Exploring Gender Bias in Nursing Evaluations of Emergency Medicine Residents

Krista Brucker, MD[1], Nash Whitaker, MD[1], Zachary S. Morgan, PhD[2], Katie Pettit, MD[1], Erynn Thinnes, MD[1], Alison M. Banta[1], and Megan M. Palmer, PhD[1]

ABSTRACT

Objectives: Nursing evaluations are an important component of residents' professional development as nurses are present for interactions with patients and nonphysician providers. Despite this, there has been few prior studies on the benefits, harms, or effectiveness of using nursing evaluations to help guide emergency medicine residents' development. We hypothesized that gender bias exists in nursing evaluations and that female residents, compared to their male counterparts, would receive more negative feedback on the perception of their interpersonal communication skills.

Methods: Data were drawn from nursing evaluations of residents between March 2013 and April 2016. All comments were coded if they contained words falling into four main categories: standout, ability, grindstone, and interpersonal. This methodology and the list of words that guided coding were based on the work of prior scholars. Names and gendered pronouns were obscured and each comment was manually reviewed and coded for valence (positive, neutral, negative) and strength (certain or tentative) by at least two members of the research team. Following the qualitative coding, quantitative analysis was performed to test for differences. To evaluate whether any measurable differences in ability between male and female residents existed, we compiled and compared American Board of Emergency Medicine in-training examination scores and relevant milestone evaluations between female and male residents from the same period in which the residents were evaluated by nursing staff.

Results: Of 1,112 nursing evaluations, 30% contained comments. Chi-square tests on the distribution of valence (positive, neutral, or negative) indicated statistically significant differences in ability and grindstone categories based on the gender of the resident. A total of 51% of ability comments about female residents were negative compared to 20% of those about male residents ($\chi^2 = 11.83$, $p < 0.01$). A total of 57% of grindstone comments about female residents were negative as opposed 24% of those about male residents ($\chi^2 = 6.03$, $p < 0.01$).

Conclusions: Our findings demonstrate that, despite the lack of difference in ability or competence as measured by in-service examination scores and milestone evaluations, nurses evaluate female residents lower in their abilities and work ethic compared to male residents.

From the [1]Department of Emergency Medicine, Indiana University, Indianapolis, IN; and [2]Institutional Research, California Institute of the Arts, Valencia, CA.
Received October 30, 2018; revision received July 15, 2019; accepted July 30, 2019.
Presented at the Society for Academic Emergency Medicine Annual Meeting, Indianapolis, IN, May 2018.
The authors have no relevant financial information or potential conflicts to disclose.
Author contributions: KB—study concept and design, analysis and interpretation of the data, drafting of the manuscript, critical revision; NW—study concept and design, acquisition of the data, analysis and interpretation of the data, critical revision of the manuscript; ZM—study concept and design, analysis and interpretation of the data, drafting of the manuscript, statistical expertise; KP—study concept and design, acquisition of the data, analysis and interpretation of the data, drafting of the manuscript, critical revision of the manuscript; ET—acquisition of the data, analysis and interpretation of the data, drafting of the manuscript; AB—acquisition of the data, analysis and interpretation of the data, administrative, technical, or material support; MP—study concept and design, analysis and interpretation of data, study supervision.
Supervising Editor: Esther K. Choo, MD, MPH.
Address for correspondence and reprints: Krista Brucker, MD; e-mail: kbrucker@gmail.com.
ACADEMIC EMERGENCY MEDICINE 2019;26:1266–1272.

