ORIGINAL RESEARCH

Impact of Gender on Clinical Evaluation of Trainees in the Intensive Care Unit

Jenna Spring[1,2], Caroline Abrahams[3], Shiphra Ginsburg[4,5,6], Dominique Piquette[1,2], Fernando Martinez Guasch[7], Alex Kiss[8], and Sangeeta Mehta[1,4,6]

[1]Interdepartmental Division of Critical Care Medicine, [3]Postgraduate Medical Education, Temerty Faculty of Medicine, [6]Department of Medicine, and [8]Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada; [2]Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada; [4]Sinai Health System, Toronto, Ontario, Canada; [5]Wilson Centre for Research in Education, Toronto, Ontario, Canada; and [7]Division of Pulmonary and Critical Care Medicine, Tufts Medical Center, Boston, Massachusetts

ABSTRACT

Background: Gender disparities in medical education are increasingly demonstrated, including in trainee assessment.

Objective: This study aimed to evaluate whether gender differences exist in trainees' evaluation during intensive care unit (ICU) rotations, which has not been previously studied.

Methods: We reviewed the in-training evaluation reports (ITERs) for trainees rotating through five academic ICUs at the University of Toronto over a 10-year period (2007–2017). We compared the mean global score for the rotation and the mean score for seven training subdomains between men and women trainees. All scores were reported on a scale of 1 (unsatisfactory) to 5 (outstanding).

Results: Over the 10-year period, there were 3,203 ITERS overall, representing 1,207 women and 1,996 men trainees. The mean overall score was lower for women than for men trainees: 4.26 (standard deviation [SD], 0.58) for women and 4.30 (SD, 0.60) for men ($P = 0.04$). This difference was driven by anesthesia trainees, in whom the mean overall score was 4.21 for women and 4.37 for men ($P < 0.001$), with men trainees scoring consistently higher across all seven training subdomains. Within surgical, internal medicine, and critical care residents, there were no differences between men and women in the overall score or the scores across any of the seven subdomains. Across all ITERS, women were less likely than men to receive an overall rating of 5 (outstanding) for the ICU rotation (33% women vs. 37% men; odds ratio, 0.83; 95% confidence interval, 0.71–0.96).

(Received in original form April 5, 2021; accepted in final form July 7, 2021)

ATS Scholar Vol 2, Iss 3, pp 442–451, 2021

DOI: 10.34197/ats-scholar.2021-0048OC

Conclusion: Overall, quantitative evaluation scores between women and men trainees in the ICU are relatively similar. Within anesthesia trainees, scores for men were consistently higher across all domains of evaluation, a finding that requires further investigation.

Keywords:
medical education; critical care medicine; gender

Data from several specialties support the concern that women medical trainees are affected by gender bias and inequities during postgraduate training. Bias in assessment is one area of particular concern, as evidence suggests that women trainees may receive different evaluations and feedback than men (1–4). For example, in an assessment of formal narrative comments provided to surgical residents, men received more positive comments overall (3). Men were also more likely to be described using standout words (e.g., excellent, outstanding, superstar, etc.) and had more positive comments about their future career potential (3). Furthermore, an analysis of qualitative feedback provided to emergency medicine residents revealed that stereotypically male traits, such as decisiveness, confidence, and a take-charge attitude, may be favored in high-acuity settings such as the emergency department (1). This study also demonstrated that men trainees received consistent feedback from attending physicians when areas for improvement were identified, whereas for women, feedback differed between evaluators (1). Discrepancies in feedback and evaluation may result in very different training experiences and may influence trainees' confidence, career choices, and trajectories. In the era of competency-based medical education, variability in evaluation may also result in delays to milestone achievement and impact a trainee's progression within their training program (2). Ensuring fair, objective, and unbiased trainee assessment is essential to providing an equitable training experience.

Critical care medicine is a high-acuity specialty that requires physicians to take a leadership role within an interdisciplinary, interprofessional team. In this type of setting, evidence has shown that men may be perceived to be more effective leaders (5). Given that traditionally male traits and behaviors may be highly valued in the critical care environment, we hypothesize that this may result in lower evaluations for women trainees. Our study aimed to determine if there was a difference between men and women trainees in assessment scores for rotations in the intensive care unit (ICU).

