for the growth of surgical trainees, we hypothesized that blinded assessments could minimize gender bias. Traditional assessments remain necessary when blinding is not possible, such as the evaluation of non-operative skills; however, blinded assessments offer an opportunity to further improve surgical education by eliminating gender-related bias. Multiple standardized assessment tools to describe surgical skills have been devised in the past. For robotic surgery, the Global Evaluative Assessment of Robotic Skills (GEARS) score, which quantifies surgical skill with a Likert-based scale, has been validated to distinguish between novice and expert surgeons [10, 11]. The GEARS score describes 5 domains of robotic skill, with a higher number suggesting better technical skills [12]. We hypothesized that there would be no difference between the technical skill of men and women surgeons in gender-blind video-based assessments.

Methods

Patients undergoing robotic hysterectomies performed between June 2019 and March 2020 at 8 hospitals within a large hospital system were captured into a prospective database for retrospective analysis. Hysterectomies were chosen as the model procedure as there was a more even distribution of men and women surgeons performing this case as compared to other robotic procedures. Our hospital system routinely sends robotic videos for GEARS scoring, regardless of the surgical subspecialty; as such, the robotic approach was utilized in this study. Surgeon gender and mean operative time were collected. Because of privacy related concerns, we were unable to associate videos with each specific patient and therefore collect patient- or outcome-specific variables. Surgeon gender was determined from preferred pronoun on the hospital website. Years of experience was determined in combination with the hospital website and graduation year from Doximity (Doximity Inc., San Francisco, CA). This study was approved by the Institutional Review Board at Northwell Health and was deemed exempt. Consent was not required.

The methodology for video-based assessment was described previously by this group [13]. Skills assessment was provided by the Crowd-Sourced Assessment of Technical Skills (C-SATS) group (Seattle, WA). Online evaluators assigned GEARS scores without accessing information identifying the patient or surgeon. Evaluators did not have access to surgeon- or patient-identifying information, including the surgeon’s name. After assessment of technical skills, the videos and score were sent to the research team through a secure application program interface. All data were stored in a secure, HIPAA-compliant database under the hospital surgical quality improvement program.

Descriptive statistics are presented as follows: for continuous data: mean ± standard deviation; for categorical data, frequencies and percentages. The mean operative time and GEARS scores were determined for each surgeon, and one-way analysis of variance was used to compare the means for men and women surgeons. Two-tailed p-value < 0.05 was considered significant.

For adjustment for confounding variables, significant variables from the univariate screen (years of experience and number of cases) were included in a final one-way analysis of covariance (ANCOVA). ANCOVA is a linear model that blends analysis of variance (ANOVA) and regression, evaluating whether the means of a dependent variable are equal across an independent variable, while controlling for covariant variables. In this case, we evaluated a continuous outcome variable (duration and GEARS scores), continuous dependent covariates (operative volume and years of experience), and a categorical independent variable (gender). The assumptions of ANCOVA were tested as follows. The dependent (operative time and GEARS scores) and covariates (years of experience and number of cases) are continuous, and the independent variable (gender) is categorical. The groups were independently observed without significant outliers (Fig. 1). The residuals are normally distributed for each independent variable. The covariates are

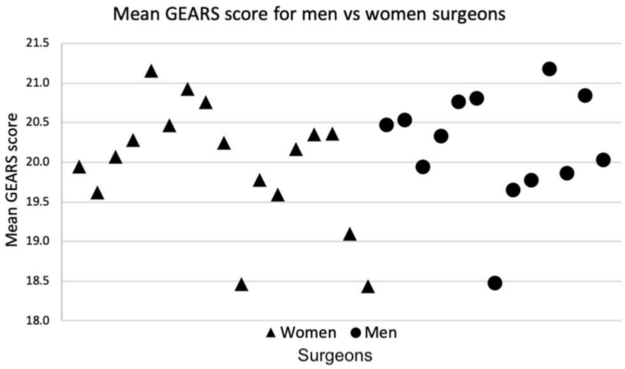

Fig. 1 Mean GEARS score for men vs women surgeons. On unadjusted analysis, there was no significant difference in mean GEARS score for men vs women surgeons

Table 1 Comparison between men and women surgeons sampled

Variable	Women ($n=17$)	Men ($n=13$)	p-value
Years of experience (years ± SD)	10.2 ± 5.4	17.6 ± 9.0	0.009
Number of cases (no ± SD)	2 ± 2.4	11.5 ± 11.1	0.003
Fellowship training	9	6	0.724
Female pelvic medicine and reconstructive surgery	4	2	
Gynecologic oncology	4	5	
Minimally invasive gynecologic surgery	1	0	