ISSN 1553-2712

doi: 10.1111/acem.13843

Evaluations are a commonly implemented tool for feedback in graduate medical education. Faculty evaluations provide important feedback on resident physician performance to guide improvement during training. Studies have demonstrated that multidisciplinary feedback can be useful and reliable.[1–4] One prospective study demonstrated that multidisciplinary evaluations improved performance of residents compared to faculty feedback alone.[5] Nursing staff are thought to be an important component of a resident 360° evaluation as they are often present for resident interactions with patients, families, and other medical personnel.[6] Effective collaboration and teamwork are essential skills for emergency medicine (EM) residents as is evidenced by their inclusion in the Accreditation Council for Graduate Medical Education (ACGME) milestones.[7]

Several studies have attempted to explore the dynamic relationships between genders in leadership positions in medicine. Keck-McNulty and Wear[8] reported that female residents most commonly expressed "excessive self-monitoring of communication style due to fears of being perceived as too demanding and not friendly enough … having to justify their orders more than their male peers … and receiving less assistance than their male peers." The study of female leadership roles during resuscitations by Linden and colleagues[9] also revealed gender discrepancies, stating that "female residents had to earn the trust and respect of the nurses more than their male counterparts." These prior studies suggest that female residents continue to face challenges in their training program that their male counterparts do not.

Furthermore, recent research has revealed the presence of a gender bias in faculty evaluations of EM residents.[10] However, few studies have sought to examine for the presence of gender bias in 360° evaluations, and results are conflicting.[2,11] Early literature found that female residents received more favorable evaluations from nursing staff,[11] whereas a more recent study in 2015 found the opposite; women received harsher feedback from nursing staff.[12] The purpose of our study was to determine if gender bias exists in nursing evaluations of our EM residents.

METHODS

This is a retrospective study at a single ACGME-accredited EM residency program in Indiana University. The nursing evaluations at our institution are used to assess professionalism, interpersonal skills, and communication. In addition, the evaluation form includes a free-text box where nurses comment on any aspect of resident performance not strictly limited to communication and professionalism. The residency program supports 69 residents working in three urban emergency departments with a combined annual patient volume of over 250,000 visits. This study was reviewed by the institutional review board and was deemed to be exempt research.

To evaluate if any measurable differences in ability between male and female residents existed, we compiled and compared American Board of Emergency Medicine (ABEM) in-training examination scores and relevant milestone evaluations between female and male residents from the same period when the residents were evaluated by nursing staff. Milestones included in our evaluation included: Systems-based Practice 2 (participates in strategies to improve healthcare delivery and flow, demonstrates an awareness of and responsiveness to the larger context and system of health care); Professionalism (PROF) 1 (demonstrates compassion, integrity, and respect for others as well as adherence to the ethical principles relevant to the practice of medicine.); PROF 2 (demonstrates accountability to patients, society, profession, and self); Interpersonal and Communication Skills (ICS) 1 (demonstrates interpersonal and communication skills that result in the effective exchange of information and collaboration with patients and their families); and ICS 2 (leads patient-centered care teams, ensuring effective communication, and mutual respect among members of the team).

To evaluate if gender bias was present in nursing evaluations, we reviewed nursing evaluations completed between March 2013 and April 2016. On a bi

annual basis, all nurses working at each of the clinical sites were sent an electronic standard evaluation form for 10 assigned residents (Data Supplement S1, Appendix S1, available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1111/acem.13843/full). A list of nurses was provided by the departments' nursing leadership. Forms were sent and completed via the residency's online evaluation platform (MedHub). Assignments of nurses to particular residents and distribution of evaluations was completed by residency administration staff. Approximately 40 requests per resident were made, and the range of completed evaluations each resident received was 10 to 18.

To ensure blinding of resident gender for the reviewers, one member of the research team obscured the names and gendered pronouns from the comments. After blinding the comments were distributed equally to two independent reviewers. The reviewers were authors and were not blinded to the hypothesis of the study, but were blinded to the gender of the resident associated with each comment. The author responsible for blinding did not participate in coding.

The coding scheme used for this study was based on prior research (see Data Supplement S1, Appendix S2). Trix and Penska[13]and Schmader et al.[14] created word lists to perform a comparison of letters of recommendation based on gender. In our study, we used the grindstone, ability, and standout categories developed by Schmader and colleagues. Schmader and colleagues also coded for communication but a list of those words was not published in their 2007 study. Thus, based on our initial hypothesis that female residents would receive more comments related to interpersonal skills, we created a fourth word list. Our category of interpersonal closely matches the concepts noted in Madera et al.[15] communal category but since a specific word list was not published we generated our own by reviewing a subset of evaluations completed by our faculty.