METHODS

In this retrospective study, we analyzed quantitative scores on the in-training evaluation reports (ITERs) for trainees rotating through five core academic ICUs (Mount Sinai Hospital, Toronto General site of University Health Network, Toronto

Author Contributions: Conception and design: J.S., S.G., and S.M. Data collection, analysis, or interpretation: J.S., C.A., S.G., D.P., F.M.G., A.K., and S.M. Drafting or revision of the manuscript: J.S., C.A., S.G., D.P., F.M.G., A.K., and S.M.

Correspondence and requests for reprints should be addressed to Jenna Spring, M.D., F.R.C.P.C., Sunnybrook Health Sciences Center, 2075 Bayview Ave, Room D-108, Toronto, ON, M4N 3M5 Canada. E-mail: jenna.spring@mail.utoronto.ca.

Western site of University Health Network, St. Michael's Hospital, and Sunnybrook Health Sciences Centre) at the University of Toronto over a 10-year period from July 2007 to June 2017. These are tertiary or quaternary medical–surgical ICUs with approximately 16–24 beds. This study was approved by the University of Toronto Research Ethics Board in February 2018. The need for informed consent was waived.

ITERs are a summative assessment done at the end of a clinical rotation. They are completed by the education lead for each ICU, with input from all physician faculty who worked with the trainee. The content of the ITERs varied between specialties, and there were some changes to the ITERs within each specialty over the 10-year study period. However, a 5-point evaluation scale was used throughout. The trainee scores correspond to the following: 1 = unsatisfactory, 2 = needs improvement, 3 = meets expectations, 4 = exceeds expectations, and 5 = outstanding. Trainees must receive a score of 3 or higher to pass the rotation without further review. To account for the variability between ITERs, we recorded the overall rating for the ICU rotation, and we generated a mean score for the components of each CanMEDS subdomain for each individual ITER. The academic year that the rotation took place and the resident's year of training were also recorded.

The study included ITERs for two groups of residents: 1) residents from internal medicine in Postgraduate Years 1–3, anesthesia, general surgery, and surgical subspecialties (neurosurgery, orthopedic surgery, urology, vascular surgery, cardiac surgery, or plastic surgery), who complete mandatory or elective rotations in the ICU as part of their core training program; and 2) critical care subspecialty residents who are training to become intensivists. In Canada, critical care medicine residents have previously completed training in another base specialty (e.g., internal medicine or anesthesiology) and thus have completed at least 3 years of postgraduate training.

Trainee evaluation data was obtained from the Office of Postgraduate Medical Education at the University of Toronto. All ITERs were built on the CanMEDS framework. This is the framework used to evaluate Canadian trainees and includes the following core competencies: medical expert, communicator, collaborator, manager/leader, health advocate, scholar, and professional. We also recorded the overall score for procedural or technical skills for critical care medicine and anesthesia trainees. This information was not explicitly included in the ITER for the other specialties.

Analysis

Descriptive statistics were calculated for all variables of interest. Continuous measures were summarized using means and standard deviations (SDs), whereas categorical measures were summarized using counts and percentages. To compare scores between women and men trainees, a two-sample two-sided *t* test was used for normally distributed scores and a Wilcoxon rank sum test for the case of a nonnormal distribution. To compare trends over time, a linear regression model was run. The model included gender, time, and a gender by time interaction term. To assess gender differences by specialty, a linear regression model was run, which included gender, specialty, and a gender by specialty interaction term. The overall rating score was also assessed as a binary measure by tabulating the proportion of trainees who received a score of 5 versus any other score. This was compared between gender groups using a logistic regression model. Results were reported as odds ratios and their associated 95% confidence intervals (CIs).

Table 1. Number of intensive care unit in-training evaluation reports by specialty between 2007 and 2017

	Critical Care	Internal Medicine	Anesthesia	Surgery	All Programs
Women	186	646	253	122	1,207
Men	438	789	489	280	1,996
Total	624	1,435	742	402	3,203

All analyses were performed using SAS Version 9.4 (SAS Institute).

RESULTS

A total of 3,203 ITERs were included in the analysis. The breakdown by gender and specialty is shown in Table 1. Overall, the proportion of ITERs was 38% for women ($n = 1,207$) and 62% for men ($n = 1,996$). For critical care medicine, 30% ($n = 186$) of ITERs were for women trainees compared with 45% ($n = 646$) in internal medicine, 30% ($n = 122$) in surgery, and 34% ($n = 253$) in anesthesia. For 18 critical care ITERs, an overall score was not recorded (8 women, 10 men). Overall scores were not missing in any other specialty.

