









Examining Gender-Based Differences in Quantitative Ratings and Narrative Comments in Faculty Assessments by Residents and Fellows

Jessica Hane , MD
 Vivien Lee , BA
 You Zhou , MA
 Taj Mustapha , MD

Susan M. Culican , MD, PhD
 G. Nic Rider , PhD
 Paul R. Sackett , PhD
 Michael J. Cullen , PhD

ABSTRACT

Background Learner assessments of faculty are widespread in medicine, yet concerns are growing about possible biases in these assessments and their associations with gender disparities.

Objective To investigate gender-based differences in how residents and fellows describe faculty (rater effect) and how faculty are described (ratee effect) in faculty assessments, and their associations with teaching effectiveness ratings.

Methods We analyzed 2164 trainee assessments of University of Minnesota Medical School faculty from 2019 to 2023 with trainee and faculty gender information and narrative comments. Using natural language processing, we categorized words and 2-word groups (n-grams) into communal (eg, caring, kind), standout (eg, outstanding, amazing), and agentic/ability (eg, assertive, controlling) groups. We examined gender-based differences in n-grams used by trainees (rater effect) and received by faculty (ratee effect), and relationships between n-gram and teaching effectiveness ratings.

Results Women trainees used more communal (rater effect, incidence rate ratio [IRR]=1.36; 95% CI, 1.27-1.47), standout (IRR=1.20; 95% CI, 1.08-1.34), and agentic/ability words (IRR=1.37; 95% CI, 1.26-1.49; $P<.001$) than men trainees. Women faculty received fewer agentic/ability words than men faculty (ratee effect, IRR=0.83; 95% CI, 0.77-0.90; $P<.001$). Women trainees used fewer communal words when describing women faculty (interaction effect, IRR=0.84; 95% CI, 0.73-0.98; $P<.05$). Teaching effectiveness ratings correlated with faculty n-gram word frequency in standout (men: $r_s=0.29$, women: $r_s=0.28$, $P<.001$) and communal categories (men: $r_s=0.23$, $P=.003$; women: $r_s=0.22$, $P=.01$).

Conclusions Women trainees used more communal, standout, and agentic/ability descriptors, while women faculty had fewer agentic/ability descriptors. Women trainees used fewer communal words when describing women faculty. Standout and communal word frequency predicted teaching effectiveness ratings for both genders.

Introduction

Learner assessments of faculty are widespread in medical education. They assess teaching quality, identify faculty strengths and weaknesses, and inform curricular development.¹⁻³ Due to their role in high-stakes employment decisions, impartial faculty assessments with high levels of validity evidence are crucial. Biased learner assessments may perpetuate gender disparities in academic medicine, impacting hiring, compensation, advancement, and leadership representation.⁴⁻⁶ They could also hinder faculty promotion and discourage women faculty from pursuing academic careers altogether.⁷

Despite an early study that found no significant gender-based differences in trainee ratings,⁸ recent research indicates women faculty receive lower ratings than men in ambulatory care settings,⁹ mandatory clerkships,^{9,10} and clinical specialties with low representation of women.¹¹ Cullen et al¹² additionally

reported lower ratings for female faculty compared to male faculty in overall teaching performance and role modeling, with small to medium effect sizes. Indeed, meta-analyses show evidence of bias against women instructors across various study types (quantitative, qualitative, experimental).^{13,14}

Qualitative studies often explore gender differences in communal (relating to care for others) and agentic (assertive, controlling) words in assessment comments. These categories stem from social-psychological research linking cultural stereotypes to male leadership traits and female relationship-oriented qualities.¹⁵ While exceptions exist,¹⁶ most studies find that women faculty are associated with communal characteristics (eg, compassionate, empathetic) and men with agentic behaviors (eg, scientific, quick learners).¹⁷⁻¹⁹

In this study, we examine whether gender-based differences exist in how trainees describe faculty (rater effect) and how faculty are described (ratee effect) in faculty assessments. Unlike previous studies focusing on either qualitative or quantitative methods and examining only one effect,^{10,11,17-20} we

DOI: <http://dx.doi.org/10.4300/JGME-D-24-00627.1>

investigate both simultaneously. This approach reveals distinct aspects of bias in ratings and how communal versus agentic traits shape perceptions of exceptional teaching.

