Published online: 2020-01-20

The Role of Gender in Ophthalmology Resident Evaluations

Faith A. Birnbaum, MD[1] Shivram Chandramouli, BS[2] Mridul K. Thomas, PhD[3] Jullia A. Rosdahl, MD, PhD[1]

[1] Department of Ophthalmology, Duke University, Durham, North Carolina
[2] Duke University School of Medicine, Durham, North Carolina
[3] Centre for Ocean Life, DTU Aqua, Technical University of Denmark, Kongens Lyngby, Denmark

Address for correspondence Jullia A. Rosdahl, MD, PhD, Department of Ophthalmology, Duke University, 2351 Erwin Road, Durham, NC 27705 (e-mail: Jullia.Rosdahl@duke.edu).

J Acad Ophthalmol 2020;12:e8–e14.

Abstract

Background Gender affects various aspects of medical training. Prior research in surgical specialties has shown that female residents are given less positive feedback, granted less autonomy in the operating room, perform fewer procedures, and achieve competency milestones at a slower rate as compared with their male counterparts.

Purpose The purpose of this research is to evaluate whether gender affects ophthalmology resident evaluations at a single institution.

Methods Ophthalmology resident evaluations at a single residency program from 2010 to 2018 were reviewed. Data were collected on faculty gender, resident gender, and year of resident training. A linear mixed-effects model was utilized to analyze the degree to which differences in evaluation scores could be predicted from demographic data, while accounting for multiple sources of nonindependence of data.

Results A total of 490 evaluations for 43 residents by 34 faculty were analyzed. Evaluations consisted of up to 23 questions graded on a scale from 0 (poor) to 9 (excellent). Female residents received marginally higher scores than male residents on average (coefficient of male residents $= -0.2$). Both male and female residents received marginally lower scores from male faculty than from female faculty on average (coefficient of male faculty $= -0.21$). Male faculty also appeared to have scored male residents lower to a greater degree than did female faculty (coefficient of male faculty by male resident interaction $= -0.14$), though this result was sensitive to model specifications. There was no significant interaction between year of resident training and gender.

Conclusion In contrast to other procedural specialties, female residents appear to have been graded at a similar level or higher than male residents on average. Male faculty gave slightly lower scores to both male and female residents than female faculty did. Male faculty also may have graded male residents marginally lower than female residents to a greater degree than female faculty did.

Keywords
- ophthalmology
- resident education
- medical education
- disparity
- gender

received
August 27, 2019
accepted after revision
December 1, 2019

DOI https://doi.org/10.1055/s-0039-3402770.
ISSN 2475-4757.

Women in medicine hit a landmark in 2017: for the first time, more women enrolled in medical school than men.[1] More women have been choosing ophthalmology, with women making up 42.6% of ophthalmology residents in 2015, up from 35.6% in 2005.[2] Despite these growing numbers, the experience of residency training can differ based on gender.[3] Success in residency training can have far reaching consequences on career path, and thus it is important to investigate how gender may impact residency education.

Qualitative analyses can help to offer insight on subtle ways gender impacts thought patterns, while quantitative research can help to investigate how gender may influence residency milestones and in surgical training. In surgical residencies, comments made in evaluations of male surgical residents reflected more positive feedback on overall performance and future potential as compared with female surgical residents.[4] The application of implicit bias on resident performance was demonstrated by a thematic analysis of descriptive terms used to evaluate residents from nine surgical subspecialties at one instutition.[5] In reference to the resident's future, language was more often passive for females, such as "seemed to be" compared with more authoritative language, such as "he is."[5] Comments like "always smiling," "an absolute gem," and "never seems to get upset or angry" were only found in female evaluations.[5] In a multivariate regression model of performance feedback on laparoscopic training, female resident gender was significantly associated with the attending perception of more required intraoperative guidance even after accounting for case difficulty, year of training, and performance in practice laboratories.[6] In a similar study of thoracic residents, residents self-reported that female residents were given meaningful autonomy in 19.3% of the cases while males were given it in 33.3% of the cases ($p < 0.001$).[7] In emergency medicine, male residents achieved 12.7% higher rates of competency milestones than female residents, translating to 3 to 4 months of additional training.[8]

