Gender Disparity in Evaluation of Internal Medicine Clerkship Performance

Deborah J. Gotth, MD, PhD; Rogan G. Magee, MD, PhD; Sarah E. Rosenberg, MD; Nina Mingioni, MD

Abstract

IMPORTANCE Women studying medicine currently equal men in number, but evidence suggests that men and women might not be evaluated equally throughout their education.

OBJECTIVE To examine whether there are differences associated with gender in either objective or subjective evaluations of medical students in an internal medicine clerkship.

DESIGN, SETTING, AND PARTICIPANTS This single-center retrospective cohort study evaluated data from 277 third-year medical students completing internal medicine clerkships in the 2017 to 2018 academic year at an academic hospital and its affiliates in Pennsylvania. Data were analyzed from September to November 2020.

EXPOSURE Gender, presumed based on pronouns used in evaluations.

MAIN OUTCOMES AND MEASURES Likert scale evaluations of clinical skills, standardized examination scores, and written evaluations were analyzed. Univariate and multivariate linear regression were used to observe trends in measures. Word embeddings were analyzed for narrative evaluations.

RESULTS Analyses of 277 third-year medical students completing an internal medicine clerkship (140 women [51%] with a mean [SD] age of 25.5 [2.3] years and 137 [49%] presumed men with a mean [SD] age of 25.9 [2.7] years) detected no difference in final grade distribution. However, women outperformed men in 5 of 8 domains of clinical performance, including patient interaction (difference, 0.07 [95% CI, 0.04-0.13]), growth mindset (difference, 0.08 [95% CI, 0.01-0.11]), communication (difference, 0.05 [95% CI, 0.01-0.12]), compassion (difference, 0.125 [95% CI, 0.03-0.11]), and professionalism (difference, 0.07 [95% CI, 0.11-0.11]). With no difference in examination scores or subjective knowledge evaluation, there was a positive correlation between these variables for both genders (women: r = 0.35; men: r = 0.26) but different elevations for the line of best fit (P < .001). Multivariate regression analyses revealed associations between final grade and patient interaction (women: coefficient, 6.64 [95% CI, 2.16-11.12]; P = .004; men: coefficient, 7.11 [95% CI, 2.94-11.28]; P < .001), subjective knowledge evaluation (women: coefficient, 6.66 [95% CI, 3.87-9.45]; P < .001; men: coefficient, 5.45 [95% CI, 2.43-8.43]; P < .001), reported time spent with the student (women: coefficient, 5.35 [95% CI, 2.62-8.08]; P < .001; men: coefficient, 3.65 [95% CI, 0.83-6.47]; P = .01), and communication (women: coefficient, 6.32 [95% CI, 3.12-9.51]; P < .001; men: coefficient, 4.21 [95% CI, 0.92-7.49]; P = .01). The model based on the men’s data also included growth mindset as a significant variable (coefficient, 4.09 [95% CI, 0.67-7.50]; P = .02). For narrative evaluations, words in context with “he or him” and “she or her” differed, with agentic terms used in descriptions of men and personality descriptors used more often for women.

CONCLUSIONS AND RELEVANCE Despite no difference in final grade, women scored higher than men on various domains of clinical performance, and performance in these domains was associated

Key Points

Question Are objective and subjective evaluations of men and women participating in third-year internal medicine clerkships significantly associated with gender?

Findings In this cohort study of 277 students from a single academic center, women received higher scores for a majority of the evaluated domains of clinical performance, but there was no difference associated with gender in final grade. In addition, the content of narrative evaluations was significantly associated with the gender of the student being evaluated.

Meaning These findings suggest that students of different genders might not be evaluated equally during internal medicine clerkships.

Author affiliations and article information are listed at the end of this article.
Abstract (continued)

with evaluators’ suggested final grade. The content of narrative evaluations significantly differed by student gender. This work supports the hypothesis that how students are evaluated in clinical clerkships is associated with gender.

Introduction

In 1966, only 6.9% of medical school graduates were women; more recently, women made up approximately half of medical school graduates. This shift happened during the lifetimes of half of currently practicing physicians. During clerkships, medical students are evaluated not only on their medical knowledge but also on how well they assume the role of physician. With that role being historically gendered, implicit bias regarding how a physician should behave may be associated with how medical students are evaluated. Clinical performance evaluations account for the largest portion of clerkship grades, which carry significant weight in residency recruitment. Understanding potential gender-associated differences in clinical evaluation is necessary to ensure equity in the house staff selection process.