Comments were coded with *standout* if they distinguished the resident from his or her peers, regardless of whether it was in a positive or negative manner. Comments were coded *ability* when the comment spoke to the technical skill, knowledge, and competence of the resident. For *grindstone* the comment spoke to the resident's work ethic, effort, or efficiency. Finally, comments were coded *interpersonal* if they described the type or quality of communication residents had with patients, nurses, families, or other caregivers. Comments could be coded in more than one category. An example of comments representative of these categories is shown in Table 1.

Members of the research team then determined the *valence* and *strength* of the comment. Options for the valence of a category included positive, neutral, or negative based on how favorably or unfavorably the evaluator described the resident on that topic. The strength of each category was coded as certain or tentative based on strength of conviction the commenter had on the topic. Discrepancies were resolved by distributing the comments coded differently by the two reviewers to a third member to make a determination based on their interpretation of the comment, which was informed by the comments of the other two members.

To test for differences in the responses to the quantitative questions on the nursing 360 evaluation we

Table 1
Examples of Positive and Negative Comments in Each Category

Category	Positive Example	Negative Example
Standout	Dr [NAME] is one of our favorites! [They] is great to work with, always staying calm in high stress situations—which of course we see a lot of those. [Their] time management is something other residents should strive for. [They] always keeps everyone up to date on the plan of care and is truly a joy to work with. Would love to see Dr [NAME] become a permanent member of our family when [their] residency is complete. [They] would be a huge asset to our team.	Extraordinarily dismissive and condescending toward nursing and support staff such as unit secretaries and not much better with patients and families.
Grindstone	I appreciate [NAME]'s desire to jump in and help with new patients. [They] also is willing to help fill the gaps on sick patients when the patient's primary resident is tied up.	I think that [NAME] does a great job explaining things to patients and addressing their concerns. I do feel that at times, from my perspective, [they] can get easily overwhelmed and get behind a bit during a busy shift. I feel like in these circumstances that [they] can get a little behind and is not always able to keep up with updating the nurses and or patients about the next steps in their care.
Interpersonal	Dr. [NAME] is extremely professional and energetic. [Their] bedside manner is phenomenal. [They] does a great job at taking the time to talk to patients even when they are difficult. [They] is great at approaching them in a very empathetic fashion.	Dr. [NAME] lacks communication skills, [they] is very condescending to RN's and ancillary staff, I don't think [they] intends for it to be that way but [their] tone and behavior comes across that way which many nurses find offensive. [They] is not very sympathetic to patient family members, [they] can be abrasive.
Ability	Dr. [NAME] is a strong resident. [They] seems very knowledgeable about cases and handles them well.	Dr. [NAME] is a very nice person and pleasant and polite in interactions with nursing and patients/families. There are many times however, that it appears [they] becomes overwhelmed easily and is not as efficient and confident in [their] decision making.

Table 2
In-service and Milestone Scores of Female and Male Residents

Score	Female (n = 33)[a]	Male (n = 50)[a]	Mean Difference	t	p-value
In-Service					
Year 2	75.64 (±19.00)	79.90 (±9.29)	4.27	1.37	0.174
Year 3	75.42 (±19.02)	81.44 (±8.65)	6.02	1.96	0.096
Milestone					
SBP 2	3.86 (±0.33)	3.77 (±0.29)	0.92	1.34	0.184
PROF 1	3.77 (±0.28)	3.78 (±0.30)	0.01	0.16	0.876
PROF 2	3.75 (±0.26)	3.75 (±0.32)	0.01	0.08	0.940
ICS 1	3.85 (±0.30)	3.80 (±0.27)	0.04	0.66	0.509
ICS 2	3.79 (±0.28)	3.76 (±0.29)	0.02	0.35	0.726

[a]Data are reported as mean (±SD).
ICS = Interpersonal and Communication Skills; PROF = Professionalism; SBP = Systems-based Practice.

used the Mann-Whitney U-test, which is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous and not normally distributed. Once qualitative coding was complete, we tested for statistically significant differences between male and female residents. To test for differences in distributions of these codes between male and female residents, due to the dichotomous nature of the present/absent codes and the bimodal nature of the other codes, chi-square tests were used using SPSS version 24 (IBM). First analyses looked for differences in the presence of each of the categories. Subsequent analyses only looked at comments with the topics present to explore differences in the valence and strength of each of the categories based on the gender of the residents. Significance was determined at a p-value < 0.05.