Overall Rotation Scores

The mean (SD) overall rating across all specialties was 4.26 (0.58) for women and 4.30 (0.60) for men ($P = 0.04$). For critical care residents, the mean (SD) overall rotation score was 4.33 (0.48) for women and 4.31 (0.59) for men ($P = 0.73$). The mean overall ratings for surgical and internal medicine trainees yielded similar results, with women surgical residents scoring 4.21 (0.67) compared with 4.20 (0.62) for men ($P = 0.85$), and women internal medicine residents scoring 4.27 (0.58) compared with 4.29 (0.60) for men ($P = 0.56$). However, for anesthesia trainees, the mean overall score was significantly higher for men at 4.37 (0.61) versus 4.21 (0.60) for women ($P < 0.001$, Figure 1).

Scores across Subdomains of Training

Looking at the seven CanMEDS subdomains, the only difference in scores across all specialties was that men trainees scored higher than women trainees for medical knowledge (4.21 [0.56] vs. 4.15 [0.54], $P = 0.01$, Table 2). When these scores were broken down by specialty, there were no differences between women and men aside from anesthesia (*see* data supplement). Men anesthesia trainees scored higher than women trainees across every subdomain (Table 2). Men anesthesia trainees also received higher scores than women for procedural skills (4.45 [0.59] vs. 4.32 [0.60], $P = 0.005$), whereas this difference was not seen for critical care trainees (men, 4.22 [0.58]; women, 4.19 [0.60]; $P = 0.58$].

Overall Rotation Scores Over Time

Trends in overall rotation evaluation scores over time are depicted in Figure 2. A significant gender discrepancy over time was demonstrated for anesthesia ($P = 0.008$) but was not present for any other specialty. Within anesthesia, the most significant gap was in 2012–2013, when the mean overall rotation score for women was 4.06 versus 4.56 for men, and in 2013–2014, when women and men scored 4.17 and 4.49, respectively. In both of these years, fewer anesthesia trainees rotating through the ICU were women; women represented 28% ITERs in 2012–2013 and 26% ITERs

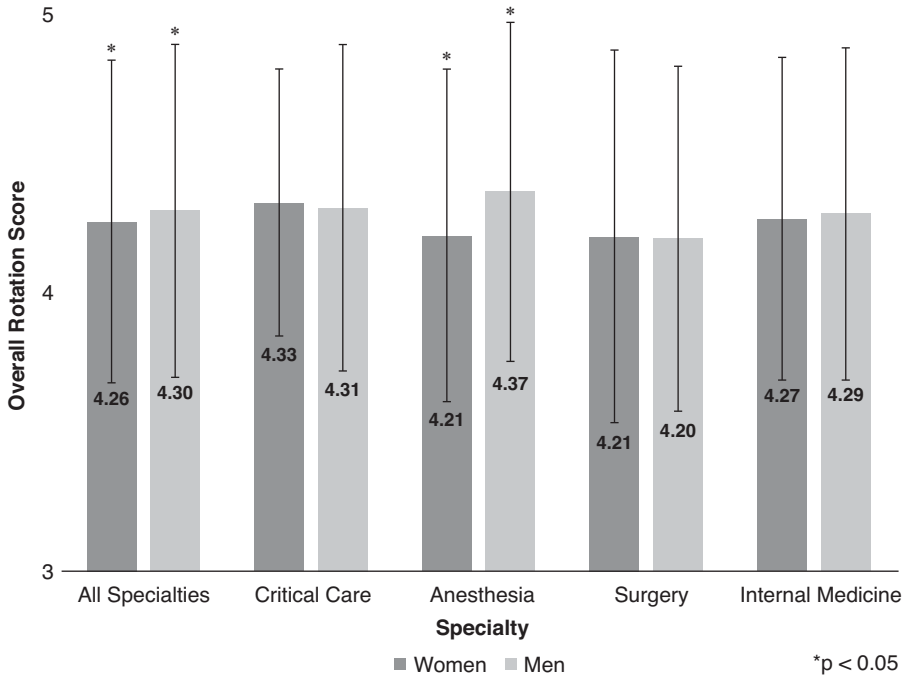

Figure 1. Mean overall in-training evaluation report score by specialty. The *P* value represents the comparison of mean overall rotation scores for women versus men.

in 2013–2014 compared with 34% over the entire 10-year period.