Rater effects examine whether men and women trainees emphasize different faculty behaviors in assessments. Social psychological theories suggest women are expected to display communal qualities (eg, friendliness, caring) and men agentic qualities (eg, assertiveness, competence).^{15,21} Expectancy violations theory posits that women emphasize communal traits to maintain social harmony.²² Thus, women trainees may prioritize warmth, while men trainees may focus on competence. Though not inherently biased, this pattern may reinforce gender stereotypes over teaching effectiveness, potentially disadvantaging faculty who defy these expectations.

Ratee effects examine whether trainees describe men and women faculty members differently. Implicit gender bias, shaped by cultural gender roles, links masculinity to assertiveness and leadership and femininity to caregiving and relationships.^{1,23} Describing men faculty with more agentic terms and women faculty with communal terms may reinforce these stereotypes, posing professional challenges for faculty who deviate from traditional norms.^{1,11,23}

We also bridge the qualitative-quantitative divide by examining how trainee descriptions relate to faculty ratings in 4 clinical teaching domains: overall teaching effectiveness, role modeling, facilitating knowledge acquisition, and teaching procedures. Previous studies on effective clinical educators have used either qualitative methods (eg, essays, interviews) or quantitative approaches (eg, faculty ratings).^{24,25}

We hypothesized gender differences in how trainees describe faculty (rater effect) and the descriptors faculty receive (ratee effect). We expected women trainees to use more communal terms (eg, warm) and men trainees more agentic terms (eg, capable), reflecting differing evaluative schemas. Similarly, we anticipated women faculty would receive more communal descriptors and men faculty more agentic ones, reinforcing stereotypes. We also examined trainee-faculty interactions but made no predictions due to inconsistent past findings.^{8,9} Given limited research on how descriptions relate to teaching domains, we did not hypothesize specific associations.

Methods

Procedure and Participants

This single-institution (University of Minnesota Medical School) retrospective cohort analysis included 14245 faculty assessments completed by residents and fellows for 1145 faculty in 18 clinical departments

KEY POINTS

What Is Known

Gender bias is well documented in the assessment literature, both related to gender of the assessor and the learner.

What Is New

This study of assessments performed by trainees on faculty showed an association between descriptor type (eg, standout words, communal words) and overall teaching effectiveness ratings and varied by faculty and trainee gender.

Bottom Line

Consumers of assessment data must understand the potential role for gender bias in descriptors and ratings of faculty.

(31 residency, 77 fellowship programs) between July 1, 2019, and June 30, 2023. Trainees routinely completed these anonymous electronic assessments during their clinical rotations using Qualtrics software (Qualtrics). Trainees voluntarily provided gender information and narrative comments when completing assessments, and faculty voluntarily provided gender information to the graduate medical education office.

Measures

Clinical Teaching Effectiveness: We employed a 22-item clinical teacher effectiveness measure, confirming a bifactor model with a general dimension of overall teaching effectiveness and 3 secondary dimensions (role modeling, facilitating knowledge acquisition, and teaching procedures).¹² Factor scores analyzed the relationship between trainee word choice and ratings.

N-Gram Analyses: To analyze trainees' narrative comments, we used natural language processing with n-gram analysis. N-grams are sequences of n words that provide insights into language patterns. For example, unigrams (ie, 1-grams) refer to individual words, and bigrams capture word pairings (ie, 2-grams). Comments underwent preprocessing steps: lowercasing, contraction expansion, lemmatization, special character removal, and stop word removal.²⁶ This process extracted unigrams and bigrams to capture contextual information.¹⁹ A document-term matrix recorded the frequency of each n-gram.