RESULTS

We did not find significant differences in skills or abilities between male and female residents on ABEM in-training examinations or selected milestone evaluations (Table 2). Reviewing the ordinal scale data available on 1,112 nursing evaluations reveals that female residents are reported to be less professional in their interactions with nurses (p = 0.041) and have less effective team leadership skills (p = 0.019) when compared to their male counterparts. Further, nurses are less likely to report being comfortable with female residents taking care of their family members (p = 0.013; Table 3).

Of the 1,112 completed evaluations, 332 (30%) contained text in the open-ended qualitative comments section. The proportion of evaluations which included free-text comments was not significantly different

Table 3
Means, Standard Deviations, Medians, and Results of the Mann-Whitney U-test for Female and Male Residents on Quantitative Review Items

	Female			Male				
Evaluation Item	N	Mean (±SD)	Median	N	Mean (±SD)	Median	Z	p-value
Is the resident responsive to patient and family needs/ questions?	443	8.34 (±1.69)	9	663	8.46 (±1.60)	9	−0.95	0.34
Does the resident effectively communicate with you?	442	8.05 (±1.98)	9	663	8.18 (±1.90)	9	−1.01	0.31
Does the resident behave professionally in their interactions with you?	442	8.49 (±1.85)	9	664	8.69 (±1.73)	9	−2.04	0.04
Does the resident behave professionally in their interactions with patients and/or families?	442	8.61 (±1.65)	9	661	8.76 (±1.57)	9	−1.62	0.11
Does the resident effectively demonstrate team leadership skills?	442	7.83 (±2.02)	8	660	8.08 (±1.97)	8	−2.35	0.02
Does the resident respond in a reasonable and timely fashion to your questions and concerns about patient care/needs?	442	8.07 (±1.93)	9	661	8.29 (±1.74)	9	−1.68	0.09
Would you be comfortable with this resident's care for you or a family member?	441	7.76 (±2.31)	8	662	8.09 (2.14)	9	−2.47	0.01

between female and male residents (33% vs 27%, $\chi^2 = 3.425$, $p = 0.06$). The length of the comments also did not differ significantly between female and male residents (medians = 23 words vs. 19 words, $p = 0.14$).

Upon review of the chi-square results, we did not find statistically significant differences between nurses' comments about male or female residents in terms of whether any of the four categories were present (Table 4). Both *standout* and *grindstone* language was relatively rare in comments of both male and female residents, being present in only about one-sixth of the coded comments. *Ability* language was more common, which was coded in one out of every three comments for both male and female residents. By far the most common category coded was *interpersonal* language; approximately four-fifths of comments spoke of residents' interpersonal skills (Table 4).

Valence of Language

Chi-square tests on the distribution of valence within each category revealed statistically significant differences based on gender in the use of *ability* and *grindstone* language, as shown in Table 5. Regarding *ability* language, 51% of female residents had negative ability comments, while only 20% of male residents had negative ability comments ($p < 0.01$).

Similarly, 57% of the *grindstone* comments about female residents were negative compared to (24%) of male residents, while more than three-quarters (76%) of male residents received positive *grindstone* language (p 0.01). The most parity in valence between male and female residents was in *interpersonal* language. Both female and male residents had nearly three-quarters (72%) of their *interpersonal* comments coded as positive ($p = 0.92$).

