Gender Effects in Assessment of Clinical Teaching: Does Concordance Matter?

Lynfa Stroud, MD, MEd
Risa Freeman, MD, MEd
Kulamakan Kulasegaram, PhD
Tulin D. Cil, MD, MEd
Shiphra Ginsburg, MD, PhD

ABSTRACT

Background Gender bias is thought to exist in the assessment of clinical teachers, yet its extent in different specialties is not well-documented nor has it been studied at the individual-dyadic level.

Objective The authors sought to determine whether gender bias exists in residents' assessments of faculty teaching in 3 clinical departments, and if present, whether this is influenced by gender concordance or discordance between the faculty and resident.

Methods Residents' ratings of faculty in internal medicine (800 faculty, 5753 ratings), surgery (377, 2249), and family medicine (672, 3438) at the University of Toronto from 2016–2017 were analyzed using the overall global rating on a 5-point scale. A mixed-effects linear regression analysis accounted for nesting of ratings within each faculty member.

Results Overall scores of teaching effectiveness showed a strong skew to favorable ratings for all faculty and a ceiling effect. However, gender effects differed across departments. In internal medicine (38.5% female faculty), no significant gender effects were detected. In surgery (16.2% female) and family medicine (53.0% female), male faculty received significantly higher scores than female faculty. In surgery this was driven by male residents giving male faculty higher ratings (4.46 vs 4.26, $P < .001$). In family medicine this was driven by male faculty receiving higher ratings regardless of resident gender (4.65 to 4.57, $P < .001$).

Conclusions Although effects were very small and inconsistent, with gender concordance mattering only for one department, it suggests that gender is a meaningful source of variance in teaching assessments.

Introduction

Assessments by learners are often the sole measure of teaching effectiveness in medical education[1] and have a significant effect on faculty performance reviews, merit pay, and promotion.[1,2] This is despite evidence in the broader literature that no significant relationship exists between student assessments of teaching effectiveness and actual learning.[3] Such assessments often ask students to rate content and teaching methods they are not qualified to judge and are vulnerable to bias.[1] For example, greater involvement with trainees,[4] charisma and physical attractiveness,[5] extraversion,[6] and even the presence of cookies during a course[7] have been associated with higher teaching effectiveness scores in medicine.

Recently, there has been a greater focus on the effect that gender bias may have on learners' ratings of faculty in medicine. Medical students rated male physicians higher on trustworthiness, competence, and professionalism compared to female physicians,[8] and being male was also a protective factor in perceptions of medical error.[8] During an objective structured clinical examination, internal medicine (IM) residents perceived male faculty providers of feedback to be more credible than female faculty providers.[9] Gender-based linguistic differences exist in how physician trainees describe faculty in their narrative assessments.[10] Morgan et al[11] found that medical students rated female faculty lower than male faculty across surgery, obstetrics, pediatrics, and IM. Fasiotto et al[12] similarly found that residents rated female faculty lower than male faculty across specialties; this was especially pronounced in specialties with low female faculty representation.

Both of these studies[11,12] examined teaching effectiveness ratings of faculty using the overall cohort of teaching assessments, with the latter also controlling for ethnicity, seniority, rank, specialty, and low female representation. However, neither accounted for the gender of the learner in the rating of teaching effectiveness, nor did they look at effects at the dyadic interaction level to determine whether concordance or discordance between learner and faculty influenced ratings. Small but significant effects of concordance of gender and underrepresented minority status on resident ratings of faculty were identified in a study across 18 clinical departments,[13] but were not analyzed within or between departments that may have had very different proportions of male

DOI: http://dx.doi.org/10.4300/JGME-D-20-00145.1

BOX Resident Assessment of Teacher Effectiveness (RATE) Forms

Department of Medicine
This teacher[a]:
1. Made themselves available to me so I had the support I needed
2. Encouraged me to explore my limits safely
3. Provided regular, meaningful, prompt feedback to me
4. Demonstrated respect for me as a learner and as a person
5. Demonstrated respect for others, including patients and team members
6. Stimulated learning as a dedicated and effective teacher
7. Was a good role model as a physician, teacher, and person
8.[b] Overall this teacher had the following impact:

Department of Surgery
This teacher[a]:
1. Made themselves available to me so I had the support I needed
2. Encouraged me to explore my limits safely
3. Provided regular, meaningful, prompt feedback to me
4. Demonstrated respect for me as a learner and as a person
5. Ensured we agreed on expectations early and did their best to meet them
6. Provided effective instruction in the operating room
7. Provided effective instruction in ward/ambulatory setting
8.[b] Overall this teacher had the following impact on me as a learner:

Department of Family & Community Medicine
This teacher[a]:
1. Made themselves available to me so I had the support I needed
2. Encouraged me to explore my limits safely
3. Provided regular, meaningful, prompt feedback to me
4. Demonstrated respect for me as a learner and as a person
5. Ensured we agreed on expectations early and did their best to meet them
6. Provided appropriate and timely supervision, support, and resources
7. Was a good role model as a physician, teacher, and person
8.[b] Overall this teacher had the following impact on me as a learner:

[a] The scale is as follows: (1) Never or very poor (this teacher needs help with this); (2) Occasionally or needs improvement; (3) Frequently and adequately; (4) Usually and skillfully; and (5) Always and exemplary (should be a role model for all teachers).

[b] The scale is as follows: (1) Terrible learning experience; (2) Unpleasant experience; (3) Good experience; (4) Very good experience; and (5) Exceptional experience.

What was known and gap
The relationship of gender concordance with residents' assessments of faculty teaching within different specialties has not been fully examined.

What is new
An analysis to determine if residents' assessments of clinical teachers within 3 departments at 1 institution contain gender bias, and if present, whether this is influenced by gender concordance or discordance between the faculty and resident.

Limitations
The available data limited gender to a dichotomous variable. Data on academic rank was missing from 1 department.

Bottom line
The effects of gender were very small and generally, but not universally, favored male faculty.

and female faculty and residents, which may lead to important differences in ratings. Therefore, this study aimed to determine if gender bias exists, and whether gender concordance or discordance between faculty and residents has an effect on residents' assessments of faculty teaching effectiveness across 3 clinical departments at 1 university faculty of medicine.

Methods

Setting

We conducted this study across the departments of medicine (DOM), surgery (DOS), and family and community medicine (DFCM) at the University of Toronto in Canada using data from a single academic year (July 2016 to June 2017). We selected these clinical departments for their large numbers of learners and faculty and for their differential in gender composition to enable interdepartmental comparisons. In each department, residents assess faculty teaching across a variety of clinical contexts using the Resident Assessment of Teaching Effectiveness (RATE) form at the end of each rotation (at least one faculty member of the resident's choosing must be evaluated per rotation). The RATE form has 7 items on a 5-point scale and each department has slightly different items on their forms (BOX), but there is commonality in the intent of each item and the same overall global.

Data Collection

We created a database containing all RATE forms for 1 academic year across the 3 departments. In each department, an independent research officer extracted, anonymized, collated, and linked data that included resident and faculty gender, resident relation to department (eg, medicine resident on DOM rotation = "on-service," psychiatry resident on DOM rotation = "off-service"), and faculty academic rank (lecturer, assistant professor, associate professor,

TABLE 1
Characteristics of Faculty, Residents, and Gender Discordance Between Resident and Faculty on Resident Assessments of Teaching Effectiveness (RATE) Forms Across 3 Departments at the University of Toronto (2016–2017)

University Department	No. of Faculty	Female Faculty (%)	No. of Individual Ratings	Ratings by Female Residents (%)	Female Concordant Ratings (%)	Male Concordant Ratings (%)	No. of Residents	Female Residents (%)	Off-Service Residents (%)	RATE Gender Discordance (%)
Medicine	800	38.5	5753	44.6	11.3	44.4	552	46	31.2	44.3
Surgery	377	16.2	2249	30.4	7.0	58.5	585	38	32.1	34.5
Family and Community Medicine	672	53	3438	58.7	33.6	20.9	423	62	NA	45.7

professor). We recognize there is a difference between biological sex and gender, which is how a person feels internally and/or identifies with publicly.[14] We use the term gender in this article as (1) our databases are based on self-report (at time of initial employment with options of male, female, other) and are therefore more representative of the concept of gender, and (2) the term gender is more commonly used in the literature.[8–12,15–17] We included faculty academic rank to adjust for supervisory experience and/or perceptions of hierarchy. Given the longitudinal nature of training in DFCM, the duration of contact between resident and faculty was also coded for this cohort (in 3 categories ranging from minimal, moderate, and extensive contact based on resident interpretation and self-report at time of form completion).