Published in 1978, one of the original studies exploring the association of gender with student performance found no difference between men and women in terms of course grades, clinical performance, and both written and oral examinations. Since that time, additional work has examined the association between gender and medical school performance with mixed results; generally, either no gender-associated difference is observed or a small increase in clinical performance is noted for women. In addition, research has found that narrative evaluations of women include more personality terms, whereas the focus for men includes more competency-related skills. These existing studies focus on 1 metric of evaluation, either overall scores or language analysis, without considering individual components of evaluations and how these metrics interact. However, whether differences in narrative evaluations may be traced to differences in clinical performance has not been studied.

We sought to examine how well individual components of clinical evaluations correlate with overall grade and whether that association is preserved when men and women are considered separately. We gathered both quantitative and qualitative clerkship evaluation data from a single school and class year and used regression analysis and natural language processing to examine gender-specific differences in these evaluations. We compared overall grades, the median scores in various domains of clinical performance for men and women, and the associations between these components to determine whether gender-associated differences were observable. We investigated potential differences in how subjective knowledge evaluation may be associated with multiple-choice examination performance. Finally, we analyzed the language in the narrative portion of the clerkship evaluations, attempting to uncover any significant differences in phrasing or word choice. To our knowledge, our study is the first to comprehensively examine the role that gender may play in clerkship evaluation, by investigating the association of gender with the interplay between overall evaluation, examination scores, and the content of narrative evaluations.

Methods

Setting, Participants, and Data

This was a single-center retrospective cohort analysis of evaluation data, including clerkship evaluations and the National Board of Medical Examiners Medicine Subject Examination (NBME MSE) scores, of third-year medical students in the 2017 to 2018 academic year at the Sidney Kimmel Medical College (SKMC) of Thomas Jefferson University, Philadelphia, Pennsylvania. The medicine
clerkship consists of 8 weeks with inpatient services, 4 weeks at the Thomas Jefferson University Hospital and 4 weeks at an affiliated community academic medical center. Similar to most institutions, SMKC uses a summative assessment of clinical performance by faculty and house staff, collected and aggregated by an online platform, as well as scores on the NBME MSE, which is taken at the end of the rotation. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline. This work was deemed exempt from review and from the need to obtain informed consent by the institutional review board of Thomas Jefferson University because data were collected as a part of the established curriculum and anonymized prior to analysis. No one received compensation or was offered any incentive for participating in this study.

The clinical performance of students is assessed using the Clerkship Evaluation Form, which consists of a free-text narrative portion, with a prompt to consider the students’ performance with the established Reporter-Interpreter-Manager-Educator model of evaluation,14 a Likert scale evaluation (1 = below expectations, 2 = expected, and 3 = exceeds expectations) of various domains in clinical practice (Box), and a suggested final grade (1 = fail, 2 = marginal, 3 = good−, 4 = good, 5 = good+, 6 = excellent, and 7 = honors). In addition, the evaluators state how much time they spent with the student (1 = superficial contact with student or minimal ability to assess this student, 2 = enough time to generally evaluate, 3 = solid amount of time or feel very comfortable about my ability to assess this student). The final clerkship grade consists of the weighted sum of the suggested final grade (70%), the NBME MSE score (10%), and timely completion of projects (20%). All narrative comments and their association with the final suggested grade are reviewed by a grading committee composed of the clerkship director and clerkship site directors from all affiliate sites. Our data set included anonymized student composite evaluations; gender was presumed by examining pronouns.

### Box. Language of Evaluation Prompts Targeting Various Domains of Clinical Performance

| Domain                                | Narrative Prompts                                                                 |
|---------------------------------------|----------------------------------------------------------------------------------|
| Patient interaction                    | Ability to establish humanistic rapport with patient.                             |
| Subjective knowledge                  | Demonstrates appropriate knowledge base and understanding of diseases.           |
| Growth mindset                        | Able to identify own strengths and areas for improvement.                        |
| Communication                         | Able to communicate with team about clinical, administrative, and personal tasks. |
| Resource utilization                  | Able to effectively utilize available resources.                                 |
| Compassion                            | Able to demonstrate compassion, integrity, and respect for others.                |
| Teamwork                              | Works with other health professionals and staff to establish and maintain a climate of mutual respect. |
| Professionalism                       | Demonstrates personal accountability.                                             |
|                                      | Manages competing needs of personal and professional responsibility.              |
|                                      | Demonstrates trustworthiness to one’s colleagues regarding the care of patients.   |