Table 4
Proportion of Nursing Comments by Category

Code	Female (n = 147)[a]	Male (n = 185)[a]	χ^2	p-value
Standout			0.00	0.99
Present	26 (17.8)	33 (17.8)		
Absent	120 (82.2)	152 (82.2)		
Ability			0.13	0.72
Present	47 (32.2)	63 (34.1)		
Absent	99 (67.8)	122 (65.9)		
Grindstone			0.25	0.62
Present	23 (15.8)	33 (17.8)		
Absent	123 (84.2)	152 (82.2)		
Interpersonal			1.20	0.27
Present	120 (82.2)	143 (77.3)		
Absent	26 (17.8)	42 (22.7)		

[a]Data are reported as *n* (%).

Table 5
Proportion of Nursing Comments by Valence

Code	Female[a]	Male[a]	χ^2	p-value
Standout	n = 26	n = 33	5.09	0.08
Positive	18 (69.2)	29 (87.9)		
Neutral	0 (0)	1 (3)		
Negative	8 (30.8)	3 (9.1)		
Ability	n = 47	n = 63	11.83	<0.01
Positive	22 (46.8)	50 (78.1)		
Neutral	1 (2.1)	1 (1.6)		
Negative	24 (51.1)	13 (20.3)		
Grindstone	n = 23	n = 33	6.03	0.01
Positive	10 (43.5)	25 (75.8)		
Neutral	0 (0)	0 (0)		
Negative	13 (56.5)	8 (24.2)		
Interpersonal	n = 120	n = 143	0.17	0.92
Positive	86 (71.7)	103 (72)		
Neutral	4 (3.3)	6 (4.2)		
Negative	30 (25)	34 (23.8)		

[a]Data are reported as *n* (%).

Strength of Language

Chi-square tests on the distributions of strength of language for each category only found a statistically significant for gender in the strength of the *ability* language. These results are shown in Data Supplement S2. Here almost one-third (30%) of *ability* comments about female residents appeared tentative while only one-seventh (14%) of those comments about male residents were tentative. Additionally, statistically significant yet moderate correlations between the valence and strength of *standout*, *ability*, and *interpersonal* language appear to show that nurses use more tentative language when giving criticism and more certain language when giving praise.

DISCUSSION

Although there is now strong evidence that gender bias exists across many areas of academic medicine,[8,12] the extent of the impact of those biases on daily professional interactions and professional training programs remains unknown. We designed this study to ask a simple question "is there gender bias in the way nurses evaluate residents?" Our study suggests gender bias in nursing evaluation of residents. Specifically, in

their written comments nurses evaluated female residents lower than their male counterparts in terms of ability and work ethic (grindstone). Although the discrepancies between male and female resident evaluations may be small and of unclear significance, they are concerning given the lack of gender differences in ability or competence as measured by in-service scores and milestone evaluations. This finding is similar to Mueller and colleagues[17] who examined the differences in qualitative feedback that male and female residents received from attendings. Interestingly, we did not find differences in the interpersonal domain.

Initially, we found our sample had similar themes for male and female residents throughout the comments. It was only in coding each to a positive or negative valence that we began to notice the differences in the nursing evaluation of male and female residents. As such, merely reporting the absence or presence of words or phrases is not enough for a study to truly evaluate if there is bias. It is important to get the qualitative nature of the comment.

LIMITATIONS

There are several limitations to consider when interpreting the results of this study. The first is that our study is limited to one residency program and its three clinical training sites. As such, the results of this study could be a result of our training environment and may not be found in other programs.

Another limitation is related to the evaluations and how they are completed. First, the nature of the relationship between the nurse and the resident they are evaluating could impact the results. Because evaluations are sent to nurses at random, there is no minimum amount of exposure to a resident required before a nurse has the ability to evaluate the resident. The evaluation provided to the nurses (Appendix S1) provides a Likert scale from 0 to 10 without specific anchors, which can lead to variability in evaluation. Further, the evaluations were constructed as a measure to get feedback for the residents and not for the purpose of this study. Therefore, the results of this study may be due to variability in nurses' exposure to the residents they are evaluating and/or their interpretation of the form. Further, surveys are completed anonymously so the gender of the nurse was not obtainable. Future work might explore the interactions between nursing gender and resident physician gender in influencing evaluations.