From 19645 unigrams and bigrams, n-grams appearing ≥ 10 times were selected (918 n-grams). The lead author (J.H.) reviewed these to identify 661 medically relevant n-grams for further analysis. While both negative and positive n-grams were initially identified, most retained n-grams had positive connotations, aligning with our emphasis on abilities, standout qualities, and communal characteristics (TABLE 1).

TABLE 1
Final List of N-Grams After Sorting by Faculty

Word Category	Relevant N-Grams
Ability/agentive words	teach, knowledge, evidence, skill, decision, autonomy, evidence base, knowledgeable, teach resident, ability, efficient, operate, create, provide feedback, precepting, job teach, clinical reason, educational, smart, confident, analytic, teach round, insight, expertise, guideline, medical knowledge, medical decision, practice evidence, teach procedure, diagnose, intelligent, teach provide, patient teach, teach style, specific feedback, knowledge base, knowledgeable, rationale, clinical knowledge, efficiently, efficiency, teach evidence, precept, teach session, explain reason, clinical scenario, fund knowledge, clinical experience, teach learn, knowledge gap, independently, technical, surgical skill, wisdom, advocate patient, detail orient, skill level, knowledge experience, educate, inquisitive, perform procedure, teach topic
Standout words	excellent, role model, wonderful, fantastic, amaze, excellent teacher, incredibly, incredible, favorite, greatly, asset, effective, leader, valuable, fantastic teacher, exceptional, super, extremely knowledgeable, excellent mentor, awesome, phenomenal, wonderful teacher, asset program, excellent job, favorite faculty, outstanding, provide excellent, perfect, excellent role, excellent faculty, exemplary, brilliant, consistently, excellent teach, invaluable, admire, tremendous, excellent attend, incredible teacher, wonderful job, impressive, excellent clinician, favorite staff, amaze faculty, excellent surgeon, model physician
Communal words	care, mentor, helpful, approachable, love, supportive, comfortable, constructive, share, interaction, support, fun, explain, constructive feedback, nice, style, communication, empathetic, teacher mentor, advocate, relate, guidance, foster, listen, compassionate, friendly, communicate, attitude, humble, respect, encourage resident, happy, empathy, pleasant, patience, professionalism, willingness, enjoy learn, bedside manner, mentorship, invest, guide, enjoyable, enjoy teach, helpful feedback, respectful, teach time, trust, feel support, care teach, supportive resident, interact, receptive, vulnerability, kindness, opportunity learn, approachable teach, warm, relationship, encouragement, personable, advocate resident, easy approach, time learn, demeanor, openness, compassion, provide helpful, teacher time, mentor teach, share knowledge, responsive, beneficial, fair, listener, learn opportunity, mentor teacher, approachable question, feedback time, mentor resident, time explain, serve, patient love, safe learn, eager teach, genuine, opportunity teach, support resident, rapport, humility, relate patient, teach feedback, easily approachable, approach patient, interaction patient, safe space, love learn, judgmental, demand, belittle, alongside, helpful learn, feel uncomfortable, teacher provide, environment learn, encourage learn, challenge resident, positive attitude, create environment, provide timely, question feel, provide support, treat resident, feel learn, time discuss, create safe, appreciative, contribute, patient encourage, feel question, care deeply, teach explain, feedback receive, question encourage, encourage critically, feel hear, teach opportunity, supportive encourage, encourage critical, positive learn

Subsequently, J.H. and 3 faculty members (T.M., S.M.C., G.N.R.) independently categorized these n-grams into 3 established groups from prior studies on gender differences in clinical performance assessments^{20,27-29}: (1) communal (eg, caring, kind, sympathetic); (2) standout (eg, outstanding, amazing, excellent); and (3) an agentive/ability category combining leadership traits (eg, self-confident, assertive) with knowledge and skills (eg, intelligent, gifted). A grindstone category focusing on work ethic was initially considered but excluded due to limited representation (9 words).