SD standard deviation, *no* number

Table 2 Difference in men and women surgeons' operative time and GEARS scores

Variable	Women ($n=17$)	Men ($n=13$)	Unadjusted p-value	Adjusted* p-value
Operative time (mins, mean ± SD)	133 ± 58	86 ± 46	0.024	0.607
Total GEARS score (score, mean ± SD)	20.0 ± 0.8	20.2 ± 0.7	0.415	0.431
Bimanual dexterity	3.98 ± 0.03	4.00 ± 0.03	0.705	0.426
Depth perception	4.04 ± 0.04	4.05 ± 0.05	0.799	0.325
Efficiency	3.79 ± 0.02	3.82 ± 0.02	0.437	0.408
Force sensitivity	4.01 ± 0.04	4.05 ± 0.05	0.533	0.407
Robotic control	4.16 ± 0.03	4.26 ± 0.01	0.079	0.854

GEARS score global evaluative assessment of robotic skill, *SD* standard deviation

*adjusted for years of experience and number of cases

linearly related to the dependent variables. There is homogeneity of variances with homoscedasticity. All analyses were performed with SPSS 26.0 (IBM, Armonk, NY) statistical software.

Results

Seventeen women and 13 men performed a total of 188 hysterectomies; women performed 34 (18%) and men performed 153 (81%). Women tended to be less experienced based on years of experience (10.2 ± 5.4 vs 17.6 ± 9.0, $p=0.009$) and submitted fewer robotic cases for scoring (2 ± 2.4 vs 11.5 ± 11.1, $p=0.003$) (Table 1).

On unadjusted analysis, women surgeons had a higher mean operative time (133 ± 58 vs 86.3 ± 46 min, $p=0.024$). There was no significant difference between the genders in total GEARS score (20.0 ± 0.77 vs 20.2 ± 0.70, $p=0.415$) (Fig. 1). Similarly, there was no significant difference between the groups in GEARS subcomponent scores: bimanual dexterity (3.98 ± 0.03 vs 4.00 ± 0.03, $p=0.705$); depth perception (4.04 ± 0.04 vs 4.05 ± 0.02, $p=0.799$); efficiency (3.79 ± 0.02 vs 3.82 ± 0.02, $p=0.437$); force sensitivity (4.01 ± 0.04 vs 4.05 ± 0.05, $p=0.533$); or robotic control (4.16 ± 0.03 vs 4.26 ± 0.01, $p=0.079$) (Table 2). There was no correlation between years of experience and GEARS score for men or women surgeons (Fig. 2) or between operative volume and GEARS score (Fig. 3).

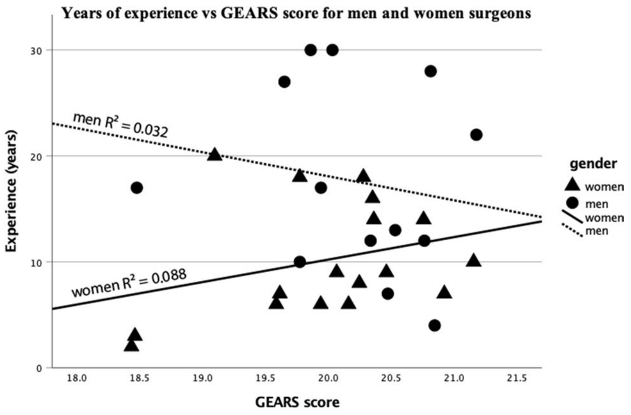

Fig. 2 Years of experience vs GEARS score for men and women surgeons. There was no significant correlation between years of experience and GEARS score for men or women surgeons

After adjustment for years of experience and number of cases, there was no significant difference between men and women in operative time ($p=0.607$), total GEARS score (0.431) or any of the GEARS subcomponent scores: bimanual dexterity ($p=0.426$); depth perception ($p=0.325$); efficiency (0.408); force sensitivity ($p=0.407$); robotic control ($p=0.854$).