Gender disparity research has also been conducted within the field of ophthalmology. An analysis of 24 residency programs found that female residents performed a mean of 15 fewer cataracts and 58.1 fewer total procedures compared with male residents, which was not explained by taking parental leave.[9] Gender differences are apparent in ophthalmology careers as well, as female ophthalmologists submitted on average 936 fewer charges annually than males to Medicaid, even after accounting for time spent on clinical activity.[10] This difference resulted in females earning $0.56 for every dollar earned by a male in 2012 and 2013.[10] Female ophthalmologists also have fewer industry ties and receive less industry payments than male counterparts.[11]

There is a need to expand the understanding of how gender may affect ophthalmology residency training. Gaps in understanding remain in studying the strength of the effect of gender, how many aspects of residency training may be affected, and the consistency of this effect among different training programs. The aim of this study was to evaluate if gender affects ophthalmology resident evaluations at a single residency program.

Methods

The study protocol was reviewed by the Duke University Institutional Review Board and found to be exempt. Evaluations of first-year (PGY-2), second-year, (PGY-3), and third-year (PGY-4) ophthalmology residents were collected from the Duke University Department of Ophthalmology (from 2010 to 2018). Evaluations consisted of 9 to 23 questions (due to variable format over the study period, ►**Appendix A**) which were graded on a scale from 0 (poor) to 9 (excellent). Two questions were excluded from the analysis as they were changed in the standard evaluation template early in the data collection period and could not be compared across multiple years. These two excluded questions were "Overall assessment of resident performance" and "Moral and ethical behavior." Competencies included were history taking and communication; physical examination; medical knowledge, decision making, and application; procedural and surgical skills; professionalism; and rotation specific skills, such as checking pupils, using the indirect and direct ophthalmoscopes, the strabismus exam, and reading neuroimaging. The faculty gender, resident gender, and year of resident training were recorded along with the evaluation score of each question. Data were analyzed using the R statistical environment version 3.5.3. We made use of the packages *lme*[12] and *lmerTest*[13] for analysis, the *tidyverse*[14] for data handling and processing, and *effects*[15] and *ggplot2*[16] for plotting.

A linear mixed-effects model was used to explain variation in evaluation score (the dependent variable). The principal independent variables were faculty gender, resident gender, and resident year. Resident year was treated as a categorical variable with three levels. The model contained five fixed effect terms with three main effects (faculty gender, resident gender, and resident year) and two interaction terms (faculty gender by resident gender and resident gender by resident year). Additionally, three random intercept terms were used to account for nonindependence caused by (1) repeated evaluations of the same resident, (2) multiple evaluations by the same faculty, and (3) use of the same questions multiple times across the dataset.

To limit model complexity, we assumed that residuals were normally distributed even though the ratings scale was bounded and used discrete values. This assumption is frequently acceptable even when not completely theoretically justified.[17] To evaluate whether this influenced our results, we also fit a more complex generalized linear mixed model assuming a more theoretically justifiable quasibinomial distribution. This more complex model returned results that were very similar quantitatively, i.e., the coefficient values were highly similar. However, there were two small differences: (1) the faculty gender by resident gender interaction returned a p-value just above the conventional significance threshold of 0.05 ($p = 0.11$), and (2) the resident gender by resident year interaction returned one p-value that was near the threshold ($p = 0.07$ for male residents in PGY-3). These small changes do not substantively alter our conclusions since the results of the two models were quantitatively very similar, significance testing is fraught with