Analysis

We analyzed each department's data separately for 3 reasons: (1) the use and instructions around completing the RATE form varied across the departments; (2) the available covariates were not consistent across departments; and (3) separate analysis allowed a clearer comparison of the effect of gender concordance across departments. The primary outcome for the analysis was the overall global rating. While the average of all the items had greater reliability and range, the overall is the rating used for teaching effectiveness reports and merit decisions in our faculty of medicine. All analyses were replicated with the average of all items on the RATE form as a sensitivity check.

Faculty and resident gender were coded along with resident relation to service, faculty academic rank (where available), and duration of contact (where applicable). The primary analysis of the data was a mixed-effects linear regression using restricted maximum likelihood estimation to account for the nesting of data within each faculty member. Accordingly, we specified a random intercept model for faculty to estimate the variance associated with nesting of scores within each faculty and entered faculty gender and gender pairing with the resident evaluator (ie, gender discord) as fixed effects. Other covariates in the analysis included whether ratings were given by on-service versus off-service residents, faculty rank (where available), and duration of contact (where applicable).

The model building approach was to hierarchically enter potential influencers of teaching assessments, including faculty academic rank, on- vs off-service resident, duration of contact, and faculty gender. Interaction terms between these factors were also entered into the model and retained if significant. Resident gender was also included in some models to better estimate if leniency or severity of rating differed

TABLE 2
Raw Means of Resident Assessments of Teaching Effectiveness (RATE) for Male and Female Faculty by Department at the University of Toronto (2016–2017)

Department	All Faculty (Mean)	*P* Value
Medicine		
Male residents	4.67	$P < .02$
Female residents	4.62	
Surgery		
Male residents	4.50	$P < .001$
Female residents	4.31	
Family & Community Medicine		
Male residents	4.56	$P < .0001$
Female residents	4.45	

significantly between female and male residents. Lastly, gender discord was built into the model. F tests were used to assess the statistical significance of each effect with the threshold set at 0.05. The final model reported for all analyses includes only variables that reached the alpha threshold and adequate model fit, compared to the model that includes only the covariates. All analyses were done in SPSS 23 (IBM Corp, Armonk, NY) and SAS 9.2 (SAS Institute Inc, Cary, NC). The University of Toronto Research Ethics Board approved this study.

Results

A demographic summary appears in TABLE 1. The DFCM had the greatest percentage of female faculty and residents, followed by DOM, then DOS. Off-service residents accounted for approximately one-third of residents on medical and surgical rotations. Gender discordance in RATE scoring was lowest in DOS and highest in DFCM.

Overall, the RATE scores showed a strong skew to favorable ratings for all faculty (TABLE 2) and a ceiling effect. Thus, these scores were reversed and log transformed prior to being converted to Z-scores to minimize skew for purposes of analysis. Because the final coefficients express effects on transformed variables, we report only the significant tests and mean differences on the original scale where applicable. As the available covariates and overall effects of gender were inconsistent across the departments, we present each's RATE score results individually (TABLE 3). Within each department there was significant covariance associated within individual faculty's ratings (DOM [ICC = 10.1%, Wald Z = 9.6, $P < .01$]; DOS [ICC = 8.8%; Wald Z = 5.4, $P < .0001$]; and DFCM [ICC = 16.1%; Wald Z = 7.8, $P < .0001$], respectively). Across all 3 departments, male residents gave higher scores than female residents (TABLE 2).

Department of Medicine

There was no significant overall difference on RATE scores given to male and female faculty. We detected a very small effect of gender discordance whereby male residents rated female faculty higher than male faculty. Female residents rated female and male faculty similarly. Off-service residents gave lower scores to faculty than on-service residents at 4.54 vs 4.72 ($F(1,7431) = 3.64$, $P < .040$). We could not determine an interaction effect with gender, as not all off-service residents' genders were known. There was no significant effect of faculty academic rank on scores.