Finally, although we compared objective measures (milestones assessing communication and professionalism skills) between male and female residents as a surrogate marker for performance, this has not been proven to correlate to bedside performance. As a result, we cannot say with certainty that there are not differences in the abilities of our residents.

Another consideration is that coders were not blinded to the hypothesis of the study, but were blinded to the gender of the resident. This could have skewed the results toward finding bias. In addition the coding process, while based on previous work, required the combination of existing lists with those created specifically for this project, which may have affected our results. Finally, in our study design, individual comments could be in multiple categories, which would give greater weight to those comments and the respective nurses than comments falling into single categories.

CONCLUSION

The data presented here suggest gender bias in nursing evaluations of residents. We undertook this systematic study as a starting point in the design of a proactive effort to mitigate gender bias and bolster support for our female residents. More work is necessary to further understand the impact these differential evaluations have on the training experience of our female residents and what role they might play in our ability to recruit and retain women in academic emergency medicine.

References

1. Yang Y, Lee F, Hsu H, et al. Validation of the behavior and concept based assessment of professionalism competence in postgraduate first-year residents. J Chin Med Assoc 2013;76:186–94.
2. Massagli TL, Carline JD. Reliability of a 360-degree evaluation to assess resident competence. Am J Phys Med Rehab 2007;86:845–52.
3. Yazdankhah A, Norooz MT, Amoli HA, et al. Using 360-degree multi-source feedback to evaluate professionalism in surgery departments: an Iranian perspective. Med J Islamic Public Iran 2015;29:284.
4. Joshi R, Ling FW, Jaeger J. Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. Acad Med 2004;79:458–63.
5. Berger JS, Pan E, Thomas J. A randomized, controlled crossover study to discern the value of 360-degree versus

traditional, faculty-only evaluation for performance improvement of anesthesiology residents. J Educ Perioper Med 2009;11:E053.
6. Ogunyemi D, Gonzalez G, Fong A, Alexander C, Finke D, Donnon T, Azziz R. From the eye of the nurses: 360-degree evaluation of residents. J Contin Educ Health Prof 2009;29:105–10.
7. Edgar L, Roberts S, Holmboe E. Milestones 2.0: a step forward. J Grad Med Educ 2018;10:367–9.
8. Keck-McNulty C, Wear D. Attitude of female nurses and female residents toward each other: a qualitative study in one U.S. teaching hospital. Acad Med 2004;79:291–301.
9. Linden JA, Breaud AH, Mathews J, et al. The intersection of gender and resuscitation leadership experience in emergency medicine residents: a qualitative study. AEM Educ Train 2018;2:162–8.
10. 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 2017;177:651.
11. Kaplan CB, Centor RM. The use of nurses to evaluate house officers' humanistic behavior. J Gen Intern Med 1990;5:410–4.
12. Galvin SL, Parlier AB, Martino E, Scott KR, Buys E. Gender bias in nurse evaluations of residents in obstetrics and gynecology. Obstet Gynecol 2015;126 Suppl 4:7S–12S.
13. Trix F, Psenka C. Exploring the color of glass: letters of recommendation for female and male medical faculty. Discourse Society 2003;14:191–220.
14. Schmader T, Whitehead J, Wysocki VH. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. Sex Roles 2007;57:509–14.
15. Madera JM, Hebl MR, Martin RC. Gender and letters of recommendation for academia: agentic and communal differences. J Appl Psychol 2009;94:1591–9.
16. Carnes M, Bartels CM, Kaatz A, Kolehmainen C. Why is John more likely to become department chair than Jennifer? Trans Am Clin Climatol Assoc 2015;126:197–214.
17. Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM. Gender differences in attending physicians' feedback to residents: qualitative analysis. J Grad Med Educ 2017;9:577–85.

Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1111/acem.13843/full

Data Supplement S1. Supplemental material.

Data Supplement S2. Supplemental material.