Distribution of Rotation Scores

The overall distribution of scores by gender is shown in Table 3. Women trainees had lower odds of receiving an overall rating of 5 for the rotation (33% women vs. 37% men; OR, 0.83; 95% CI, 0.71–0.96). When broken down by specialty, women anesthesia trainees were less likely to receive an overall score of 5 than men (30% vs. 43%; OR, 0.57; 95% CI, 0.41–0.78), but there were no differences within other specialties. The proportion of trainees receiving a score of 1 or 2 was very low across the entire sample (0.3% for women vs. 0.5% for men, *P* value not calculated).

DISCUSSION

In this retrospective study of trainees rotating through five ICUs at a large academic medical center, women were less likely to receive a rating of "outstanding" for the rotation, and the overall rotation scores were also slightly lower for women. However, when broken down by specialty, scores differed only between men and women anesthesia trainees with no statistically significant differences for critical care, internal medicine, or surgical trainees. Men trainees in anesthesia received higher global evaluation scores and higher scores across all seven core training competencies. We also noted that mean overall evaluation scores were uniformly high across all specialties, with relatively small absolute differences between groups (e.g., less than 0.2 on a score of 5) and an upward trend across the 10-year study period.

Taken as a whole, our results are reassuring. For the majority of trainees rotating through five ICUs, there was no difference in the quantitative evaluation scores between men and women. Within anesthesia trainees, the gender gap in evaluation scores appears to be driven by

Table 2. Comparison of in-training evaluation report scores between men and women for major training competencies

Specialty and Competency Domain	Women	Men	*P* Value
All specialties*			
Medical expert	4.15 (0.54)	4.21 (0.56)	0.01
Communicator	4.35 (0.61)	4.34 (0.60)	0.66
Collaborator	4.38 (0.59)	4.40 (0.59)	0.36
Manager/leader	4.14 (0.56)	4.16 (0.60)	0.18
Health advocate	4.18 (0.59)	4.17 (0.63)	0.74
Scholar	4.23 (0.59)	4.26 (0.61)	0.13
Professional	4.56 (0.58)	4.56 (0.59)	0.93
Procedural skills†	4.27 (0.60)	4.34 (0.59)	0.03
Anesthesia‡			
Medical expert	4.16 (0.56)	4.32 (0.55)	<0.001
Communicator	4.18 (0.65)	4.30 (0.63)	0.01
Collaborator	4.30 (0.60)	4.44 (0.60)	0.003
Manager/leader	4.01 (0.59)	4.12 (0.63)	0.02
Health advocate	4.10 (0.63)	4.23 (0.67)	0.01
Scholar	4.20 (0.58)	4.37 (0.57)	<0.001
Professional	4.47 (0.65)	4.57 (0.60)	0.04
Procedural skills	4.32 (0.60)	4.45 (0.58)	0.005

Data are shown as mean (standard deviation).
Data for critical care medicine, internal medicine, and surgical trainees are presented in the data supplement; there were no differences between women and men in the major competencies within these specialties.
*n = 1,207 for women and n = 1,996 for men.
†A score for procedural skills was only available for critical care medicine and anesthesia trainees.
‡n = 253 for women and n = 489 for men.

the greater likelihood of men receiving an "outstanding" evaluation compared with women. This finding may represent gender bias that may impact residents' training experience and potentially their future careers. Notwithstanding, these quantitative scores represent a gross surrogate of educational interactions between trainees and faculty and do not encapsulate the entire assessment process. Our findings highlight the need to evaluate the entire assessment process to better understand subtle differences that may have a meaningful impact on trainees.