Department of Surgery

On overall RATE scores, male faculty received higher ratings than female faculty. When resident gender was entered into the model, we detected a significant effect of gender discordance. Further exploration showed that this effect was driven by male residents giving higher scores to male versus female faculty. Female residents rated male and female faculty similarly. Off-

TABLE 3
Adjusted Means for Gender Concordant and Discordant Resident Assessments of Teaching Effectiveness (RATE) for Male and Female Faculty by Department at the University of Toronto (2016–2017)

Department	Male Faculty Mean (SD)	Female Faculty Mean (SD)	F Statistic, *P* Value
Medicine	4.64 (0.82)	4.65 (0.84)	$F(1,594) = 0.17$, $P < .78$
Male residents	4.62	4.65	$F(1,5673) = 5.63$, $P < .02$
Female residents	4.62	4.64	$F(1,5711) = 1.9$, $P < .20$
Surgery	4.46 (0.73)	4.28 (0.78)	$F(1,313) = 13.76$, $P < .001$
Male residents	4.46	4.26	$F(1,2211) = 34.2$, $P < .001$
Female residents	4.41	4.48	$F(1,2202) = 1.2$, $P < .30$
Family & Community Medicine	4.65 (0.65)	4.57 (0.72)	$F(1,505) = 12.5$, $P < .0001$
Male residents	4.68	4.56	$F(1,3304) = 11.45$, $P < .001$
Female residents	4.63	4.57	$F(1,3406) = 12.35$, $P < .0001$

service residents gave lower scores to faculty than on-service residents at 4.19 vs 4.51 ($F(1,2045) = 60.1$, $P < .0001$); however, there was no significant interaction between this factor and the gender of residents or of faculty. Academic rank was unavailable as a covariate for the DOS faculty.

Department of Family and Community Medicine

On overall RATE scores, male faculty received higher ratings than female faculty. In the DFCM, both male and female residents rated male faculty higher than female faculty. Increased duration of contact between faculty and residents was also associated with higher ratings of faculty ($F(2, 3386)=61.2$, $P<.00001$), and there was an effect of faculty rank with assistant professors receiving higher ratings than lecturers or associate/full professors ($F(3,3386) = 48.2$, $P < .001$). Observed gender differences remained significant after controlling for differences in the distribution of male and female faculty across academic rank and duration of contact.

Discussion

In this study of the effect of dyadic gender pairing as a construct irrelevant source of variance in resident teaching assessments of IM, FM, and surgical faculty, we found that gender bias and effects of concordance varied significantly by department. While no overall gender bias was detected in the DOM, in DFCM both male and female residents rated male faculty slightly higher, and in the DOS, gender bias was driven by the concordance of slightly higher ratings of male faculty by male residents. This complex pattern of results suggests that gender bias effects require further exploration.

The lack of gender bias in the DOM contrasts with recent studies that have reported bias against female faculty.[8,9,11,12] However, a closer examination of Morgan et al[11] and Fassiotto et al[12] also demonstrates less gender bias in IM. Morgan et al[11] observed the least discrepancy in scores in IM, and in Fassiotto et al[12] after controlling for academic rank and proportional representation of women in specialty (IM being high), gender differences did not persist in IM ratings. Reasons for why less gender bias was detected in our DOM are unclear. Almost 40% of our faculty are female and we have had longstanding female chairs of medicine and IM program directors, so it is possible that role-modeling and female leadership may create a culture of equality and mitigate bias to some extent.

However, if role-modeling fully explained our results, we would not have expected the significant though albeit small difference in DFCM, as at the time of the study they also had a female chair and greater than 50% of faculty are female. Neither Morgan et al[11] nor Fassiotto et al[12] included family medicine in their studies. In DFCM, the duration of the relationship between resident and faculty was important, with higher ratings associated with a greater duration of interaction regardless of gender. This is not entirely unexpected and may be attributed to greater time to develop a coaching relationship with the benefits that may arise from this[18,19]; conversely, this finding may also reflect another form of bias, the "mere exposure effect," a heuristic by which individuals favor things which are more familiar to them,[20] which has been observed in other studies of faculty teaching effectiveness.[21]

The effect of gender in surgery was heavily influenced by gender concordance, with men giving male faculty higher teaching ratings. However, as only 16% of faculty are female, the 65.5% concordance rate observed between resident and faculty was predominantly male resident and male faculty. Although female surgical residents rated male and female faculty similarly, due to the very small concordance rates of female resident–female surgeon dyads, the numbers may have been insufficient to demonstrate a difference. Morgan et al[11] observed the greatest difference between faculty in surgery; similarly, the effect of gender bias only persisted in the study by Fassiotto et al[12] for specialties with low female representation, which were largely surgical. In our study, the combination of low representation at the faculty level and the high concordance with men may have contributed to the largest difference being observed here. This suggests that female surgical faculty are most vulnerable to gender bias influencing their teaching effectiveness ratings and most susceptible to the potential consequences of low ratings.

