A Novel Method for Blinding Reviewers to Gender of Proceduralists for the Purposes of Gender Bias Research

Urgent message: Gender bias, whether overt or subconscious, may be to blame for disparities in hiring practices, salary, and advancement in medical schools, the urgent care setting, and any healthcare workplace. Recognizing the value of gender-neutral assessment may not only "even the playing field," but increase the likelihood of identifying the best candidates for clinical positions.

MICHAEL PALLACI, DO; JENNIFER BECK-ESMAY, MD; ADAM R. ALUISIO, MD, MSC; MICHAEL B. WEINSTOCK, MD; ALLEN FRYE, NP; ASHLEY SEE, DO; AND JEFF RIDDELL, MD

Abstract

Introduction
Gender disparity has been demonstrated on Accreditation Council for Graduate Medical Education (ACGME) milestone evaluations, with the largest differences in procedural competency. There are currently no validated methods by which researchers can blind reviewers to gender to evaluate for bias in procedural evaluations.

Objective
Our objective was to determine if a novel video-based evaluation method could blind evaluators to the gender of trainees performing simulated procedures, thereby facilitating more objective assessment of gender disparity.

Methods
After removing all jewelry from their hands, proceduralists were gowned, double-gloved, and filmed by a professional videographer while performing simulated procedures. Only their double-gloved hands, gowned forearms and lower torsos, and the procedural field were visible in the videos. Five residents (two male and three female) performed three procedures each (lumbar puncture, chest tube thoracostomy and central venous catheter placement), yielding 15 videos. Seven graduate medical educators with experience evaluating residents watched 30-45 second video clips and evaluated the perceived gender of the proceduralist on a Likert-type scale (1=definitely male, 3=likely male, 5=can't tell, 7=likely female, 9=definitely female). A response concordant with proceduralist gender with a confidence level of likely or higher (1-3 for males or 7-9 for females) was considered correct gender identification. Responses discordant with proceduralist gender or in the "can't tell" category, were considered incorrect.

Author affiliations: **Michael Pallaci, DO,** Northeast Ohio Medical University; Ohio University Heritage College of Osteopathic Medicine. **Jennifer Beck-Esmay, MD,** Mount Sinai Morningside – Mount Sinai West. **Adam R. Aluisio, MD,** The Warren Alpert Medical School of Brown University. **Michael B. Weinstock, MD,** Adena Health System Emergency Medicine Residency Program; Wexner Medical Center at The Ohio State University. **Allen Frye, NP,** Adena Health System. **Ashley See, DO,** Adena Health System Emergency Medicine Residency Program. **Jeff Riddell, MD,** Keck School of Medicine of the University of Southern California.

Reviewer scores were summarized, and one-sample proportion tests were used to assess for significant differences in correctly identifying the proceduralist gender assuming a null hypothesis of >50% correct gender identification (ie, greater than chance in the binary categorization).

Results

Of 105 total responses, 56 (53.3%) expressed confidence in the gender of the proceduralist (1-3 or 7-9), and 62.5% of those assessments (35/56) were accurate. Across all reviewers and procedures, the proceduralist's gender was correctly identified in 33.3% (95% CI: 25.1% to 42.8%) of videos. This proportion was not statistically significant compared to the null of >50% correct gender identification (p=1.00).

Conclusion

The method used was effective in blinding reviewers to the gender of the proceduralist, and represents an innovative approach to facilitate research into gender bias in procedural evaluations.

Introduction

Gender disparities exist throughout medicine.[1-3] Jena and colleagues have demonstrated significant differences in both salary[1] and academic rank[2] for women in U.S. medical schools, even after adjusting for age, experience, specialty, and productivity. While women now outnumber men in medical school classes, just over one-third of emergency medicine residents are female,[4] and women remain significantly underrepresented as faculty in academic medicine.[3,5,6] Additionally, there is a significant perception of bias among female physicians; a 2016 survey of over 1,000 academic physicians found that 70% of female physicians perceived gender bias, whereas only 22% of male physicians reported such bias.[7] It has been hypothesized that the greatest attrition in academia for women occurs during residency,[8] which may be potentiated by implicit bias and/or explicit discrimination experienced across all levels of training.

A recent study found an attainment gap between male and female emergency medicine (EM) residents in evaluations of performance on Accreditation Council for Graduate Medical Education (ACGME) milestones, as well as qualitative differences in the kind of feedback that male and female EM residents received from attending physicians.[9] The largest differences occurred in the evaluation of procedural competency.[9,10] A large study examining longitudinal milestone ratings for all EM residents reported to the ACGME found males were rated as performing better than females for four of the 22 subcompetencies at graduation, including three procedural subcompetencies.[11] As such, the procedural subcompetencies are appropriate target domains to study the etiology and significance of these gaps.

