Gender Bias in Simulation-Based Assessments of Emergency Medicine Residents

Jeffrey N. Siegelman, MD
Michelle Lall, MD, MHS
Lindsay Lee, MPH
Tim P. Moran, PhD
Joshua Wallenstein, MD
Bijal Shah, MD

ABSTRACT

Background Gender-related disparities persist in medicine and medical education. Prior work has found differences in medical education assessments based on gender.

Objective We hypothesized that gender bias would be mitigated in a simulation-based assessment.

Methods We conducted a retrospective cohort study of emergency medicine residents at a single, urban residency program. Beginning in spring 2013, residents participated in mandatory individual simulation assessments. Twelve simulated cases were included in this study. Rating forms mapped milestone language to specific observable behaviors. A Bayesian regression was used to evaluate the effect of resident and rater gender on assessment scores. Both 95% credible intervals (CrIs) and a Region of Practical Equivalence approach were used to evaluate the results.

Results Participants included 48 faculty raters (25 men [52%]) and 102 residents (47 men [46%]). The difference in scores between male and female residents (M = –0.58, 95% CrI –3.31–2.11), and male and female raters (M = 2.87, 95% CrI –0.43–6.30) was small and 95% CrIs overlapped with 0. The 95% CrI for the interaction between resident and rater gender also overlapped with 0 (M = 0.41, 95% CrI –3.71–4.23).

Conclusions In a scripted and controlled system of assessments, there were no differences in scores due to resident or rater gender.

Introduction

Despite the adoption of Title IX of the Education Amendments of 1972 nearly 45 years ago, gender disparities still exist in medicine, including compensation,[1–3] academic rank, retention,[4] and leadership positions.[5–7] In 2013–2014, women accounted for 47% of US medical students and one-third of all full-time academic physicians. Emergency medicine (EM) was 1 of the top 10 specialties for women entering residency, with 38% of EM residency positions being filled by women. However, EM is among the specialties with the lowest percentage of department chairs who are women (10%).[6]

The medical literature on the presence and impact of gender biases in medical education, explicit or implicit, is sparse. A recent study that investigated for differences in medical student evaluations of male and female faculty physicians on 4 required clinical rotations (obstetrics and gynecology, pediatrics, surgery, and internal medicine) found female faculty physicians received lower scores on the evaluation item "overall quality of teaching" in all 4 rotations.[8] No differences were found in faculty evaluations based on medical student gender.[8] The findings suggest the transient relationships between medical students and faculty physicians may be subject to unconscious gender bias. These findings are in contrast to a large study of resident evaluations of faculty, which showed no overall gender difference in mean ratings,[9] although there was a significant interaction effect with female faculty rated highest by female residents, and male faculty rated highest by male residents. Differential performance by gender has also been shown in medical students on a clerkship rotation[10] and on a high-stakes procedural simulation.[11]

There is a paucity of studies on gender differences in milestone assessment. One recent large multi-site cohort study of EM residents evaluated bias in end-of-shift evaluations and found a significant gender bias based on resident gender.[12] This may be because shift evaluations usually represent subjective assessments. Residents are evaluated only on the cases they saw during a particular shift, resulting in considerable variation with respect to which competencies were assessed across residents and rated by faculty.

Simulation allows for a more structured, consistent evaluation environment in which residents can be tested on identical clinical problems, and in which specific competencies can be assessed. We hypothesized

DOI: http://dx.doi.org/10.4300/JGME-D-18-00059.1

Editor's Note: The online version of this article contains a figure showing traces and histograms of parameter estimates, further statistical analyses details, and secondary analyses examining validity of the simulation cases.

that simulation, being a more objective assessment tool, may mitigate gender disparities in resident assessment.

Methods

We conducted a retrospective cohort study of EM residents at a single urban residency program with 21 residents per class.

Beginning in spring 2013, residents participated in mandatory semiannual individual simulation assessments. Each assessment included 2 cases as well as a debriefing session. Data for this study consist of testing scores from fall 2013 through spring 2016 and include residents from 5 class years. Subcompetencies to be tested were defined prior to case development (4 to 5 subcompetencies per case, presented in TABLE 1). Cases were developed by a group of simulation faculty de novo, or adapted from prior published cases. Content represented the breadth of the EM curriculum. Three cases, with validity evidence and example assessment tools,[13] were developed as part of a national collaboration that included our institution. For cases that were developed locally, a similar case and rating form development process was used.