A post hoc power analysis was performed to determine the necessary sample size to identify a difference in mean GEARS scores. Based on 80% power and α of 0.05, four

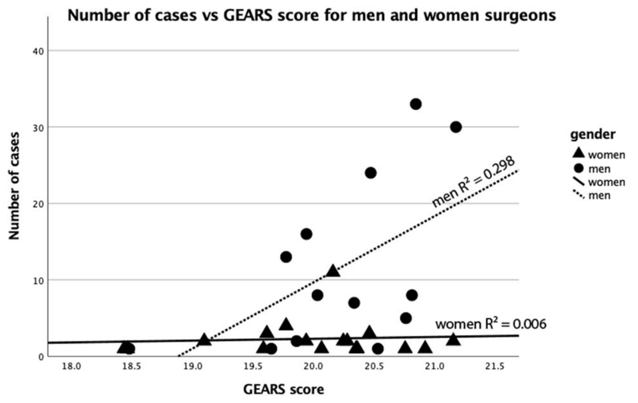

Fig. 3 Number of cases vs GEARS score for men and women surgeons. There was no significant correlation between number of cases and GEARS score for men or women surgeons

women and four men would be required to detect a 10% difference in GEARS scores, assuming an anticipated mean GEARS score of 20 ± 1 for men.

Discussion

This study aimed to examine gender-based discrepancies in intraoperative technical skill as evaluated by blinded, video-based assessment. We found that women surgeons tended to have a longer operative time; however, there were no differences in technical skill based on the GEARS score assigned by blinded, crowd-sourced evaluators. In a post hoc power analysis, this study was appropriately powered to identify a 10% difference in GEARS score. While this early work should be replicated in future multi-institutional studies across procedures and specialties, we suggest that blinded, video-based assessments may help to minimize gender bias in surgical evaluation.

Few studies consider gender as a potential bias in the evaluation of surgical trainees [14]. Currently, surgical education is based on qualitative, holistic feedback from attending surgeons, and self-assessment. However, unconscious gender biases affect this type of assessment, threatening the validity of the current way that we assess surgical trainees [14, 15]. Men tend to receive more positive narrative feedback from faculty [5]. Women tend to receive conflicting feedback, particularly regarding autonomy, assertiveness and receptiveness to oversight, making it difficult for women to improve their performance based on this feedback [3]. Gender bias is seen across specialties; in internal medicine, men faculty members rate men residents higher in clinical judgment, history, procedures, relationships, medical care and overall [9]. It is also seen in evaluations of students by faculty [6, 16] and in evaluations of faculty by trainees [4] and by students, with the largest gender discrepancy seen in surgical specialties [17]. This gender biased is internalized. In a blinded, video-based assessment of simulated surgical skills, women tended to underrate their performance as compared to expert assessment [7]. While not a solution to the problem of gender bias in surgery, blinded video-based assessment may help to minimize its impact on the judgment of surgical trainees.

Our current system may undervalue women surgeons. Repeated undervaluation as a trainee can result in women receiving less operative autonomy, impacting confidence, training quality, and performance [18]. In addition to the barriers faced in an unfavorable work environment with harassment, insufficient mentoring and leadership, and a male-dominated culture [19], these stereotypes may lead women to undervalue themselves, pursuing fewer leadership opportunities or shying away from more difficult cases, perpetuating a cycle where women are underrepresented in surgical decision-making.

While this study is among early work addressing gender bias in surgery, there are several important limitations to discuss. At the onset of the study, we sought to address gender disparities in general surgery; however, the majority of the robotic surgeries performed outside of gynecology were performed by men. The gender disparity at the attending level unfortunately resulted in our inability to study gender disparities. In the future, with a more equal distribution of men and women surgeons in all specialties, this study should be repeated outside of a gynecologic procedure. There is no current way to accurately describe the technical complexity of a surgical procedure. While the link between technical skill and outcomes has yet to be established [13, 20], potential future studies could use medical history or clinical outcomes as proxies for case difficulty and skill. Given that this was a small study, we were unable to perform subgroup analyses. In addition, we were not able to assess outcomes data as videos were deidentified. Although videos are routinely uploaded to the C-SATS database at our institution, surgeons still have the option of not doing so and the degree of compliance with uploading to C-SATS is unknown. Additionally, the vast majority of procedures were performed by men (153 vs 34), and we were unable to address whether men are busier, perform more robotic surgeries, or submit their surgeries more frequently to C-SATS. Taken together, then, this may suggest a selection bias. Despite these limitations, this early work establishes that video-based assessment provides a unique opportunity for gender-blind technical skill evaluation.

Conclusion

In this analysis of gender-blind video-based assessments in robotic hysterectomies, there were no differences identified between men and women surgeons in operative time, GEARS score or GEARS subcomponent scores as determined by crowd-sourced review. Incorporating video-based assessments into holistic review may minimize the impact of gender bias in surgical trainee evaluations.