After sorting the n-grams, faculty reached consensus on categorization via discussion, focusing on specific n-grams and category definitions. For clarity, only n-grams unanimously assigned to a single category were included, resulting in a final 238 n-grams (121 unigrams and 117 bigrams; see TABLE 1).

Analyses

We employed hierarchical Poisson regression to analyze gender differences in trainees' use of n-grams (rater effect) and faculty's received n-grams (ratee effect) in communal, standout, and agentive/ability categories. This method controlled for rater and ratee effects while examining interactions. N-gram frequencies were summed per category in each comment. Poisson regression was used due to the count nature of the dependent variable. Coefficients were exponentiated to obtain incident rate ratios, which represent the factor by which the expected count of words in each category changes per unit increase in predictor variables. Given anonymous assessments, each comment was treated as independent, though faculty could receive multiple ratings. Predictors were added in 2 stages: first trainee and faculty gender, then their interactive effect to examine potential interactions.

We used logistic regressions to examine associations between trainee and faculty gender and n-gram usage in assessments. Given the large number of n-grams, we applied Elastic Net regression¹⁹ for variable selection, optimizing hyperparameters through grid search and 10-fold cross-validation based on ROC (receiver operating characteristic) curve performance. This yielded 120 n-grams (75 unigrams and 45 bigrams) for predicting trainee gender and 71 n-grams (54 unigrams and 17 bigrams) for predicting faculty gender. Separate logistic regression models were then fitted using these selected n-grams as predictors.

Finally, we used Spearman correlations to examine the relationship between faculty n-gram (ie, ratee) word frequencies in each word category and performance scores in the 4 clinical performance dimensions, and the Z-test to examine gender differences in those relationships.

All statistical and text analyses were performed using R Version 4.2.2 (R Core Team), and alpha was set to .05.

The University of Minnesota Institutional Review Board exempted this study from review.

Results

From the original sample, we retained only records with available trainee and faculty gender information and narrative comments. Our final data included 2164 assessments (1113 [51.4%] by women trainees and 1051 [48.6%] by men trainees) for 384 faculty members (188 [49.0%] women faculty and 196 [51.0%] men faculty). Assessments spanned 18 clinical departments (20 residency, 47 fellowship programs). The highest proportions were from family medicine and community health (698 of 2164 [32.3%] assessments), pediatrics (308 of 2164 [14.2%]), diagnostic radiology (305 of 2164 [14.1%]), and medicine (186 of 2164 [8.6%]); the remaining 667 assessments (30.8%) were from other departments, including anesthesiology, neurology, obstetrics and gynecology and women’s health, and urology, representing various hospital-based, medical, and surgical programs.

TABLE 2 shows significant effects for raters and ratees in the agentic/ability category (IRR=1.37 and 0.83; 95% CI, 1.26-1.49 and 0.77-0.90; $P<.001$), and a rater effect for standout and communal categories (IRR=1.20 and 1.36; 95% CI, 1.08-1.34 and 1.27-1.47; $P<.001$). Women trainees used 37% more agentic/ability, 20% more standout, and 36% more communal words compared to men trainees, controlling for faculty gender. Women faculty received 17% fewer agentic/ability words than men faculty, controlling for trainee gender. Interaction

TABLE 2 Hierarchical Moderated Poisson Regression of Frequency of N-grams by N-gram Categories in Trainees’ Ratings