Aside from our study, there are no studies in critical care evaluating gender bias in trainee assessment. However, there is reason for concern for potential gender bias within our specialty. In critical care medicine, women remain underrepresented. This includes the overall proportion of intensivists who are women and the proportion of women who are

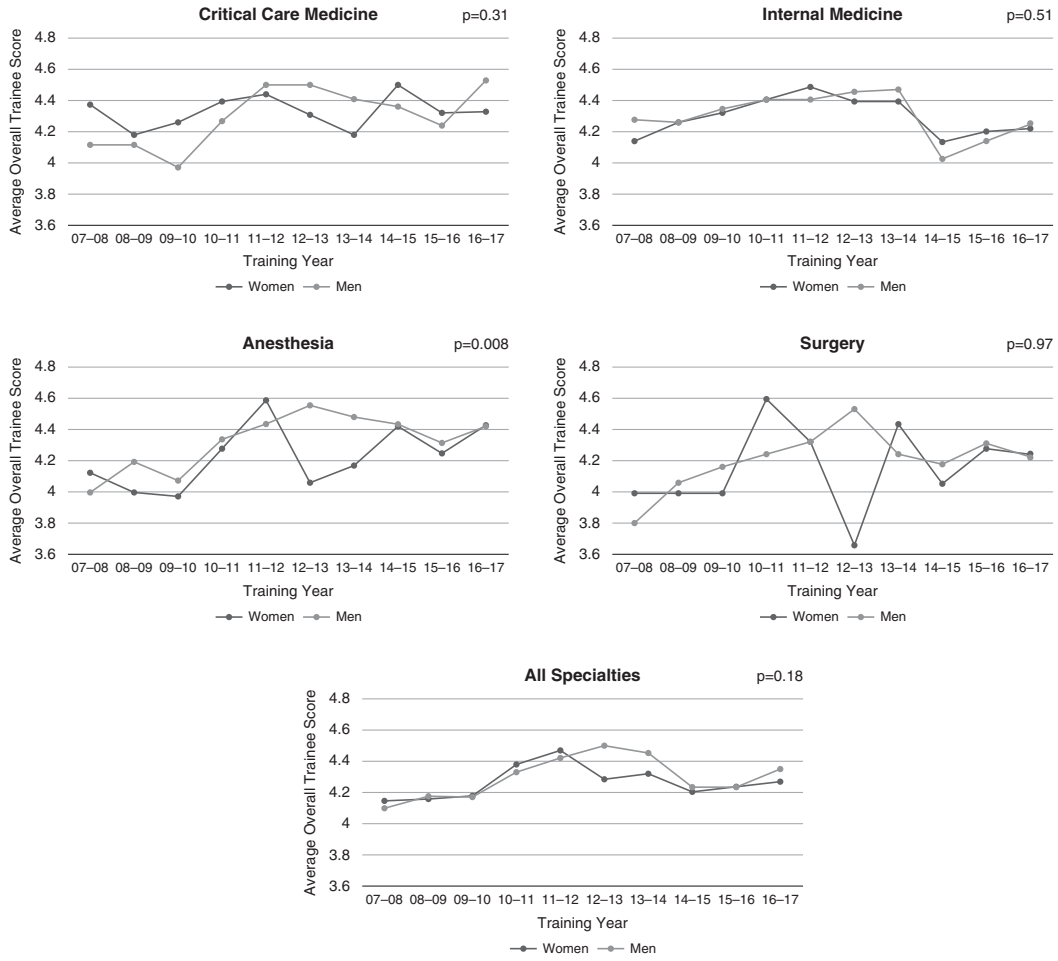

Figure 2. Trends in overall rating over time by specialty. The *P* value represents the comparison in overall rotation scores between women and men over the 10-year study period.

conference speakers, lead authors, and peer reviewers and in leadership positions within critical care societies (6–10). A recent analysis revealed that fewer than one-third of fellows in US critical care training programs are women and that the proportion of women trainees appears to have plateaued over the last decade (11), In a 2019 Canadian qualitative study examining the drivers of gender inequity in critical care medicine, respondents highlighted the inflexible and long work hours, the lack of women role models, and the overt value placed on traditionally masculine traits (12). The latter factor may also have an impact on trainee evaluations.