Table 1 Parameters of a linear mixed-effects model of resident evaluation score

Linear mixed-effects model parameters of variation in resident evaluation score			
Predictors	**Estimates**	**CI**	***p*-Value**
(Intercept)	6.94	6.11–7.76	**<0.001**
Faculty male gender	−0.21	−1.27 to 0.85	0.700
Resident male gender	−0.24	−0.62 to 0.15	0.243
PGY-3	0.55	0.45–0.65	$<\mathbf{2 \times 10^{-16}}$
PGY-4	0.86	0.74–0.98	$<\mathbf{2 \times 10^{-16}}$
Faculty male gender: Resident male gender	−0.14	−0.25 to −0.03	**0.013**
Resident male gender:PGY-3	0.12	−0.01 to 0.25	0.070
Resident male gender:PGY-4	−0.03	−0.20 to 0.14	0.710
Random effects			
σ^2	1.03		
$\tau_{Resident}$	0.39		
$\tau_{Faculty}$	2.46		
$\tau_{Question}$	0.04		
$N_{Resident}$	43		
$N_{Faculty}$	34		
$N_{Question}$	23		
Observations	6193		
Marginal R^2/ Conditional R^2	0.043/0.750		

Abbreviations: :, interaction term; CI, confidence interval; *N*, number of levels; PGY, postgraduate year; R^2, coefficient of determination; σ^2, residual variance; τ, variance of associated random effect.
Note: Independent variables consisted of faculty gender, resident gender, and resident year. Random effects consisted of resident, faculty, and question number. PGY-2 year and female gender have been arbitrarily designated as the baseline group.
Note: The *p*-values for Intercept, PGY-3, and PGY-4 are less than 2X 10^{-16}.

challenges in such models, and small changes across the arbitrary $p = 0.05$ threshold are not meaningful. Therefore, we choose to present results from the simpler, more easily interpretable model (►**Table 1**). However, we note that the significance of the gender-based difference is sensitive to model specification and would need additional data to resolve conclusively.

Results

A total of 490 evaluations were analyzed for 43 residents and 34 faculty. There were 23 (53%) male residents and 20 (47%) female residents, 18 (53%) male faculty, and 16 (47%) female faculty. There were 221 (45%) PGY-2 evaluations, 185 (38%) PGY-3 evaluations, and 84 (17%) PGY-4 evaluations. The mean and distribution of the scores by gender of resident and faculty is depicted by boxplot (►**Fig. 1**).

Contrary to our expectation, male residents received marginally lower scores than female residents (►**Fig. 2**, coefficient of male residents = −0.24). Both male and female residents received marginally lower scores from male faculty than from female faculty (coefficient of male faculty = −0.21). Furthermore, male faculty appeared to score male residents slightly lower than female residents to a greater degree than female faculty did (coefficient of male faculty by male resident interaction term = −0.14), although this result was sensitive to model specifications (see methods section for details). These fixed effects jointly accounted for just 4% of the evaluation score variance (Marginal $R^2 = 0.04$).

Evaluation scores improved strongly from PGY-2 to PGY-3 (coefficient = 0.55, 95% confidence interval [CI] 0.45–0.65), and also from PGY-2 to PGY-4 (coefficient = 0.86, 95% CI 0.74–0.99; ►**Fig. 3**). There was no evidence of a different rate of progression from PGY-2 through PGY-4 between male and female residents ($p > 0.05$).

Most of the variance in the evaluation scores was driven by differences between faculty members (variance = 2.5) followed by residents (0.39), and lastly, by question number (0.04). This indicates that approximately 70% of the variance in evaluation score was explained by individual differences between faculty members (Conditional $R^2 = 0.75$, Marginal $R^2 = 0.04$). This was because of a high degree of variability in faculty scoring practices (►**Fig. 4**): for example, among the two faculty members that gave the lowest evaluation scores, the highest score was 4, and the faculty member that gave the highest average score gave 7 as a minimum score.