Predictors	Agentic/Ability				Standout				Communal						
	IRR	B	SE	P value	95% CI for IRR	IRR	B	SE	P value	95% CI for IRR	IRR	B	SE	P value	95% CI for IRR
First step															
Intercept	0.98	-0.02	0.04	.65	0.92-1.06	0.56	-0.59	0.05	<.001	0.51-0.61	1.17	0.16	0.03	<.001	1.1-1.25
Trainee gender	1.37	0.31	0.04	<.001	1.26-1.49	1.20	0.18	0.06	<.001	1.08-1.34	1.36	0.31	0.04	<.001	1.27-1.47
Faculty gender	0.83	-0.18	0.04	<.001	0.77-0.9	1.02	0.02	0.05	.74	0.91-1.13	0.97	-0.03	0.04	.40	0.9-1.04
Pseudo R ²	0.02					0.00					0.02				
Adjusted Pseudo R ²	0.02					0.00					0.02				
Second step															
Trainee x faculty gender	1.06	0.06	0.09	.48	0.9-1.26	0.98	-0.02	0.11	.89	0.79-1.22	0.84	-0.17	0.07	.02	0.73-0.98
Pseudo R ²	0.02					0.00					0.02				
Adjusted Pseudo R ²	0.02					0.00					0.02				
Δ Pseudo R ²	0.00					0.00					0.00				

Abbreviation: IRR, incidence rate ratio; B, beta; SE, standard error.

effects for communal words were statistically significant such that women trainees used significantly fewer communal words when describing women faculty (IRR=0.84; 95% CI, 73-98; $P<.05$). Additionally, women trainees used more words per assessment on average (25.72 vs 20.50 words, respectively).

FIGURE 1 indicates that n-grams like “trust” (OR=11.95; 95% CI, 2.27-63.04), “bedside manner” (OR=7.48; 95% CI, 1.69-33.01), “guidance” (OR=3.88; 95% CI, 1.42-10.55), and “constructive feedback” (OR=2.01; 95% CI, 1.04-3.89) were significantly associated with assessments from women trainees. In contrast, n-grams including “professionalism” (OR=11.17; 95% CI, 2.05-60.72), “insight” (OR=5.28; 95% CI, 1.13-24.73), “teach procedure” (OR=3.24; 95% CI, 1.06-9.88), “fun” (OR=2.75; 95% CI, 1.49-5.08) and “teacher mentor” (OR=2.59; 95% CI, 1.26-5.35) were significantly associated with assessments from men trainees.

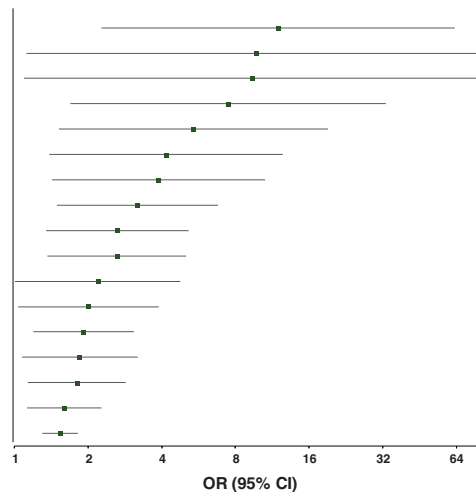
FIGURE 2 indicates that n-grams such as “opportunity teach” (OR=9.91; 95% CI, 1.27-77.50), “perfect” (OR=6.61; 95% CI, 1.24-35.12), “relate” (OR=2.25; 95% CI, 1.07-4.72) and “supportive” (OR=1.68;

95% CI, 1.03-2.76) were significantly associated with assessments of women faculty. In contrast, words like “operate” (OR=2.43; 95% CI, 1.26-4.70), “teacher mentor” (OR=2.05; 95% CI, 1.05-4.00), and “evidence base” (OR=1.78; 95% CI, 1.11-2.87) were significantly associated with assessments of men faculty.

TABLE 3 indicates that overall teaching effectiveness ratings correlated positively with standout and communal word frequencies for both men and women faculty ($r_s=0.29$ and 0.28 , $P<.001$), and ($r_s=0.23$ and 0.22 , $P=.003$ and $.01$), respectively. Role modeling ratings correlated positively with agentic/ability and standout word frequencies for both men and women faculty ($r_s=0.22$ and 0.26 , $P=.005$ and $.001$) and ($r_s=0.15$ and 0.30 , $P=.05$ and $<.001$), respectively, and with communal word frequencies for men faculty ($r_s=0.21$, $P=.01$). Teaching procedures ratings correlated positively with standout and communal word frequencies for both men and women faculty ($r_s=0.19$ and 0.16 , $P=.01$ and $.04$) and ($r_s=0.18$ and 0.19 , $P=.02$ and $.02$), respectively. Facilitating knowledge acquisition ratings correlated negatively with communal word frequency for men faculty ($r_s=-0.16$, $P=.04$).