Multiple studies have found that delayed video procedural assessment is equivalent to direct observation. Aggarwal, Driscoll, and Dath all demonstrated good inter-rater reliability between scores based on real-time assessment and those based on delayed video review.[12-14] While Hassanpour and colleagues found scores on video review to be significantly lower than on real-time assessment, they found a strong correlation between the two methods (r=0.89) and acceptable inter-rater reliability.[15] The authors also made the point that, unlike real-time assessments, video reviews can potentially be blinded, thereby limiting the potential for bias on the part of the reviewer.

Blinding of reviewers and evaluators to gender in other academic domains has helped reduce gender disparities.[16,17] As such, blinding faculty members to gender in the evaluation of simulated procedures could provide a methodologically rigorous way to investigate the etiology and magnitude of evaluation disparities and may help further mitigate gender bias in procedural evaluations.

There are currently no validated methods by which researchers can reliably blind reviewers to gender to evaluate for bias in these evaluations. The objective of this study was to evaluate if a novel method using video-based evaluations could blind faculty evaluators to the gender of trainees performing simulated procedures.

Methods

We conducted a prospective evaluation to determine the effectiveness of a novel video-based method of blinding reviewers to the gender of residents performing simulated procedures. This study was approved by the Adena Health System Institutional Review Board (protocol number: 18-05-011).

Video Content Development

Resident trainees in the emergency medicine residency program at the Adena Health System, a community-based residency program in the Midwest, volunteered to perform procedures for the video content. Two self-identified males and three self-identified females performed three procedures each (lumbar puncture, tube thoracostomy, and central venous catheter placement) on patient simulators, yielding 15 procedure videos. The proceduralists removed all jewelry from their hands and wrists. Each proceduralist donned a pair of opaque blue gloves followed by a standard Association for the Advancement of Medical Instrumentation (AAMI) Level 3

Figure 1. Proceduralist performing right internal jugular central line placement

Figure 2. Proceduralist preparing for lumbar puncture.

Figure 3. Proceduralist performing tube thoracostomy

The scale was coded as 1=definitely male, 3=likely male, 5=can't tell, 7=likely female, 9=definitely female. A response with a confidence level of "likely" or higher (1-3 for males or 7-9 for females) was considered correct gender identification if the reviewers' scores were concordant with the proceduralist self-reported genders. All proceduralists self-reported binary gender identification as either female or male. If the reviewers score was discordant with the proceduralist self-reported gender or if they were unsure (score 4-6) they were coded as not having correctly identified the gender of the proceduralist.

Analysis

Reviewer scores were reported per video assessment, and percentage correct was calculated for scores that were concordant with the proceduralist self-reported genders. The primary outcome was correct identification of the proceduralist self-reported gender by the reviewer. As correct gender identification would be 50% based on chance alone, it was hypothesized that blinding would be effective if the reviewers did not correctly identify proceduralists' genders in ≥50% of assessments. One sample equality of proportions was used to assess significance against the null of >50% correct gender identification in the overall sample. Similar one sample equality of proportions was also performed stratified by the three procedures evaluated by the reviewers. Data analysis was carried out with blinding to the gender of the proceduralists.

Results

Fifteen clinical and core residency faculty (11 males and four females) were offered the opportunity to participate, seven of which (five males and two females) chose to participate (**Table 1**).

Table 2 details the responses of each reviewer to each video. Of the 105 total responses, approximately half (56; 53.3%) of the reviewers' responses expressed confidence in the gender of the proceduralist (1-3 or 7-9),

opaque blue surgical gown with tapered elastic wrist cuffs that extended past the wrist. Lastly, a pair of surgical latex gloves was donned on top of the opaque blue gloves, covering the gown wrist cuffs. A professional videographer (Blue Skies Video & Film Productions, LLC) was contracted to film the simulated procedures. A high-resolution video camera was used to record each procedure. Only the procedural field, the double-gloved hands of the proceduralists, and portions of their gowned forearms and lower torsos were visible in the videos (**Figures 1, 2,** and **3**). Prior to the procedure, each video was assigned a unique number and entered into a log. The raw video was then edited, using Premiere Pro video editing software (Adobe Corporation) to create video segments approximately 5 seconds in duration, which were compiled into a single 3:02 video and posted to a password-protected website (vimeo.com).