Critical actions were developed that mapped milestone language to specific observable behaviors with binary responses. Cases were reviewed for content validity by topical experts and tested prior to implementation. Standardized patients, nurses, and simulation operators were trained through the institution's clinical skills center and by pilot testing the case. Cases were adjusted prior to the first assessment to ensure standardization and appropriate focus on the specific behaviors of interest. Faculty raters (board-certified or board-eligible EM physicians) received general information about the assessments, and were provided the case and tool approximately 1 week ahead of the simulation. On the day of the assessments, raters received verbal training on the use of the form with further instruction on how to grade specific items on the form itself. Residents were tested over a 4-day period in 2-hour blocks, and were asked to keep cases confidential. Resident and rater gender were assigned in a binary fashion based on internal department records.

The study was reviewed by the Emory University School of Medicine's Institutional Review Board and determined to be exempt.

Residents and rater demographic variables were described using means and standard deviations for continuous variables and frequencies/percentages for categorical variables. The primary aim of the study was to determine whether simulation assessment scores resulted in equivalent scores for male and female residents. Because Bayesian statistical methods are better suited to provide evidence for equivalence than standard null hypothesis significance tests and *P* values,[14] we examined the effect of gender on simulation scores using a Bayesian mixed model regression. A mixed model was used to account for clustering within the data. Each resident completed multiple assessments and raters evaluated multiple residents.

What was known and gap
There are gender-related disparities in medicine and medical education, with prior work having found gender-based differences in assessments of learners.

What is new
A retrospective cohort study of emergency medicine residents at a single program assessed for gender-based (faculty and trainee) differences in resident ratings on a series of mandatory simulation cases.

Limitations
Single specialty, single institution study limits generalizability.

Bottom line
This study of assessment based on simulated cases did not find practically significant differences in assessments by resident or rater gender.

Our decision rule used a region of practical equivalence (ROPE),[14–16] representing the largest difference between male and female residents' scores that would be considered unimportant for practical purposes. If the 95% credible interval (CrI)—the Bayesian analog of the confidence interval—falls entirely within the ROPE, the 95% most credible values for the difference between male and female residents' scores are practically unimportant. The ROPE was determined by expert consensus of board-certified EM physicians without input by a statistician and before the results of the regression were known, and set at 0 ± 5. Statistical analyses were conducted using R version 3.4.1 (The R Foundation for Statistical Computing, Vienna, Austria) and MCMCglmm.[17] Details of the analysis are presented in the online supplemental material.

Results

Twelve cases were included in this analysis, with 48 faculty and 102 residents participating over the 3-year study period. Two or fewer residents were missing from each examination due to vacation or personal leave. Resident and rater demographics are presented in TABLE 2. Overall, the mean score (percentage of total checklist items observed) on the simulation assessments was 65.43 (95% CrI 63.27–67.49) with

TABLE 1
Subcompetencies Assessed by Case

Topic	PC-1	PC-2	PC-3	PC-5	PC-6	PC-9	PC-10	PC-11	PC-13	ICS-1	ICS-2	SBP-1
ID	x			x	x							x
GYN	x	x	x	x						x	x	
ENT	x	x	x	x		x	x					
Procedures	x	x		x		x		x			x	
Peds	x	x	x	x			x			x		
Tox	x	x	x	x			x					x
Neuro	x	x	x	x						x		
Trauma	x			x		x	x		x		x	
CV	x	x	x			x					x	
ID	x	x	x							x		
Peds	x									x		x
Trauma	x	x					x				x	

Abbreviations: PC, patient care; ICS, interprofessional and communication skills; SBP, systems-based practices; ID, infectious disease; GYN, gynecology; ENT, otolaryngology; Peds, pediatrics; Tox, toxicology; Neuro, neurology; CV, cardiovascular.

male (M = 64.60, 95% CrI 62.12–66.76) and female (M = 64.95, 95% CrI 62.63–67.67) residents obtaining similar scores.

Results of the regression analysis are presented in TABLE 3 and in the online supplemental figure. The main effects of "resident gender" and "resident gender by rater gender interaction" were nonsignificant, and the 95% CrIs were completely contained within the ROPE. The main effect of "rater gender," though nonsignificant, was larger, and the 95% CrI extended beyond the ROPE, indicating that female raters may tend to rate residents more favorably.