Although existing effects were small, our findings suggest that gender is a source of variance in teaching effectiveness scores. It is often on these tiny razor-thin edges that decisions regarding merit are made. Gender bias that disadvantages women has also been observed in many other areas of medicine.[22–24] Efforts have recently been directed to better understand gender bias toward women in academic medicine[17] and to increase awareness of and decrease bias toward women.[15,16] However, while specific interventions may diminish gender bias, they do not eliminate it.[15] The phenomena of implicit bias as it affects gender is pervasive and hard to address.

The inability to erase bias entirely suggests that additional approaches should be taken to mitigate their potentially harmful influence. Teaching assessments are often one of the main determinants of teaching awards, academic rewards, and promotion.[1,2] Although effects of gender may be small, differences can appear large, given the ceiling effect on ratings and very narrow standard deviations around means.

Additionally, gender is not the only bias that influences teaching assessments: attractiveness, charisma,[5] extraversion,[6] and race[13,25] are just a few other examples that may unfairly disadvantage some teachers, both male and female. There have recently been calls to reduce such heavy reliance on learner ratings as the only means to rate teaching effectiveness,[26] and instead to include multiple data sources beyond just student assessments.[1] Other sources might include peer assessment, external expert input, learning outcomes, portfolios,[1] and a greater inclusion of narrative assessment.[27] None of these alone are a perfect measure of teaching effectiveness, but using collective input from multiple sources should improve the rigor and validity of teaching effectiveness measurement.

Our study is limited by the use of a dichotomous variable for gender, as this was the format of data available to us; we recognize this is no longer acceptable. We were missing data on academic rank for the DOS so could not examine the effect of seniority in this department. Our rating scale used 5 points, and scores were positively skewed. It is possible with a wider response range (eg, a 10-point scale) we may have detected more subtle differences in ratings, as previously documented in a recent study of gender bias.[28] Additionally, residents' ratings might have differed among the specialties due to other, not accounted for, reasons. Our study included 3 academic departments over 1 year, which may limit its generalizability to other settings. Finally, our analysis identified correlations but cannot determine causation.

Deeper exploration to understand the effects of bias, including gender, on assessments in education is an area for further study, particularly when somewhat surprising effects are observed as in the DFCM in our study (with a predominance of both female faculty and residents). Additionally, given the limitations to eliminating biases, attention should be turned to improving measurements of teaching that combine multiple methods and sources to both provide meaningful feedback to faculty and to inform robust decisions about teaching awards, merit, and career advancement.

Conclusions

In this study examining the relationship of gender concordance with residents' assessments of faculty teaching in IM, FM, and surgery at one institution, the effects of gender were very small and generally, but not universally, favored male faculty. Gender concordance only mattered for the DOS, whereby male surgical residents rated male faculty; in the DFCM both male and female residents rated male faculty more favorably.