A. N-gram associations with women trainees

N-gram	OR (95% CI)
trust	11.95 (2.27-63.04)
wonderful job	9.77 (1.12-84.94)
efficiency	9.41 (1.10-80.72)
bedside manner	7.48 (1.69-33.01)
smart	5.38 (1.52-19.12)
teach round	4.17 (1.39-12.49)
guidance	3.88 (1.42-10.55)
create	3.18 (1.50-6.77)
incredible	2.64 (1.35-5.15)
incredibly	2.62 (1.37-5.02)
job teach	2.2 (1.01-4.78)
constructive feedback	2.01 (1.04-3.89)
wonderful	1.92 (1.19-3.07)
explain	1.85 (1.07-3.2)
love	1.8 (1.13-2.86)
approachable	1.6 (1.13-2.27)
teach	1.54 (1.30-1.82)



B. N-gram associations with men trainees

N-gram	OR (95% CI)
professionalism	11.17 (2.05-60.72)
insight	5.28 (1.13-24.73)
teach procedure	3.24 (1.06-9.88)
phenomenal	2.84 (1.11-7.30)
fun	2.75 (1.49-5.08)
teacher mentor	2.59 (1.26-5.35)
greatly	2.33 (1.05-5.15)
knowledge	1.59 (1.11-2.29)

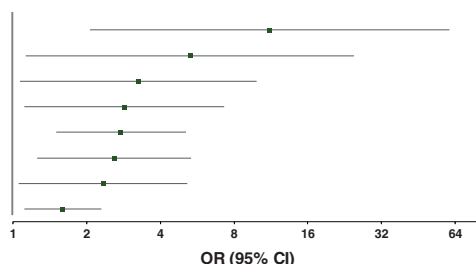


FIGURE 1
Significant N-Gram Associations by Trainee Gender

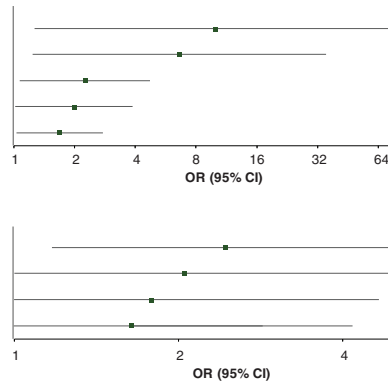
Note: Only n-grams significant at $P<.05$ were included. An exponential axis with log2 was used to better visualize effects. Abbreviations: OR, odds ratio; CI, confidence interval.

A. N-gram associations with women faculty

N-gram	OR (95% CI)
opportunity teach	9.91 (1.27-77.50)
perfect	6.61 (1.24-35.12)
relate	2.25 (1.07-4.72)
efficient	1.98 (1.02-3.87)
supportive	1.68 (1.03-2.76)

B. N-gram associations with men faculty

N-gram	OR (95% CI)
operate	2.43 (1.26-4.70)
teacher mentor	2.05 (1.05-4.00)
evidence base	1.78 (1.11-2.87)
autonomy	1.64 (1.07-2.52)

**FIGURE 2****Significant N-Gram Associations by Faculty Gender**

Note: Only n-grams significant at $P < .05$ were included. An exponential axis with log₂ was used to better visualize effects. Abbreviations: OR, odds ratio; CI, confidence interval.

A significant gender difference was observed in the correlation between role modeling ratings and communal word frequency for men and women faculty ($r_s = 0.21$ and 0.00 , $P = .010$ and $.99$; $z = 1.927$, $P = .027$).