Studies to date in other specialties have presented conflicting results regarding the effect of gender on assessment in medicine. A systematic review reported evidence of gender bias in postgraduate trainee evaluation in five of the nine studies analyzed (13). Even within a single specialty, investigators have reported discordant findings. Within surgery, a 2020 study found that trainee gender did not impact evaluations of meaningful autonomy in the operating room (14). In contrast, other studies have reported lower autonomy among women surgical trainees (15, 16). Whether gender differences are detected may also depend on the type of

Table 3. Distribution of overall in-training evaluation report scores by gender

Overall Score	1	2	3	4	5
Women	1 (0.1)	3 (0.3)	72 (6)	730 (61)	393 (33)
Men	2 (0.1)	7 (0.4)	121 (6)	1,120 (56)	736 (37)
Total	3 (0.1)	10 (0.3)	193 (6)	1,850 (58)	1,129 (35)

18 ITERs were missing an overall score and were not included in this analysis.
Data are presented as *n* (%).

competencies being assessed as well as the timing of these evaluations. A retrospective study of more than 12,000 evaluations in an internal medicine training program found that men and women trainees received similar overall scores across most domains, with women scoring higher in areas related to interpersonal skills (17). Another recent study evaluated the trajectory of internal medicine evaluations over time and found that evaluation scores for women trainees peaked early in residency and then plateaued, whereas men trainees had sustained improvement over the course of their training (18). Looking at the literature as a whole, discrepancies in ratings of men and women are often reported and may represent subtle bias, but the magnitude and direction of these effects may depend on context—what is being evaluated and by whom.

There are few additional studies of gender bias in the evaluation of anesthesia trainees. However, a recent study of 356 anesthesia trainees in Australia and New Zealand demonstrated that men were significantly more likely to rate their competency above their level of training and exaggerate their procedural experience (19). If also true in our setting, men's confidence, particularly regarding procedures, may be viewed more favorably by evaluators and may lead to more outstanding ratings.

Our study has several strengths. To our knowledge, this is the first study to examine gender discrepancies in evaluation during critical care medicine rotations. We included a large number of evaluations across numerous specialties with trainees at various levels of training, increasing the generalizability of the results. The 10-year study period with data from multiple hospital sites and faculty evaluators also helps to ensure that our results were not related to discrepancies within a specific trainee cohort or at one site.

Our study has limitations. We evaluated quantitative scores and did not assess verbal or written qualitative feedback provided to trainees at the end of the rotation. Qualitative studies looking at trainee evaluations of faculty members have shown evidence of discrepancies based on gender, such as differences in the wording used in free-text evaluations for men and women faculty (20). Similarly, gender differences have also been demonstrated in the comments medical students receive on evaluations (21). We lack information about the gender of the assessor because it is not included in the final ITER, but this may have had an impact on evaluations as well (22). However, each ITER is an aggregate assessment of performance with input from all faculty members who worked with the trainee within that ICU. This method of assessment, with input from multiple faculty members, may reduce gender evaluation bias. The evaluation scores were also uniformly high in this cohort, with most trainees receiving a global rating of 4 on a 5-point scale. This resulted in a relatively

low SD of scores across all specialties, making small discrepancies challenging to interpret. There was an upward trend in overall scores across the study period, which may have reduced the likelihood of detecting a difference between men and women. The "halo effect" and grade inflation are well described in medical education and may be potential underlying contributors to this trend (23–25). Finally, faculty members do not receive standardized training regarding completion of ITERS; this training could potentially reduce bias in evaluations.

The information obtained from this study adds to a growing body of evidence that could have an important impact on trainee education and feedback in the ICU, especially as we transition to a competency-based medical education model. Although these results are reassuring in that the gender gap in rotation evaluation was largely absent across all specialties except for anesthesia, the anesthesia data suggest that implicit bias may play a role in the evaluation of women in certain specialties. This issue necessitates further investigation, and an evaluation of qualitative comments and verbal feedback would be helpful to further explore potential explanations for our findings.

Author disclosures are available with the text of this article at www.atsjournals.org.