Able to maintain a clear, legible, and appropriate medical record.
Able to engage patients in education.
Demonstrates sensitivity and responsiveness to a diverse patient population.
Demonstrates integrity and commitment to ethical principles.
Respects patient confidentiality.
Advocates for patient safety.
Aware of concepts of cost, quality, and patient safety.
Regression Analysis

Univariate linear regression was used to assess the association between individual domains of clinical performance and the overall median score given by evaluators. Correlation was assessed (Pearson) and significance determined using analysis of covariance (GraphPad). Multivariate linear regression was used to assess the relative importance of each domain to overall evaluation using the lme4 method and relaimpo package in RStudio. This method has been used, for example, in medical education research to attempt to predict clinical performance based on undergraduate records. Relative importance uses modeling to determine which variables—in our case, score in each individual domain in clinical practice—are more important when determining an outcome—the overall suggested score in this study.

Natural Language Processing

Natural language processing is an umbrella term for a number of quantitative, machine learning approaches to automated analysis of written documents. Word embeddings use a statistical model to describe how often individual words are used together in a context. This analysis generates numeric vectors that represent the relationships between words and phrases throughout the entire text under consideration. These representations allow for words to be arranged such that the distance between individual vectors represents their typical context; the smaller the distance between words, the more often they are used in context with each other. For example, in calculated word embeddings from a biology textbook, the words frog and amphibian might be “closer” in the embedding space than the word frog would be to the word mammal. A similar method has been applied to examine the language of evaluations used for underrepresented minority students. Word embedding analysis was conducted by replacing all xxx’s that originally replaced the students’ names with the appropriate gendered pronouns (she, her, he, or his). Word embeddings were generated 1000 times using the word2vec and text2vec packages in RStudio. For each word embedding, we then queried the top 50 closest word vectors by the euclidean distance to the words he, she, her, or his. These top 50 word vectors represented the 50 closest words in context to the pronouns. We tallied the unique words in each of the 1000 lists. Finally, we performed 2-sided χ² tests on each word to identify which were differentially present in context with the target words (P < .05 after Benjamini-Hochberg correction).

Statistical Analysis

We assessed for normality with the Shapiro-Wilk test and evaluated differences with the Mann-Whitney test for nonparametric continuous data, analysis of covariance with the Kruskal-Wallis test for multiple continuous variables, and the χ² test for categorical variables (GraphPad). Data were analyzed from September to November 2020.

Results

Overall Grade Distributions

In total, 2589 evaluations of 277 students (140 [51%] presumed women with a mean [SD] age of 25.5 [2.3] years and 137 [49%] presumed men with a mean [SD] age of 25.9 [2.7] years) were collected. There was no difference in final clerkship grade distribution between men and women (Figure 1A). However, women had a higher suggested final score (difference, 0.21 [95% CI, 0.06-0.28]; P = .003) (Figure 1B). There was a trend for an increase in subjective knowledge evaluation scores as the academic year progressed, although this trend was not statistically significant (Figure 1C). The NBME MSE scores were not different between genders (Figure 1D). Subjective knowledge evaluation was positively correlated with NBME MSE score for both men and women (women: r = 0.35 and P < .001; men: r = 0.26 and P = .004), and the slope was not significantly different by gender. However, the correlation elevation was lower for women (y-intercept for women, 1.63 [95% CI, 1.29-1.97]; y-intercept for men, 1.87 [95% CI, 1.58-2.17]; P < .001) (Figure 1E). The length of the narrative
The final overall grade distribution awarded considering final clinical grade and National Board of Medical Examiners (NBME) score; χ² test shows no significant difference. B. The median value of the evaluator-suggested final grade (1 = fail, 2 = marginal, 3 = good−, 4 = good, 5 = good+, 6 = excellent, 7 = honors). C. Subjective knowledge evaluation for each quarter (Q) of the academic year. D. Internal medicine NBME examination score, taken at the end of the clinical rotation. E. Linear correlation between NBME score and the evaluator-assessed demonstrated knowledge base. Positive correlation between NBME performance and subjective knowledge evaluation is significant (women: r = 0.35, P < .001; men: r = 0.26, P = .004) for both genders, and the calculated slope of the correlation is not different. However, the y-intercept is significantly lower for women (women: 1.63 [95% CI, 1.29-1.97]; men: 1.87 [95% CI, 1.58-2.17]; P < .001); thus, a man with a low NBME score is more likely than a woman with the same score to be rated as having a better knowledge base. F. There was no difference in the number of words used in the narrative evaluation. G. Evaluators were more likely to use men’s names than women’s names when writing narrative evaluations. H. Students who earned honors grades were more likely to have their name mentioned more often in narrative evaluations.