Discussion

We found that the gender of the residents and faculty had a small impact on evaluation score, and the relationship was in the opposite direction to our expectation based on prior studies. In literature about resident education, female residents have been scored lower than male residents at the same level of training, have met milestones at a later stage of training compared with their male counterparts, and have reported fewer procedures specifically in ophthalmology residency.[4–9] In contrast, we found no evidence to support that female residents were scored lower than male residents nor that female residents met objectives later than male residents. In fact, our findings suggest that female residents may be scored higher than male residents by both male and female faculty.

The reasons for the difference in our findings compared with other published literature on surgical resident training are not clear. Although outright gender discrimination is not common in the medical professional workplace, implicit, also known as unconscious, bias has been found to be more common in medicine and involves a complex interaction of learned behavior, societal expectations, and cognitive associations.[18,19] Unconscious bias occurs when preconceived gender schemas are applied to situations and influence thoughts and behavior without conscious realization or intention.[19] Psychological experiments can illustrate how this may occur: college students were shown pictures of five people sitting around a table and were asked to identify the leader.[20] In mixed-gender groups, a man sitting at the head of the table was always identified as the leader. In comparison, a woman sitting at the head of the table had about equal odds of being identified as the

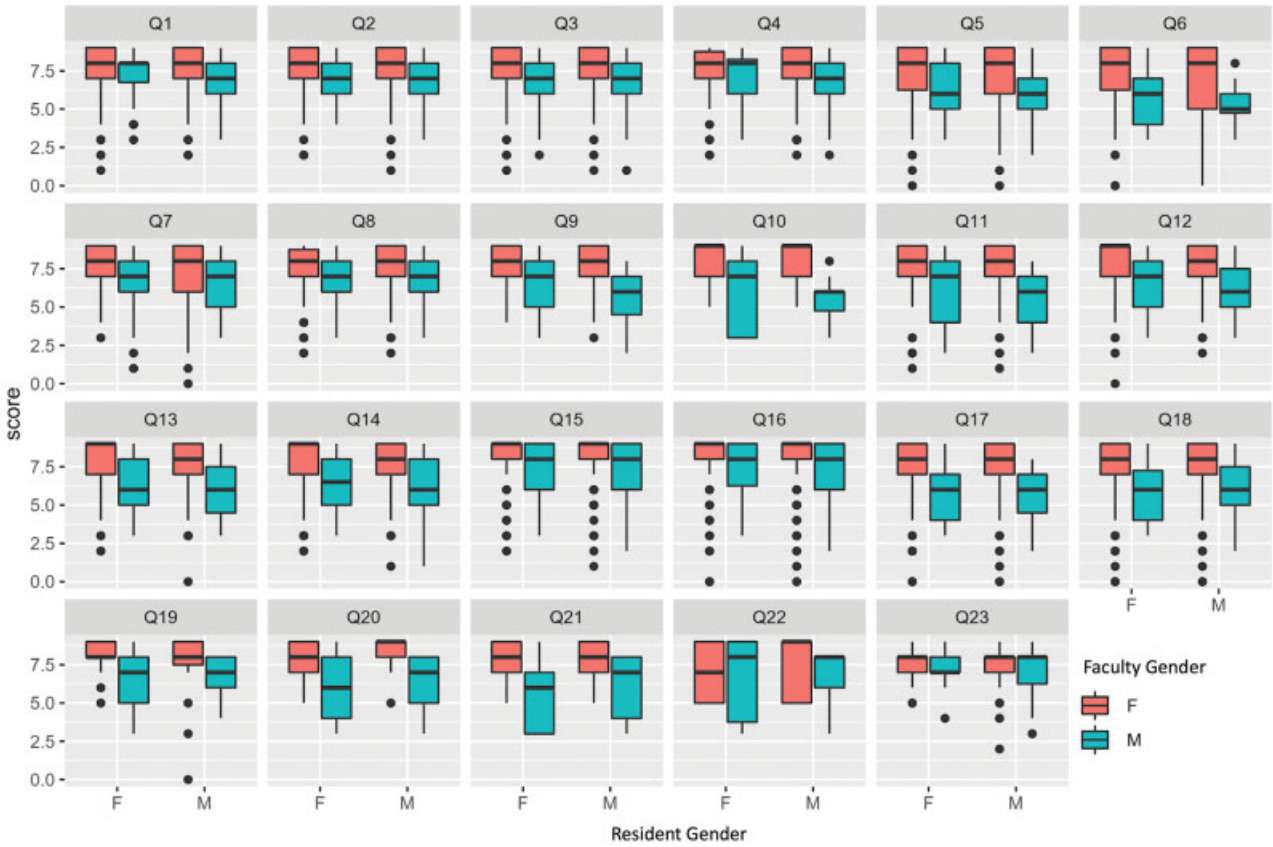

Fig. 1 Mean evaluation scores, by question. The boxplot shows the mean (*horizontal black line*), 25 and 75% percentile (*colored box*) and outliers (*black dots*) of each question score (Q1 through Q23) by male and female resident and male (teal) and female (peach) faculty. Note that sample sizes differ between questions and individual boxplots.

leader as a man sitting at the side of the table. It is possible that unconscious bias such as this may translate into giving females less autonomy in the operating room as found by prior mentioned studies.