Protocol

The study participants were seven graduate medical educators with experience evaluating residents recruited by the primary author from both within and outside the study site. A template email contact developed by the authors was used for recruitment.

The reviewers scored each video using a Likert-type scale to identify the perceived gender of the proceduralist.

Table 1. Recruitment of Participants

Program	Males Recruited (Participants)	Females Recruited (Participants)	Total Recruited (Participants)
Emergency medicine at primary site	3 (1)	1 (0)	4 (1)
Family medicine at primary site	3 (2)	1 (0)	4 (2)
Internal medicine at primary site	1 (1)	0 (0)	1 (1)
Emergency medicine at other sites	4 (1)	2 (2)	6 (3)
Total	11 (5)	4 (2)	15 (7)

Table 2. Reviewer Responses

Procedure	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Mean score per individual operator	Operator gender
CVL	3	5	8	6	5	7	9	6.3	Female
LP	5	3	1	3	1	3	1	2.4	Male
CVL	5	7	4	7	7	2	5	5.3	Male
TT	5	5	6	2	5	2	5	4.3	Male
TT	7	6	1	8	6	2	2	4.6	Male
LP	5	7	9	4	5	1	7	5.4	Male
CVL	4	5	9	7	5	6	6	6.0	Male
TT	7	4	1	7	5	6	6	6.3	Female
TT	7	4	1	7	5	2	4	4.3	Female
LP	7	7	9	5	5	2	3	5.4	Male
CVL	3	7	1	2	5	5	4	4.0	Male
TT	6	6	9	7	5	7	6	6.6	Female
LP	6	7	3	7	5	3	1	4.6	Female
LP	6	5	1	5	5	3	2	3.9	Female
CVL	5	5	9	5	5	1	5	5.0	Female
Correctly identified	20%	20%	53%	53%	7%	47%	33%		

Score: 1, definitely male; 3, likely male; 5, unable to determine; 7, likely female; 9, definitely female. Key: Green, gender correctly identified (1-3, 7-9, concordant with proceduralist gender); pink, gender unable to be identified (4-6); red, gender incorrectly identified (1-3, 7-9, discordant with proceduralist gender)
Procedures: CVL, central venous line; LP, lumbar puncture; TT, tube thoracostomy

but when they did, they were incorrect in 37.5% (21/56) of the assessments. Across all reviewers and procedures, 33.3% (95% CI: 25.1-42.8%) correctly identified the gender of the proceduralist. This proportion was statistically nonsignificant as compared with the null of >50% correct gender identification (p=1.00). The same nonsignificant differences were maintained when the data were stratified by each of the three procedures assessed. The mean response for four of the five proceduralists were between 5-5.7 (5=can't tell); the mean for the fifth (a male) was 3.7, outside of the range that indicated a confidence level of likely or higher.

Discussion

Assessment of procedural competency should be objective and performed without regard to gender. We believe that research in this area is required in order to better define the extent of gender bias in medical education and the effectiveness of interventions to counter it. Before such research in the area of procedural eval-

uations can be performed, it is essential to develop and validate a method that could effectively conceal the gender of the proceduralist. This study, which found that a novel technique is effective in blinding evaluators to the gender of the proceduralist, lays the groundwork for potential future studies.

An effective blinding method for procedural skills assessment allows for more accurate evaluation for implicit gender bias. A recent systematic literature review identified nine studies examining the presence and influence of gender bias on resident assessment.[18] Eight examined faculty evaluation of residents in a real-world setting, inherently unblinded. Another examined faculty assessment of resident performance in simulated, standardized encounters, but in an unblinded fashion.[19] With an effective blinding method, further research focusing specifically on procedural skills assessment can be obtained.

This method also has potential for use outside of the research setting. Training programs can easily replicate this blinding method to limit the effect of gender bias in evaluations for procedural competency. The gloves and gowns required to conceal the gender of a trainee performing a procedure are readily available, and most training institutions have the ability to use video to record procedures in a simulation laboratory. Isaak and colleagues found video-based simulation reviews for the assessment of milestones in anesthesia residents to be as reliable as real-time assessment.[20] Moving some procedural evaluations from the bedside to video review, with the identity and gender of the proceduralist effectively concealed, would limit if not remove the effect of gender bias from these evaluations and, if bias is indeed a significant contributor, may reduce the documented gender disparity.[8,9]