References

1. Berk RA. Top five flashpoints in the assessment of teaching effectiveness. *Med Teach*. 2013;35(1):15–26. doi:10.3109/0142159X.2012.732247.
2. Snell L, Tallett S, Haist S, Hays R, Norcini J, Prince K, et al. A review of the evaluation of clinical teaching: new perspectives and challenges. *Med Educ*. 2000;34(10):862–870. doi:10.1046/j.1365-2923.2000.00754.x.
3. Uttl B, White CA, Wong Gonzalez DW. Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. *Studies Educ Eval*. 2017;54:22–42. doi:10.1016/j.stueduc.2016.08.007.
4. Irby DM, Gillmore GM, Ramsey PG. Factors affecting ratings of clinical teachers by medical students and residents. *J Med Educ*. 1987;62(1):1–7. doi:10.1097/00001888-198701000-00001.
5. Rannelli L, Coderre S, Paget M, Woloschuk W, Wright B, McLaughlin K. How do medical students form impressions of the effectiveness of classroom teachers? *Med Educ*. 2014;48(8):831–837. doi:10.1111/medu.12420.
6. Scheepers RA, Lombars KMJMH, van Aken MAG, Heineman MJ, Arah OA. Personality traits affect teaching performance of attending physicians: results of a multi-centre observational study. *PLoS One*. 2014;9(5):e98107. doi:10.1371/journal.pone.0098107.
7. Hessler M, Popping DM, Hollstein H, Ohlenburg H, Arnemann PH, Massoth C, et al. Availability of cookies during an academic course session affects evaluation of teaching. *Med Educ*. 2018;52(10):1064–1072. doi:10.1111/medu.13627.
8. Ladha M, Bharwani A, McLaughlin K, Stelfox HT, Bass A. The effects of white coats and gender on medical students' perceptions of physicians. *BMC Med Educ*. 2017;17(1):93. doi:10.1186/s12909-017-0932-1.
9. Stroud L, Sibbald M, Richardson D, McDonald-Blumer H, Cavalcanti R. Feedback credibility in a formative postgraduate OSCE: effects of examiner type. *J Grad Med Educ*. 2018;10(2):185–191. doi:10.4300/JGME-D-17-00578.1.
10. Heath KJ, Weissman GE, Clancy CB, Haochang S, Farrar JT, Dine J. Assessments of gender-based linguistic differences in physician trainees evaluations of medical faculty using automated text mining. *JAMA Newt Open*. 2019;2(5):e193520. doi:10.1001/jamanetworkopen.2019.3520.
11. Morgan HK, Purkiss JA, Porter AC, Lypson ML, Santen SA, Christner JG, et al. Student evaluations of faculty physicians: gender differences in teaching evaluations. *J Womens Health (Larchmt)*. 2016;25(5):453–456. doi:10.1089/jwh.2015.5475.
12. Fassiotto M, Li J, Maldonado Y, Kothary N. Female surgeons as counter stereotype: the impact of gender perceptions on trainee evaluations of physician faculty.

J Surg Educ. 2018;75(5):1140–1148. doi:10.1016/j.jsurg.2018.01.011.
13. McOwen KS, Bellini LM, Guerra CE, Shea JA. Evaluation of clinical faculty: gender and minority implications. *Acad Med*. 2007;82(10 suppl):94–96. doi:10.1097/ACM.0b013e3181405a10.
14. Statistics Canada. Gender of Person. http://www23.statcan.gc.ca/imdb/p3Var.pl?Function=DEC&Id= 410445. Accessed October 29, 2020.
15. Girod S, Fassiotto M, Grewal D, Candy Ku M, Sriram N, Nosek BA, et al. Reducing implicit gender leadership bias in academic medicine with an educational intervention. *Acad Med*. 2016;91(8):1143–1150. doi:10.1097/ACM.0000000000001099.
16. Carnes M, Devine PG, Baier Manwell L, Byars-Winston A, Fine E, Ford CE, et al. The effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized controlled trial. *Acad Med*. 2015;90(2):221–230. doi:10.1097/ACM.0000000000000552.
17. Fassiotto M, Hamel EO, Ku M, Correll S, Grewal D, Lavori P, et al. Women in academic medicine: measuring stereotype threat among junior faculty. *J Womens Health (Larchmt)*. 2016;25(3):292–298. doi:10.1089/jwh.2015.5380.
18. Hauer KE, Iverson N, Quach A, Yuan P, Kaner S, Boscardin C. Fostering medical students' lifelong learning skills with a dashboard, coaching and learning planning. *Perspect Med Educ*. 2018;7(5):311–317. doi:10.1007/s40037-018-0449-2.
19. Harrison CJ, Könings KD, Dannefer EF, Schuwirth LWT, Wass V, van der Vleuten CPM. Factors influencing students' receptivity to formative feedback emerging from different assessment cultures. *Perspect Med Educ*. 2016;5(5):276–284. doi:10.1007/s40037-016-0297-x.
20. Marsh B. Heuristics as social tools. *New Ideas Psych*. 2002;20(1):49–57. doi:10.1016/S0732-118X(01)00012-5.
21. Potisek NM, Page L, Narayan A, McNeal-Trice K, Steiner MJ. The association between pediatric faculty factors and resident physician ratings of teaching effectiveness. *Glob Pediatr Health*. 2019;6:2333794X18822996. doi:10.1177/2333794X18822996.
22. Trix F, Psenka C. Exploring the color of glass: letters of recommendation for female and male medical faculty. *Discourse & Society*. 2003;14(2):191–220. doi:10.1177/0957926503014002277.
23. Magua W, Zhu X, Bhattacharya A, Filut A, Potvien A, Leatherberry R, et al. Are female applicants disadvantaged in National Institute of Health peer review? Combining algorithmic text mining and qualitative methods to detect evaluative differences in R01 reviewers' critiques. *J Womens Health (Larchmt)*. 2017;26(5):560–570. doi:10.1089/jwh.2016.6021.
24. Wong A, McKey C, Baxter P. What's the fuss? Gender and academic leadership. *J Health Organ Manag*. 2018;32(6):779–792. doi:10.1108/JHOM-02-2018-0061.
25. Merrit DJ. Bias, the brain, and student evaluations of teaching. *St. John's Law Rev*. https://scholarship.law.stjohns.edu/lawreview/vol82/iss1/6. Accessed October 29, 2020.
26. Ontario Confederation of University Faculty Associations. Report of the OCUFA Student Questionnaires on Courses and Teaching Working Group. https://ocufa.on.ca/assets/OCUFA-SQCT-Report.pdf%22. Accessed October 12, 2020.
27. Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: a quantitative reliability analysis of qualitative data. *Acad Med*. 2017;92(11):1617–1621. doi:10.1097/ACM.0000000000001669.
28. Rivera LA, Tilcsikb A. Scaling down inequality: rating scales, gender bias, and the architecture of evaluation. *Am Sociol Rev*. 2019;84(2):248–274. doi:10.1177/0003122419833601.