[9,18] Our department has a large number of faculty and is composed of 47% female faculty members; perhaps this translates to less implicit bias against female residents as female faculty have positions of power, authority, and competency. Our residency program was also composed of 47% female residents over the course of the study, which was higher than the national average. Possibly, female residents perform better in an environment that has a substantial population of other female residents and role models. Measuring unconscious bias is possible through self-assessments provided by Project Implicit, a nonprofit organization founded by a collaboration of researchers, and has been applied to health care providers in other research.[21,22] A future direction of our study could be to measure the unconscious bias across different residency programs and institutions and correlate it with gender differences in evaluations.

In addition to the differences in evaluations by resident gender, we also found modest evidence of differences based on faculty gender, with male faculty evaluating residents using marginally lower scores than female faculty. Perhaps this is due to gender differences in faculty expectations, evaluation styles, or communication styles. These systematic differences were substantially smaller than differences at the individual faculty level. The large variation in scores given by individual faculty was not surprising to us, given that our department is large and there is significant variation in personalities, grading standards, and teaching methods amongst our faculty.

Our findings regarding the lack of evidence that female residents are scored lower than male residents are encouraging for female residents. Evaluations are one aspect of training, however, and gender disparities are complex. Different metrics of gender bias in the field of ophthalmology indicate that gender disparities affect ophthalmology as in other medical specialties; women authored approximately 30 to 37% of ophthalmologic academic journal articles from 2002 to 2014[23] and received only 26.6% of National Institute of Health Grants in ophthalmology from 2011 to 2014.[24]

A limitation of our study in exploring gender bias may be that the standard evaluations used by our institution did not capture potential perceived differences—for example, our evaluations did not include questions at the same level of detail with regards to surgical skill assessment as those used in other studies that found gender disparities.[6,7] We also did not analyze comments, which have been shown to apply different descriptive terms based on gender in other studies.[5] Additionally, our underlying assumption in this analysis is that female and male residents actually do perform at the equivalent levels, and that differences in evaluations could be the result of bias. This underlying assumption may not be true; it could also be that one group performs at a higher level, say only