Limitations

The study participants were disproportionately male. The Likert scale employed was developed de novo and was not validated, and the reviewers did not receive formal training on the use of the scale. It is possible that longer clips, if available, would have revealed the gender of the proceduralist to the reviewers. There are stereotypical traits for each gender (eg, larger hands or fast and aggressive movements assigned to the male gender by some reviewers) that no blinding could conceal. It is possible that movements associated with procedures not included in this sample may reveal these or other stereotypical gender characteristics and therefore produce different results, which could theoretically have a negative effect on the applicability of our results to other procedures. Because previous studies on gender

bias have focused on male and female and the proceduralists self-identified in that manner, a binary gender paradigm was used. It is not certain how trainees with nonbinary gender identity would be identified and evaluated using the methods described in this study. However, expansion of the identification scale to include additional identifications is feasible, and is a potential area for further study.

Conclusions

The method used was effective in blinding reviewers to the gender of the proceduralist, and represents an innovative approach to facilitate research into gender bias in procedural evaluations. The method could also potentially be used to mitigate the effect of gender bias on trainee procedural milestone evaluations. ■

References
1. Jena AB, Olenski AR, Blumenthal DM. Sex differences in physician salary in US public medical schools. *JAMA Intern Med.* 2016;176(9):1294-1304.
2. 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.
3. Lautenberger DM, Dander V, Raezer CL, et al. The state of women in academic medicine: the pipeline and pathways to leadership. 2013-2014. American Academy of Medical Colleges. Available at: https://store.aamc.org/downloadable/download/sample/sample_id/228. Accessed January 29, 2021.
4. Nelson L, Keim S, Baren J, et al. American Board of Emergency Medicine report on residency and fellowship training information (2017-2018). *Ann Emerg Med.* 2018;71(5):636-648.
5. Bennett CL, Raja AS, Kapoor N, et al. Gender differences in faculty rank among academic emergency physicians in the United States. *Acad Emerg Med.* 2019;26(3):281-285.
6. Madsen T, Linden J, Ronds K, et al. Current status of gender and racial/ethnic disparities among academic emergency medicine physicians. *Acad Emerg Med.* 2017;24(10):1182-1192.
7. Jagsi R, Griffith KA, Jones R, et al. Sexual harassment and discrimination experiences of academic medical faculty. *JAMA.* 2016;315(19):2120-2121.
8. Edmunds L, Ovseiko P, Shepperd S, et al. Why do women choose or reject careers in academic medicine? A narrative review of empirical evidence. *Lancet.* 2016;388(10062): 2948-2958.
9. Mueller AS, Jenkins TM, Osborne M, et al. Gender differences in attending physicians' feedback to residents: A qualitative analysis. *J Grad Med Educ.* 2017;9(5):577-585.
10. 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.
11. Santen S, Yamazaki K, Holmboe E, et al. Comparison of male and female resident milestone assessments during emergency medicine residency training: a national study. *Acad Med.* 2020;95(2):263-268.
12. Aggarwal R, Grantcharov T, Moorthy K, et al. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann Surg.* 2008;247(2):372-379.
13. Driscoll PJ, Paisley AM, Paterson-Brown S. Video assessment of basic surgical trainees' operative skills. *Am J Surg* 2008;196(2):265-272.
14. Dath D, Regehr G, Birch D, et al. Toward reliable operative assessment – the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc.* 2004;18(12):1800-1804.
15. Hassanpour N, Chen R, Baikpour M, et al. Video observation of procedural skills for assessment of trabeculectomy performed by residents. *J Curr Ophthalmol.* 2016;28(2):61-64.
16. Budden A, Tregenza T, Aarssen L, et al. Double-blind review favours increased representation of female authors. *Trends Ecol Evol.* 2008;23(1):4-6.
17. Tricco AC, Thomas SM, Antony J, et al. Strategies to prevent or reduce gender bias in peer review of research grants: a rapid scoping review. *PLoS ONE* 2017;12:e0169718.
18. Klein R, Julian KA, Snyder ED, et al. Gender bias in resident assessment in graduate medical education: review of the literature. *J Gen Intern Med.* 2019;34(5):712-719.
19. Holmboe ES, Huot SJ, Brienza RS, Hawkins RE. The association of faculty and residents' gender on faculty evaluations of internal medicine residents in 16 residencies. *Acad Med.* 2009;84(3):381-384.
20. Isaak R, Stiegler M, Hobbs G, et al. Comparing real-time versus delayed video assessments for evaluating ACGME sub-competency milestones in simulated patient care environments. *Cureus.* 2018;10(3):e2267.