TABLE 2
Demographic Characteristics of Residents and Raters

Characteristic	Male	Female
Residents		
Age, M (SD)	30.05 (2.72)	29.67 (2.23)
Gender, n (%)	47 (46.08)	55 (53.92)
Race, n (%)		
Asian	11 (10.78)	16 (15.68)
Black	8 (7.84)	11 (10.78)
Hispanic	1 (0.98)	2 (1.96)
White	27 (26.47)	27 (26.47)
Raters		
Age, M (SD)	40.26 (7.56)	39.21 (6.77)
Gender, n (%)	25 (52.08)	23 (47.92)
Race, n (%)		
Asian	2 (4.17)	6 (12.5)
Black	3 (6.25)	4 (8.33)
Hispanic	0 (0)	2 (4.17)
White	20 (41.67)	11 (22.92)

Abbreviation: M, mean.

Discussion

In this retrospective analysis of resident scores from 12 simulation assessments at a single program, there was no main effect of "resident gender" and no interaction between "resident and rater gender." Additionally, the 95% CrIs were entirely contained within the ROPE, which was determined a priori. This finding indicates a high degree of probability that any gender differences in simulation assessment scoring are small and likely not practically significant. On average, female raters tended to rate residents, regardless of gender, 2.9% higher than their male colleagues. While nonsignificant, the 95% CrI extended beyond the ROPE, thereby providing weaker evidence for equivalence.

Much of the prior literature examined medical student evaluations both in the clinical environment and in structured clinical assessments. Most of these studies have demonstrated a gender bias: clerkship grades differed by gender,[10,18] empathy assessments by standardized patients in an EM clerkship assessment favored women,[19] and a surgical study found superior performance on laparoscopic trainers by men.[11] These studies relied on global clerkship ratings, procedural simulation ratings,

TABLE 3
Results of the Bayesian Regression

Predictor	Posterior Slope	95% Credible Interval
Resident gender	–0.58	–3.31–2.11
Rater gender	2.87	–0.48–6.30
Resident X rater gender interaction	0.41	–3.71–4.23

and empathy tools. A recent study of milestone-based, end-of-shift evaluations of EM residents showed a gender disparity,[12] with men advancing to higher levels of competency more quickly than women.

To our knowledge, this is the first study assessing the impact of gender bias on resident assessments in a simulated environment. Our results support our hypothesis that simulation may provide a more objective assessment environment possibly due to carefully standardized scenarios and the use of objective binary behavior-based assessment tools. Given the artificial environment of simulation, these assessments are part of a larger portfolio of assessment tools for the Clinical Competency Committee and residency program directors to consider when assigning semiannual competency ratings.

Our study has several limitations. It was conducted in a single institution, and in a department that conducts significant education on issues of diversity and inclusion, including implicit bias. The results may not generalize to other settings and should be replicated in a large multi-center study. Gender was categorized in a binary fashion, and assigned by the research team based on internal records, which may not reflect the full spectrum of gender identity. Due to the large number of unplanned comparisons that it would entail, we did not evaluate whether individual cases/subcompetencies were associated with greater or lesser gender bias.

Finally, not all cases used had a thorough psychometric evaluation. Three cases had high interrater reliability and increased as residents progressed through their residency program,[15] while the other 9 cases did not undergo this assessment. Future research will need to examine the reliability and factor structure of simulation assessments. Research also should examine whether cases characterized by more concrete and easily operationalized behaviors would be associated with less bias. Future work comparing individual cases will help determine which cases provide the fairest assessment.

Conclusion

In a scripted and controlled assessment environment such as simulation, we demonstrated that scores did not differ as a function of resident or rater gender, and that there was no interaction between resident and rater gender in this EM residency study. Our findings suggest that simulation assessment may represent a less biased method for evaluating resident competency.