a Difference, 0.21 (95% CI, 0.06-0.28); P = .003.
b Difference, 0.23 (95% CI, 0.08-0.36); P = .002.
c Difference, 0.24 (95% CI, 0.05-0.42); P = .007.
comments did not differ between genders (Figure 1F), but evaluators more often referenced men by name (difference, 0.23 [95% CI, 0.08-0.36]; \( P = .002 \)) (Figure 1G). Students receiving a grade of honors were referred to by name more often than students receiving a grade of excellent (\( P = .007 \)) (Figure 1H).

Women received better Likert scale scores on various individual domains of clinical performance, including patient interaction (difference, 0.07 [95% CI, 0.04-0.13]; \( P < .001 \)), growth mindset (difference, 0.08 [95% CI, 0.01-0.11]; \( P = .01 \)), communication (difference, 0.05 [95% CI, 0-0.12]; \( P = .01 \)), compassion (difference, 0.13 [95% CI, 0.03-0.11]; \( P < .001 \)), and professionalism (difference, 0.07 [95% CI, 0-0.11]; \( P = .02 \)) (Figure 2). There was no difference in the subjective knowledge evaluation, teamwork, or resource utilization.

Regression Analysis

In the univariate linear regression analysis, there was no correlation between patient interaction (Figure 3A), knowledge evaluation (Figure 3B), growth mindset (Figure 3C), communication (Figure 3D), or reported time spent with the student and evaluator-suggested final grade. Compassion was positively correlated with the evaluator-suggested final grade (women: \( r = 0.46, P < .001 \); men: \( r = 0.69, P < .001 \)) (Figure 3E), along with resource utilization (women: \( r = 0.56, P < .001 \),
Data are shown as linear regression lines plotted with 95% CIs as dotted black lines. In univariate analysis, patient interaction (A), subjective knowledge (B), growth mindset (C), and communication (D) Likert scores had no correlation with the overall score given to men and women. Compassion (women: $r = 0.46$; men: $r = 0.69$; $P < .001$ for slope) (E), resource utilization (women: $r = 0.56$; men: $r = 0.70$; $P < .001$ for slope) (F), teamwork (women: $r = 0.60$; men: $r = 0.67$; $P < .001$ for slope) (G), and professionalism (women: $r = 0.37$; men: $r = 0.76$; $P < .001$ for slope) (H) all had a significant correlation with the overall evaluation ($P < .001$). The relative importance of each variable determined by multivariate analysis is plotted for both models' patient interaction, subjective knowledge evaluation, reported time spent with the student, and communication. The men's data additionally included growth mindset as a significant variable (I). *$P \leq .05$. **$P \leq .01$. ***$P < .001$. 

*JAMA Network Open. 2021;4(7):e2115661. doi:10.1001/jamanetworkopen.2021.15661 (Reprinted) July 2, 2021 7/13*
Although univariate linear regression is valuable in that the resulting 2-dimensional
environment is easily visualized, multivariate regression analysis accounts for potential
interdependence of the composite variables.

Both multivariate regression models included the following significant variables: patient
interaction (women: coefficient, 6.64 [95% CI, 2.16-11.12]; P = .004; men: coefficient, 7.11 [95% CI,
2.94-11.28]; P < .001), subjective knowledge evaluation (women: coefficient, 6.66 [95% CI,
3.87-9.45]; P < .001; men: coefficient, 5.45 [95% CI, 2.43-8.43]; P < .001), reported time spent with
the student (women: coefficient, 5.35 [95% CI, 2.62-8.08]; P < .001; men: coefficient, 3.65 [95% CI,
0.83-6.47]; P = .01), and communication (women: coefficient, 6.32 [95% CI, 3.12-9.51]; P < .001; men:
coefficient, 4.21 [95% CI, 0.92-7.49]; P = .01). The model based on the men's data additionally
included growth mindset as a significant variable (coefficient, 4.09 [95% CI, 0.67-7.50]; P = .02). The
relative importance of each of these variables reveals similar patterns for men and women (Figure 3I).