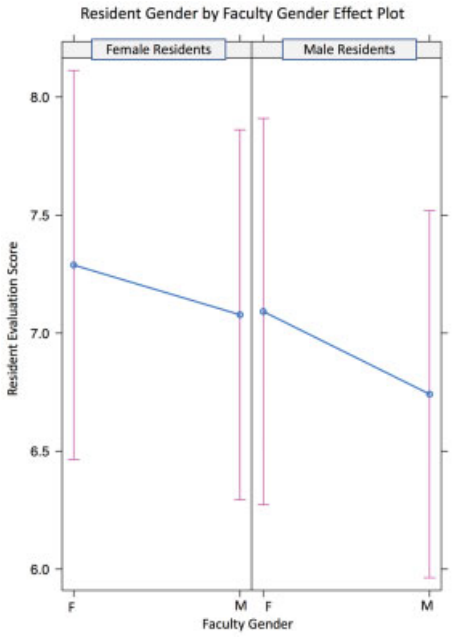

Fig. 2 Model-predicted average scores of male and female residents, by female and male faculty. The fixed effects plot shows the average score of female and male residents by female and male faculty. Male residents on average, were more likely to receive lower scores than female residents from all faculty (coefficient of male residents −0.24). Male faculty scored all residents lower than female faculty on average (coefficient of male faculty −0.21). Also, male faculty appeared to score male residents lower than female faculty to a greater degree than female faculty did (coefficient of male faculty by male resident interaction term: −0.14, 95% confidence interval −0.25 to −0.03, $p = 0.013$).

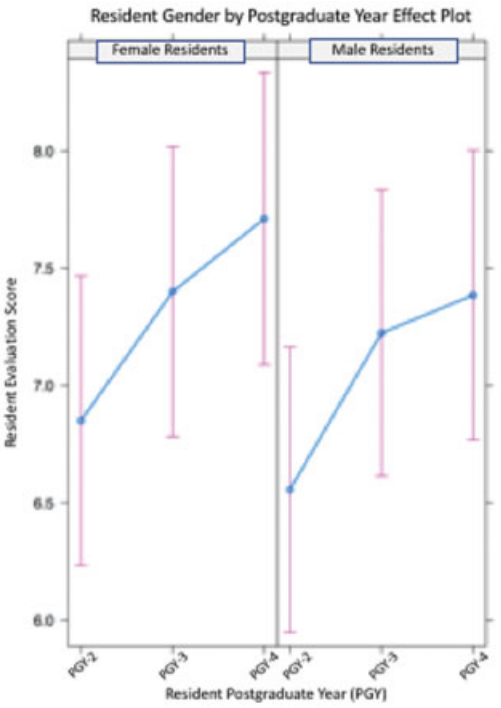

Fig. 3 Model-predicted average scores of male and female residents, by postgraduate year (PGY) of training. The fixed effects plot shows the average score of female and male residents by postgraduate year (PGY). Evaluation scores improved from PGY-2 to PGY-3 (coefficient 0.55, confidence interval [CI] 0.45–0.65), and from PGY-3 to PGY-4 (coefficient 0.86, CI 0.74–0.99). There was no evidence of a different rate of progression from PGY-2 through PGY-4 between male and female residents ($p > 0.05$).