Discussion

In this mixed-methods study, we examined differences in faculty assessments, analyzing both trainee descriptions (rater effect) and how faculty were described (ratee effect) using agentic/ability, standout, and communal categories. Women trainees used more of these descriptive words than men trainees, while women faculty received fewer agentic/ability words than men faculty. Standout and communal word frequency predicted teaching effectiveness ratings for both men and women faculty.

This study contributes to the literature by revealing gender-based differences in how men and women trainees assess faculty (rater effect). Women trainees incorporated a broader range of descriptive words, which may reflect gender-based differences in conscientiousness, attention to detail, and socialization.^{30,31} Women are more likely to emphasize relational dynamics, which could explain greater use of communal, agentic, and standout language in assessments.^{15,22} Women trainees also used fewer communal descriptors for women faculty—an unexpected result requiring further study. This may reflect sociocultural expectations, where communal traits are presumed in women but rewarded in men. Recognizing these distinct evaluative schemas can help training programs interpret assessments more accurately and support diverse evaluative priorities.

Our analysis of n-gram use further highlights these distinctions. Women trainees emphasized relational elements such as trust and bedside manner, whereas

men trainees focused more on knowledge and technical skills. For example, women trainees were 11.95 times more likely to use “trust,” while men trainees were 11.17 times more likely to use “professionalism.” These patterns underscore the distinct evaluative lenses between genders.

This study also highlights potential implicit gender bias in narrative comments (ratee effect),⁹⁻¹² which may contribute to women faculty underrepresentation in academic medicine.^{6,32} Women faculty received fewer agentic/ability descriptors, which does not necessarily indicate bias—men faculty may have exhibited more agentic behaviors. However, prior studies suggest no gender differences in knowledge or skill, challenging this assumption.³³⁻³⁵ The tendency to associate men faculty with agentic qualities and women faculty with communal traits may reinforce gender stereotypes, affecting hiring, promotion, and recognition.

Additionally, this study highlights qualities associated with effective educators. The positive correlation between communal and standout word frequency with teaching effectiveness suggests that supportive, relationship-oriented behaviors are valued in teaching. However, this does not imply causation. Faculty development programs could promote these humanistic qualities through training on empathy, active listening, and responsiveness.

Agentic/ability words are positively associated with role modeling, reinforcing the importance of clinical knowledge as a key component of effective role modeling.³⁶ However, these words show a modest negative association with teaching procedures, suggesting that procedural instruction may rely more on different competencies, such as patience and clarity. Faculty training should explicitly differentiate the skills essential for role modeling (eg, expertise, decisiveness) from

TABLE 3
The Correlation Between Word Frequencies and Teaching Effectiveness Ratings Separately by Faculty Gender

Faculty Gender	Agentic/Ability				Standout				Communal			
	Overall Teaching Effectiveness	Role Modeling	Teaching Procedures	Facilitating Knowledge Acquisition	Overall Teaching Effectiveness	Role Modeling	Teaching Procedures	Facilitating Knowledge Acquisition	Overall Teaching Effectiveness	Role Modeling	Teaching Procedures	Facilitating Knowledge Acquisition
Men, <i>r</i>	0	0.22 ^a	-0.11	0.03	0.29 ^a	-0.15 ^a	0.19 ^a	-0.10	0.23 ^a	0.21 ^a	0.18 ^a	-0.16 ^a
Women, <i>r</i>	0.08	0.26 ^a	-0.08	-0.12	0.28 ^a	0.30 ^a	0.16 ^a	-0.16	0.22 ^a	0	0.19 ^a	0.02

^a *P* value < .05.

those needed for procedural instruction (eg, patience, clarity) to ensure balanced assessment.