Natural Language Processing

Word embedding analysis revealed numerous differences in the words associated with he vs she and
his vs hers. Figure 4 shows the significant words separated by word category. Among these
differences, the pronouns she or her were more often used in context with the words professional,
wonderful, eager, helpful, and team. The pronouns he or his were more often used in context with
improve/improved/improving, notes, rounds, communication, plan, presentation, skills, fund, and
performance.

Discussion

Despite scoring higher on 5 of 8 metrics, women did not have higher grades than men. There was no
difference in NBME MSE scores, subjective evaluation of student knowledge, or the slope of their
positive correlation, but the elevations of the lines of best fit were different. Regression analysis
revealed that time, patient interaction, communication, and knowledge were all associated with the
overall evaluation for groups, but a positive score in growth mindset was also associated with the
score for men. Narrative comments for men included more agentic terms, whereas narrative
comments for women focused on personality.

Our finding of no difference in final clerkship grade with differences in score subcomponents is
consistent with existing medical education literature. In 1978, Holmes and colleagues found no
gender-associated difference in overall course grades; however, the small number of women in
medical school made this work susceptible to a type II error. More recent work has found higher
clinical grades for women, with gender concordance or discordance in evaluator-student pairings
associated with outcomes. Our data set consisted of composite evaluations, each including
evaluators of all genders. As such, we were unable to examine whether a particular evaluator-student
gender concordance was associated with the observed differences. However, similar to those
previous studies, our results do show higher clinical evaluation scores for women.

Although women scored higher than men on many internal medicine clerkship subcomponents,
this achievement did not translate into a difference in final assigned grade. This phenomenon has
also been recorded in other medical student evaluations. For example, despite women performing
better than men on obstetrics and gynecology written examinations and clinical skills examinations,
faculty evaluation of students did not reflect the higher performance of women. One factor
potentially contributing to the discordance between higher clinical grades and no difference in grade
distribution may be indicated by the free-text evaluations. Specifically, familiarity with men, as
suggested by more frequent first name use, may be associated with better evaluations. Our data
showed that frequency of name use was higher in honors grading.
One of the more interesting findings was the difference in the correlation between NBME MSE score and subjective knowledge evaluation. Examination scores and median knowledge evaluation scores were the same for both genders, and there was a positive correlation between those variables for both genders with no difference in the slope of their best fit line. This finding shows that students who perform better on the NBME MSE were more likely to be evaluated as having a better fund of knowledge, regardless of gender. However, the best fit lines for men and women had different

![Figure 4. Word Embedding Differences](image-url)

Significantly different words (P < .05 after Benjamini-Hochberg adjustment for multiple comparison) of the top 50 closest words to either he or she or his or her.
elevations, suggesting that if a man and woman had the same NBME MSE score, the woman would be rated lower in subjective knowledge evaluation than the man.

The regression analysis suggests that, although there are many commonalities in how medical students are evaluated, there are differences in the expectations for men and women studying medicine. Although the time that the evaluator spent with the student was not correlated with overall evaluation in the univariate analysis, it was significantly correlated in the multivariate analysis. In addition, this finding was stronger for women than for men. A clear causal relationship cannot be inferred from this correlation owing in part to the possibility of hidden or confounding variables. That said, this correlation agrees with previous research showing that longer observation time is associated with higher grades.6,19 This finding raises the possibility of an actionable intervention that may substantively improve learner outcomes and equity—cognizant consideration of interaction time between faculty and students.

Patient interaction, medical knowledge, and communication were significant for both genders in the multivariate analysis. The importance of these clinical domains is reflected in existing literature; in a survey of faculty, clinical reasoning and professionalism were 2 of the most influential factors when grading students.20 Our univariate analysis found that professionalism, resource utilization, compassion, and teamwork were all positively correlated with overall evaluation, but none of these factors were significantly associated with this outcome in the multivariate analysis. These data suggest that, although these factors were significant and positively correlated with individual performance, overall grade was more likely associated with alternative factors in this cohort.