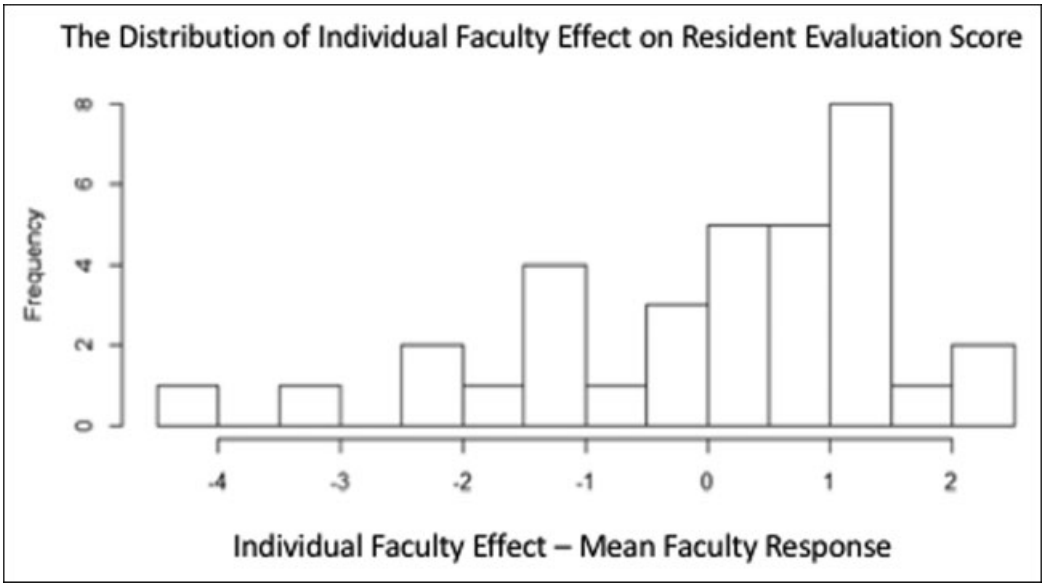

Fig. 4 Individual faculty effects on scores (individual faculty effect – mean score). The histogram shows the distribution of the differences between individual faculty responses and the average faculty response. This demonstrates a negatively skewed distribution of faculty evaluators around 0, where 0 represents the average faculty evaluator.

higher performing females have been selected at an earlier stage in their training, and perform higher than males. This would indicate that a lack of difference in evaluations actually reflects bias. However, we are limited in investigating this assumption, as we do not have available data on prior performance in medical school and college.

Other limitations of our study pertain to the retrospective design which did not account for possibility of transgender or intersex individuals. A mixed-methods design, with quantitative and qualitative components, for example with resident and faculty feedback, could provide more nuanced and potentially meaningful differences. Also, using evaluations across multiple ophthalmology residency programs would help to determine if our findings are consistent across ophthalmology as a specialty. Continued research on the effect of gender in medical education is warranted to promote supportive environments for all physicians to lead and thrive in medicine.

Conclusion

In contrast to other procedural specialties, female ophthalmology residents are not graded less than, and may be graded higher than, male residents. Female faculty graded residents of both genders slightly higher than male faculty. Male faculty may have graded male residents lower than female residents to a greater degree than female faculty did. There was a large degree of variation in scores by individual faculty.

Funding
M.K.T. received funding from the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska Curie grant agreement TROPHY No. 794264.

Conflict of Interest
None declared.

Acknowledgments
We are grateful to Renee Wynne, Program Director of Continuing Medical Education and Special Events, for her considerable time spent in collecting the data.