REFERENCES

1. 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;9:577–585.
2. 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;177:651–657.
3. Gerull KM, Loe M, Seiler K, McAllister J, Salles A. Assessing gender bias in qualitative evaluations of surgical residents. *Am J Surg* 2019;217:306–313.
4. Choo EK. Damned if you do, damned if you don't: bias in evaluations of female resident physicians. *J Grad Med Educ* 2017;9:586–587.
5. Ju M, van Schaik SM. Effect of professional background and gender on residents' perceptions of leadership. *Acad Med* 2019;94:S42–S47.
6. Canadian Medical Association. Critical care medicine profile. Ottawa, Canada: Canadian Medical Association; 2019 [accessed 2021 March 21]. Available from: https://www.cma.ca/sites/default/files/2020–10/critical-care-e.pdf.
7. Venkatesh B, Mehta S, Angus DC, Finfer S, Machado FR, Marshall J, *et al.* Women in Intensive Care study: a preliminary assessment of international data on female representation in the ICU physician workforce, leadership and academic positions. *Crit Care* 2018;22:211.
8. Mehta S, Rose L, Cook D, Herridge M, Owais S, Metaxa V. The speaker gender gap at critical care conferences. *Crit Care Med* 2018;46:991–996.
9. Vranas KC, Ouyang D, Lin AL, Slatore CG, Sullivan DR, Kerlin MP, *et al.* Gender differences in authorship of critical care literature. *Am J Respir Crit Care Med* 2020;201:840–847.
10. Goldstone K, Edgley C, Mehta S, Leslie K. Peer review for the Canadian Journal of Anesthesia in 2016 and 2017: a retrospective analysis by reviewer and author gender. *Can J Anaesth* 2020;67:336–342.

11. Santhosh L, Babik JM. Diversity in the pulmonary and critical care medicine pipeline: trends in gender, race, and ethnicity among applicants and fellows. *ATS Scholar* 2020;1:152–160.

12. Leigh JP, Grood C, Ahmed SB, Ulrich AC, Fiest KM, Straus SE, *et al.* Toward gender equity in critical care medicine: a qualitative study of perceived drivers, implications, and strategies. *Crit Care Med* 2019; 47:e286–e291.

13. Klein R, Julian KA, Snyder ED, Koch J, Ufere NN, Volerman A, *et al.*; From the Gender Equity in Medicine (GEM) workgroup. Gender bias in resident assessment in graduate medical education: review of the literature. *J Gen Intern Med* 2019;34:712–719.

14. Lane SM, Young KA, Hayek SA, Dove JT, Sharp NE, Shabahang MM, *et al.* Meaningful autonomy in general surgery training: exploring for gender bias. *Am J Surg* 2020;219:240–244.

15. Meyerson SL, Sternbach JM, Zwischenberger JB, Bender EM. The effect of gender on resident autonomy in the operating room. *J Surg Educ* 2017;74:e111–e118.

16. Joh DB, van der Werf B, Watson BJ, French R, Bann S, Dennet E, *et al.* Assessment of autonomy in operative procedures among female and male New Zealand general surgery trainees. *JAMA Surg* 2020; 155:1019–1026.

17. Sulistio MS, Khera A, Squiers K, Sanghavi M, Ayers CR, Weng W, *et al.* Effects of gender in resident evaluations and certifying examination pass rates. *BMC Med Educ* 2019;19:10.

18. Klein R, Ufere NN, Rao SR, Koch J, Volerman A, Snyder ED, *et al.*; Gender Equity in Medicine workgroup. Association of gender with learner assessment in graduate medical education. *JAMA Netw Open* 2020;3:e2010888.

19. Pearce G, Sidhu N, Cavadino A, Shrivathsa A, Seglenieks R. Gender effects in anaesthesia training in Australia and New Zealand. *Br J Anaesth* 2020;124:e70–e76.

20. Heath JK, Weissman GE, Clancy CB, Shou H, Farrar JT, Dine CJ. Assessment of gender-based linguistic differences in physician trainee evaluations of medical faculty using automated text mining. *JAMA Netw Open* 2019;2:e193520.

21. Rojek AE, Khanna R, Yim JWL, Gardner R, Lisker S, Hauer KE, *et al.* Differences in narrative language in evaluations of medical students by gender and under-represented minority status. *J Gen Intern Med* 2019;34:684–691.

22. Rand VE, Hudes ES, Browner WS, Wachter RM, Avins AL. Effect of evaluator and resident gender on the American Board of Internal Medicine evaluation scores. *J Gen Intern Med* 1998;13:670–674.

23. Sherbino J, Norman G. On rating angels: the halo effect and straight line scoring. *J Grad Med Educ* 2017;9:721–723.

24. Schiel KZ, Everard KM. Grade inflation in the family medicine clerkship. *Fam Med* 2019;51:806–810.

25. Fazio SB, Papp KK, Torre DM, Defer TM. Grade inflation in the internal medicine clerkship: a national survey. *Teach Learn Med* 2013;25:71–76.