All authors are with the University of Toronto, Toronto, Ontario, Canada. **Lynfa Stroud, MD, MEd,** is Associate Professor, Department of Medicine, and Centre Researcher, Wilson Centre for Education; **Risa Freeman, MD, MEd,** is Associate Professor and Vice-Chair Education, Department of Family and Community Medicine; **Kulamakan Kulasegaram, PhD,** is Assistant Professor, Department of Family and Community Medicine; and Scientist, Wilson Centre for Education; **Tulin D. Cil, MD, MEd,** is Associate Professor, Department of Surgery, and Centre Researcher, Wilson Centre for Education; **Shiphra Ginsburg, MD, PhD,** is Professor, Department of Medicine, Canada Research Chair in Health Professions Education, and Scientist, Wilson Centre for Education.

Funding: The authors report no external funding sources for this study.

The authors would like to thank Dr. Sharon Strauss, Vice-Chair Mentorship, Equity, and Diversity Department of Medicine; Dr. Stuart Murdoch, Program Director Department Family and Community Medicine; and Dr. Najma Ahmed, Vice-Chair Education Department of Surgery for their support of the project and access to data. The authors would also like to thank information officers Mr. Edmund Lorens, Department of Medicine; Mr. Haitao Zhang, Department of Family and Community Medicine; and Ms. Tess Weber, Department of Surgery for their assistance with extracting and compiling the data.

Conflict of interest: The authors declare they have no competing interests.

Peer-reviewed oral abstract presentation at the Canadian Conference on Medical Education, Niagara Falls, Ontario, Canada, April 13–16, 2019; Society of Teachers of Family Medicine Annual Spring Conference, Toronto, Ontario, Canada, April 27–May 1, 2019; Surgical Education Week, Chicago, Illinois, April 23–27, 2019; and International Conference on Residency Education, Ottawa, Ontario, Canada, September 26–28, 2019.

Corresponding author: Lynfa Stroud, MD, MEd, Sunnybrook HSC, Rm C412, 2075 Bayview Ave, Toronto, ON M4N 3M5, Canada, 416.480.6100, ext 83627, fax 416.480.6191, lynfa.stroud@sunnybrook.ca

Received February 20, 2020; revisions received August 19, 2020, and September 1, 2020; accepted September 16, 2020.