References

1 More women than men enrolled in U.S. medical schools in 2017. AAMCNews. 2017. Available at: https://news.aamc.org/press-releases/article/applicant-enrollment-2017/. Accessed June 20, 2019
2 Distribution of residents by specialty, 2005 compared to 2015. AAMC. 2016. Available at: https://www.aamc.org/download/481180/data/2015table2/. Accessed June 20, 2019
3 Serrano K. Women residents, women physicians and medicine's future. WMJ 2007;106(05):260–265
4 Gerull KM, Loe M, Seiler K, McAllister J, Salles A. Assessing gender bias in qualitative evaluations of surgical residents. Am J Surg 2019;217(02):306–313
5 Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM. Gender differences in attending physicians' feedback to residents: a qualitative analysis. J Grad Med Educ 2017;9(05):577–585
6 Hoops H, Heston A, Dewey E, Spight D, Brasel K, Kiraly L. Resident autonomy in the operating room: does gender matter? Am J Surg 2019;217(02):301–305
7 Meyerson SL, Sternbach JM, Zwischenberger JB, Bender EM. The effect of gender on resident autonomy in the operating room. J Surg Educ 2017;74(06):e111–e118
8 Dayal A, O'Connor DM, Qadri U, Arora VM. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. JAMA Intern Med 2017;177(05):651–657
9 Gong D, Winn BJ, Beal CJ, et al. Gender differences in case volume among ophthalmology residents. JAMA Ophthalmol 2019 (e-pub ahead of print). Doi: 10.1001/jamaophthalmol.2019.2427
10 Reddy AK, Bounds GW, Bakri SJ, et al. Differences in clinical activity and medicare payments for female vs male ophthalmologists. JAMA Ophthalmol 2017;135(03):205–213
11 Reddy AK, Bounds GW, Bakri SJ, et al. Representation of women with industry ties in ophthalmology. JAMA Ophthalmol 2016;134(06):636–643
12 Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw 2015;67(01):1–48. Doi: 10.18637/jss.v067.i01
13 Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: tests in linear mixed effects models. J Stat Softw 2017;82(13):1–26. Doi: 10.18637/jss.v082.i13
14 Fox J, Weisberg S. An R Companion to Applied Regression, 3rd ed. Thousand Oaks, CA: Sage Publications Inc; 2019
15 Wickham H. Tiddyverse: Easily Install and Load the 'Tidyverse'. 2017. R package version 1.2.1. Available at: https://tidyverse.tidyverse.org/. Accessed date July 25, 2019
16 Wickham H. Ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag; 2016
17 Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002;23:151–169
18 Phillips NA, Tannan SC, Kalliainen LK. Understanding and overcoming implicit gender bias in plastic surgery. Plast Reconstr Surg 2016;138(05):1111–1116
19 Valian V. Beyond gender schemas: improving the advancement of women in academia. Hypatia 2005;20(03):198–213
20 Porter N, Geis FL. Women and nonverbal leadership cues: when seeing is not believing. In: Mayo C, Henley N, eds. Gender and Nonverbal Behavior. New York, NY: Springer; 1981
21 Project Implicit. Implicit social cognition 2011. Available at: https://www.projectimplicit.net/index.html. Accessed November 24, 2019
22 Maina IW, Belton TD, Ginzberg S, Singh A, Johnson TJ. A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test. Soc Sci Med 2018;199:219–229
23 Mimouni M, Zayit-Soudry S, Segal O, et al. Trends in authorship of articles in major ophthalmology journals by gender, 2002–2014. Ophthalmology 2016;123(08):1824–1828
24 Svider PF, D'Aguillo CM, White PE, et al. Gender differences in successful National Institutes of Health funding in ophthalmology. J Surg Educ 2014;71(05):680–688

Appendix A Resident rotational evaluation questions

Patient interview.
Patient examination.
Office diagnostic procedures.
Disease diagnosis.
Nonsurgical therapy.
Nonoperating room surgery/consultation.
OR surgery.
Demonstrate level appropriate knowledge applied to patient management.
Incorporate cost-effectiveness, risk/benefit analysis, and IT to promote safe and effective patient care.
Work in interprofessional teams to enhance patient safety, identify system errors, and implement solutions.
Self-directed learning (1. Identify strengths, deficiencies and limits in one's knowledge and expertise, 2. Set learning and improvement goals, 3. Identify and perform appropriate learning activities, 4. Use information technology to optimize learning).
Compassion, integrity, and respect for others: sensitivity and responsiveness to diverse patient populations.
Responsiveness to patient needs that supersedes self-interest.
Respect for patient privacy and autonomy.
Accountability to patients, society, and the profession.
Communicate effectively with patients and families with diverse socioeconomic and cultural backgrounds (1. Rapport development, 2. Interview skills, 3. Counsel and educate, 4. Conflict management).
Communicate effectively with physicians, other health professionals, and health-related agencies (1. Comprehensive, timely and legible medical records, 2. Consultation requests, 3. Care transitions, 4. Conflict management).
Work effectively as a member or leader of a health care team or other professional group (1. Clinical team [outpatient clinic, inpatient consult service], 2. OR team, 3. Professional work group (e.g. QI committee)).
Patient examination-specific skills: slit lamp, ophthalmoscopy, ocular motility.
Nonoperating room surgery specific procedures: lasers, pupils/patient examination: external.
Operating room surgery-specific procedures: cataract/office diagnostic procedures: Neuroimaging/patient examination: slit lamp.
Nonoperating room surgery: chalazion, excision, biopsy, lid lesion, temporal artery biopsy/patient examination: direct and indirect.
Patient care: surgical judgment.

Abbreviations: IT, Information Technology; OR, operating room; QI, quality improvement.