The words associated with gendered pronouns show interesting connections both to our other data and to the existing literature. For men, growth mindset was significant in the multivariate model of evaluator-suggested final grade, and the words improve/improved/improving were associated with the pronouns he or his. This finding agrees with a 2010 study that showed that men were more likely than women to be described as quick learners.12 This finding persisted in our work despite the women outperforming men on the component of the evaluation addressing growth mindset. This discordance between the growth mindset score, the overall clinical score, and free-text evaluation suggests that potential is more valued in men, a previously documented phenomenon in the selection for business leadership positions.21

Women were more likely than men to be described as professional, and women outperformed men on the quantitative evaluation of professionalism. However, our univariate and multivariate analyses results suggest that professionalism has a stronger positive association with overall grade for men. This discordance indicates that, although professionalism for women was more commonly documented, it was less associated with overall grade than it was for men. This finding raises the possibility of a baseline disparity in a priori assumptions regarding professionalism across genders.

Existing literature suggests that women are more likely than men to be described as compassionate and enthusiastic and described in relation to their teamwork.12,13,22 Neither the word compassion nor empathy was significantly different in our analysis, but the terms wonderful and eager were more readily associated with women. Furthermore, although there was no difference in the Likert scale evaluation of teamwork, the words team and helpful were more often used in context with she or her. Although comments on compassion were not significantly different in free text, women outperformed men on the component of the evaluation addressing compassion and patient interaction. In addition, a study of letters of recommendation for sciences graduate students found that standout words, such as wonderful and fabulous, were more often associated with men,23 but in our work, wonderful and fantastic were both associated with women. In contrast to these personality terms associated with she or her, words describing medical proficiencies (ie, notes, rounds, communication, plan, and presentation) were used in context with the pronouns he or his. This result agrees with a previous study that found evaluations for men were more likely to include competency-related words.11 Including a prompt prior to narrative evaluation collection suggesting focus on clinical proficiencies may be an effective approach to standardize evaluation content.
Limitations
Although the study included data collected in both academic and community settings, this work was conducted at a single institution, during 1 academic year, and with a limited sample size. Although evaluation practices at the source institution are similar to common practices at other institutions, future work should repeat these observations in independent cohorts to assess whether the findings of the present study are generalizable to other institutions and clerkships. Our study treated gender as a binary variable, which does not acknowledge the true spectrum of gender identity and presentation. We presumed gender based on the evaluators’ interpretation of the students’ gender presentations, not considering sex assigned at birth nor the students’ gender identity. Race is another important factor that was not considered in this work. Our data did not include that demographic information, which could be used to examine how to create more racial equity in medical student evaluations.

Conclusions
This work highlights gender disparity in medical student evaluations. Additional qualitative research evaluating free-text evaluations is necessary to further understand the context of the differences identified by this work— with a focused reading on growth mindset, personality, and skill components— potentially providing insight into questions raised by our results. Applying these methods to other clerkships may provide further information regarding potential gendered expectations in other medical specialties.
REFERENCES

1. Association of American Medical Colleges. The state of women in academic medicine: 2015-2016. Accessed February 26, 2020. https://www.aamc.org/data-reports/faculty-institutions/data/state-women-academic-medicine-pipeline-and-pathways-leadership-2015-2016

2. Young A, Chaudhry H, Pei X, Arnhart K, Dugan M, Steingard S. FSMB census of licensed physicians in the United States. 2018. J Med Regul. 2019;105(2):7-23. doi:10.30770/2572-1852-105.2.7

3. Hernandez CA, Daroowalla F, LaRochelle JS, et al. Determining grades in the internal medicine clerkship: results of a national survey of clerkship directors. Acad Med. 2021;96(2):249-255. doi:10.1097/ACM.0000000000003815

4. McDade W, Vela MB, Sánchez JP. Anticipating the impact of the USMLE Step 1 pass/fail scoring decision on underrepresented-in-medicine students. Acad Med. 2020;95(9):1318-1321. doi:10.1097/ACM.0000000000003490

5. Holmes FF, Holmes GE, Hassanein R. Performance of male and female medical students in a medicine clerkship. JAMA. 1978;239(21):2259-2262. doi:10.1001/jama.1978.03280480051020

6. Riese A, Rappaport L, Alverson B, Park S, Rockney RM. Clinical performance evaluations of third-year medical students and association with student and evaluator gender. Acad Med. 2017;92(6):835-840. doi:10.1097/ACM.0000000000001565

7. Linn BS, Zeppa R. Sex and ethnicity in surgical clerkship performance. J Med Educ. 1980;55(6):513-520. doi:10.1097/00001888-198006000-00007

8. Rutala PJ, Witzke DB, Leko EO, Fulginiti JV. The influences of student and standardized patient genders on scoring in an objective structured clinical examination. Acad Med. 1991;66(9)(suppl):S28-S30.