Funding This work was supported by a 2020 SAGES Robotic Surgery Grant.

Declarations

Disclosures Dr. Teixeira is on the clinical advisory board for C-SATS, Inc. Dr. Filicori is a consultant for Cambridge Medical Robotics, Boston Scientific and Digital Surgery. Mr. Dechario, Ms. Kutana, Drs. Addison, Bitner, Chung, Husk, Jarrett, and Antonacci have no conflicts of interest or financial ties to disclose.

References

1. Association of American Medical Colleges. Active Physicians by Sex and Specialty, 2019. Physician Specialty Data Report Web site. https://www.aamc.org/data-reports/workforce/interactive-data/active-physicians-sex-and-specialty-2019. Published 2019. Accessed 07/26/2021.
2. Choo EK (2017) Damned if you do, damned if you don't: bias in evaluations of female resident physicians. J Grad Med Educ 9(5):586–587
3. Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM (2017) Gender differences in attending physicians' feedback to residents: a qualitative analysis. J Grad Med Educ 9(5):577–585
4. Angelo JL, Moazzez A, Neville A, Dauphine C, Lona Y, de Virgilio C (2019) Investigating gender differences in faculty evaluations by trainees in a gender-balanced general surgery program. J Surg Educ 76(6):e132–e137
5. Gerull KM, Loe M, Seiler K, McAllister J, Salles A (2019) Assessing gender bias in qualitative evaluations of surgical residents. Am J Surg 217(2):306–313
6. Turrentine FE, Dreisbach CN, St Ivany AR, Hanks JB, Schroen AT (2019) Influence of gender on surgical residency applicants' recommendation letters. J Am Coll Surg 228(4):356-365.e353
7. Miller BL, Azari D, Gerber RC, Radwin R, Le BV (2020) Evidence That Female urologists and urology trainees tend to underrate surgical skills on self-assessment. J Surg Res 254:255–260
8. Minter RM, Gruppen LD, Napolitano KS, Gauger PG (2005) Gender differences in the self-assessment of surgical residents. Am J Surg 189(6):647–650
9. Rand VE, Hudes ES, Browner WS, Wachter RM, Avins AL (1998) Effect of evaluator and resident gender on the American Board of Internal Medicine evaluation scores. J Gen Intern Med 13(10):670–674
10. Aghazadeh MA, Jayaratna IS, Hung AJ et al (2015) External validation of global evaluative assessment of robotic skills (GEARS). Surg Endosc 29(11):3261–3266
11. White L, Kowalewski T, Dockter R, Comstock B, Hannaford B, Lendvay T (2015) Crowd-sourced assessment of technical skill: a valid method for discriminating basic robotic surgery skills. J Endourol 29(11):1295–1301
12. Goh Alvin C, Goldfarb David W, Sander James C, Miles Brian J, Dunkin BJ (2012) Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 187(1):247–252
13. Addison P, Yoo A, Duarte-Ramos J et al (2020) Correlation between operative time and crowd-sourced skills assessment for robotic bariatric surgery. Surg. Endosc. 35(9):5303–5309
14. Klein R, Julian KA, Snyder ED et al (2019) Gender bias in resident assessment in graduate medical education: review of the literature. J Gen Intern Med 34(5):712–719
15. Antonoff MB, Feldman H, Luc JGY et al (2021) Gender bias in the evaluation of surgical performance: results of a prospective randomized trial. Ann Surg. https://doi.org/10.1097/SLA.0000000000005015
16. Axelson RD, Solow CM, Ferguson KJ, Cohen MB (2010) Assessing implicit gender bias in medical student performance evaluations. Eval Health Prof 33(3):365–385
17. Morgan HK, Purkiss JA, Porter AC et al (2016) Student evaluation of faculty physicians: gender differences in teaching evaluations. J Womens Health (Larchmt) 25(5):453–456
18. Meyerson SL, Odell DD, Zwischenberger JB et al (2019) The effect of gender on operative autonomy in general surgery residents. Surgery 166(5):738–743
19. Lim WH, Wong C, Jain SR et al (2021) The unspoken reality of gender bias in surgery: A qualitative systematic review. PLOS ONE 16(2):e0246420
20. Birkmeyer JD, Finks JF, O'Reilly A et al (2013) Surgical Skill and Complication Rates after Bariatric Surgery. N Engl J Med 369(15):1434–1442