References

1. Jagsi R, Griffith KA, Stewart A, et al. Gender differences in the salaries of physician researchers. *JAMA*. 2012;307(22):2410–2417.
2. Sege R, Nykiel-Bub L, Selk S. Sex differences in institutional support for junior biomedical researchers. *JAMA*. 2015;314(11):1175–1177.
3. Lo Sasso AT, Richards MR, Chou CF, et al. The $16,819 pay gap for newly trained physicians: the unexplained trend of men earning more than women. *Health Aff (Project Hope)*. 2011;30(2):193–201.
4. Jena AB, Khullar D, Ho O, et al. Sex differences in academic rank in US medical schools in 2014. *JAMA*. 2015;314(11):1149–1158.
5. Conrad P, Carr P, Knight S, et al. Hierarchy as a barrier to advancement for women in academic medicine. *J Womens Health (Larchmt)*. 2010;19(4):799–805.
6. Lautenberger DM, Dander V, Raezer CL, et al. The State of Women in Academic Medicine: The Pipeline and Pathways to Leadership. 2013–2014. https://members.aamc.org/eweb/upload/The%20State%20of%20Women%20in%20Academic%20Medicine%202013-2014%20FINAL.pdf. Accessed July 6, 2018.
7. Carr PL, Gunn CM, Kaplan SA, et al. Inadequate progress for women in academic medicine: findings from the National Faculty Study. *J Womens Health (Larchmt)*. 2015;24(3):190–199.
8. Morgan HK, Purkiss JA, Porter AC, et al. Student evaluation of faculty physicians: gender differences in teaching evaluations. *J Womens Health (Larchmt)*. 2016;25(5):453–456.
9. McOwen KS, Bellini LM, Guerra CE, et al. Evaluation of clinical faculty: gender and minority implications. *Acad Med*. 2007;82(suppl 10):94–96.
10. Bienstock JL, Martin S, Tzou W, et al. Medical students' gender is a predictor of success in the obstetrics and gynecology basic clerkship. *Teach Learn Med*. 2002;14(4):240–243.
11. Thorson CM, Kelly JP, Forse RA, et al. Can we continue to ignore gender differences in performance on simulation trainers? *J Laparoendosc Adv Surg Tech A*. 2011;21(4):329–333.
12. Dayal A, O'Connor DM, Qadri U, et al. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. *JAMA Intern Med*. 2017;177(5):651–657.
13. Hart D, Bond W, Siegelman JN, et al. Simulation for assessment of milestones in emergency medicine residents. *Acad Emerg Med*. 2018;25(2):205–220.
14. Kruschke J. *Doing Bayesian Data Analysis: A Tutorial Using R, JAGS, and STAN*. 2nd ed. New York, NY: Academic Press; 2014.
15. Kruschke JK, Liddell TM. The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and

power analysis from a Bayesian perspective. *Psychon Bull Rev.* 2018;25(1):178–206.
16. Lesaffre E. Superiority, equivalence, and non-inferiority trials. *Bull NYU Hosp Jt Dis*. 2008;66(2):150–154.
17. Hadfield JD. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R Package. *J Stat Softw.* 2010;33(2):1–22.
18. Burgos CM, Josephson A. Gender differences in the learning and teaching of surgery: a literature review. *Int J Med Educ*. 2014;5:110–124.
19. Berg K, Blatt B, Lopreiato J, et al. Standardized patient assessment of medical student empathy: ethnicity and gender effects in a multi-institutional study. *Acad Med*. 2015;90(1):105–111.

All authors are with Department of Emergency Medicine, Emory University. **Jeffrey N. Siegelman, MD,** is Assistant Professor and Associate Residency Director; **Michelle Lall, MD, MHS,** is Assistant Professor and Associate Residency Director; at the time of the study, **Lindsay Lee, MPH,** was an MPH Student, Emory University School of Public Health, and is now a Medical Student, School of Medicine; **Tim P. Moran, PhD,** is Associate Research Scientist; **Joshua Wallenstein, MD,** is Associate Professor; and **Bijal Shah, MD,** is Associate Professor and Chair, Clinical Competency Committee.

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

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

This work was previously presented at the Southeastern Regional Meeting of the Society for Academic Emergency Medicine, Jacksonville, Florida, February 10–11, 2017; Council of Emergency Medicine Residency Directors Academic Assembly, Fort Lauderdale, Florida, April 27–30, 2017; and Society for Academic Emergency Medicine Annual Meeting, Orlando, Florida, May 16–19, 2017.

The authors would like to thank the staff of the Emory Clinical Skills and Simulation Centers, especially to Kim Fugate, Deb Laubscher, and Reggie Adams for their leadership and skill, and to Jess Bowling for technical assistance. It is only through their organizational prowess and technical know-how that these assessments are possible.

Corresponding author: Jeffrey N. Siegelman, MD, Emory University, Department of Emergency Medicine, 49 Jesse Hill Jr Drive SE, Atlanta, GA 30303, 404.251.8868, fax 404.778.2630, jsiegelman@emory.edu

Received January 18, 2018; revision received April 6, 2018; accepted May 15, 2018.