9. Wijesekera TP, Kim M, Moore EZ, Sorenson O, Ross DA. All other things being equal: exploring racial and gender disparities in medical school honor society induction. Acad Med. 2019;94(4):562-569. doi:10.1097/ACM.0000000000002463

10. Wang-Cheng RM, Fullkerson PK, Barnas GP, Lawrence SL. Effect of student and preceptor gender on clinical grades in an ambulatory care clerkship. Acad Med. 1995;70(4):324-326. doi:10.1097/00001888-199504000-00018

11. Rojek AE, Khanna R, Yim JWL, et al. Differences in narrative language in evaluations of medical students by gender and under-represented minority status. J Gen Intern Med. 2019;34(5):684-691. doi:10.1007/s11606-019-04889-9

12. Axelson RD, Solow CM, Ferguson KJ, Cohen MB. Assessing implicit gender bias in Medical Student Performance Evaluations. Eval Health Prof. 2010;33(3):365-385. doi:10.1177/0163278710375097

13. Ross DA, Boatright D, Nunez-Smith M, Jordan A, Chekrourd A, Moore EZ. Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. PLoS One. 2017;12(8):e0181659. doi:10.1371/journal.pone.0181659

14. Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. Acad Med. 1999;74(11):1203-1207. doi:10.1097/00001888-199911000-00012

15. Grömping U. Relative importance for linear regression in R: The package relaimpo. J Stat Softw. 2006;17(1):1-27. doi:10.18637/jss.v017.i01

16. Stegers-Jager KM, Themmen APN, Cohen-Schotanus J, Steyerberg EW. Predicting performance: relative importance of students’ background and past performance. Med Educ. 2015;49(9):933-945. doi:10.1111/medu.12779

17. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Preprint. Posted online October 16, 2013. arXiv 1310.4546. https://arxiv.org/abs/1310.4546

18. Bienstock JL, Martin S, Tsou W, Fox HE. Medical students’ gender is a predictor of success in the obstetrics and gynecology basic clerkship. Teach Learn Med. 2002;14(4):240-243. doi:10.1207/S15328015TLM1404_7

19. Ingram MA, Pearlman JL, Estrada CA, Zinski A, Williams WL. Are we measuring what matters? how student and clerkship characteristics influence clinical grading. Acad Med. 2021;96(2):241-248. doi:10.1097/ACM.0000000000003616

20. Herrera LN, Khodadadi R, Schmit E, et al. Which student characteristics are most important in determining clinical honors in clerkships? a teaching ward attending perspective. Acad Med. 2019;94(10):1581-1588. doi:10.1097/ACM.0000000000002836

21. Player A, Randsley de Moura G, Leite AC, Abrams D, Tresh F. Overlooked leadership potential: the preference for leadership potential in job candidates who are men vs. women. Front Psychol. 2019;10(MAR):755. doi:10.3389/fpsyg.2019.00755
22. Li S, Fant AL, McCarthy DM, Miller D, Craig J, Kontrick A. Gender differences in language of standardized letter of evaluation narratives for emergency medicine residency applicants. *AEM Educ Train*. 2017;1(4):334-339. doi:10.1002/aet2.10057

23. 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(7-8):509-514. doi:10.1007/s11199-007-9291-4

24. Hyde JS, Bigler RS, Joel D, Tate CC, van Anders SM. The future of sex and gender in psychology: five challenges to the gender binary. *Am Psychol*. 2019;74(2):171-193. doi:10.1037/amp0000307

25. Schudson ZC, Beischel WJ, van Anders SM. Individual variation in gender/sex category definitions. *Psychol Sex Orientat Gend Divers*. 2019;6(4):448-460. doi:10.1037/sgd0000346