Limitations

This study's single-institution scope limits generalizability, particularly as assessments were drawn primarily from specialties with higher women faculty representation (eg, family medicine, pediatrics), providing limited insight into fields that are dominated by men faculty, like surgery. Future research should include a broader specialty mix and examine faculty rank, length of service, trainee program, postgraduate year, and specialty. Due to data limitations, we could not control for rank or service length, and it remains unclear how specialty-specific gender distributions influence the rater effect. Analyses were restricted to assessments where both trainee and faculty disclosed gender, potentially introducing bias; future studies should compare those who disclose gender with those who do not. Anonymity prevented tracking whether the same trainees repeatedly evaluated faculty, possibly affecting standard error estimations. Additionally, gender identity data were limited to men and women, highlighting the need for further research on gender-diverse groups.

Conclusions

This study reveals gender-based patterns in faculty assessments. Women trainees used more agentic/ability, communal, and standout descriptors, while women faculty received fewer agentic/ability words than men faculty. Women trainees used fewer communal words for women faculty than men faculty. Standout and communal word frequency predicted teaching effectiveness ratings for both genders.

References

1. Babal JC, Webber S, Nacht CL, et al. Recognizing and mitigating gender bias in medical teaching assessments. *J Grad Med Educ*. 2022;14(2):139-143. doi:10.4300/JGME-D-21-00774.1
2. Vaughan B. Clinical educator self-efficacy, self-evaluation and its relationship with student evaluations of clinical teaching. *BMC Med Educ*. 2020;20(1):347. doi:10.1186/s12909-020-02278-z
3. Rosen AS. Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data. *Assess Eval Higher Educ*. 2018;43(1):31-44. doi:10.1080/02602938.2016.1276155
4. Jena AB, Olenki AR, Blumenthal DM. Sex differences in physician salary in US public medical schools. *JAMA*

- Intern Med.* 2016;176(9):1294-1304. doi:10.1001/jamainternmed.2016.3284
5. Richter KP, Clark L, Wick JA, et al. Women physicians and promotion in academic medicine. *N Engl J Med.* 2020;383(22):2148-2157. doi:10.1056/NEJMsa1916935
 6. Association of American Medical Colleges. Department chairs by medical school, department type, and gender, 2021. Accessed March 20, 2025. <https://www.aamc.org/media/9876/download?attachment>
 7. Mengel F, Sauermann J, Zölitz U. Gender bias in teaching evaluations. *J Euro Econ Assoc.* 2019;17(2):535-566. doi:10.1093/jea/ajvx057
 8. McOwen KS, Bellini LM, Guerra CE, Shea JA. Evaluation of clinical faculty: gender and minority implications. *Acad Med.* 2007;82(suppl 10):94-96. doi:10.1097/ACM.0b013e3181405a10
 9. Leone-Perkins M, Schnuth R, Kantner T. Preceptor-student interactions in an ambulatory clerkship: gender differences in student evaluations of teaching. *Teach Learn Med.* 1999;11(3):164-167. doi:10.1207/S15328015TL110307
 10. 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. doi:10.1089/jwh.2015.5475
 11. Fassiotto M, Hamel EO, Ku M, et al. Women in academic medicine: measuring stereotype threat among junior faculty. *J Womens Health (Larchmt).* 2016;25(3):292-298. doi:10.1089/jwh.2015.5380
 12. Cullen MJ, Zhou Y, Sackett PR, Mustapha T, Hane J, Culican SM. Differences in trainee evaluations of faculty by rater and ratee gender. *Acad Med.* 2023;98(10):1196-1203. doi:10.1097/ACM.0000000000005260
 13. Kreitzer RJ, Sweet-Cushman J. Evaluating student evaluations of teaching: a review of measurement and equity bias in SETs and recommendations for ethical reform. *J Acad Ethics.* 2022;20(1):73-84. doi:10.1007/s10805-021-09400-w
 14. Heffernan T. Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching. *Assess Eval Higher Educ.* 2022;47(1):144-154. doi:10.1080/02602938.2021.1888075
 15. Eagly AH. *Sex Differences in Social Behavior: A Social-Role Interpretation.* Erlbaum; 1987.
 16. Bhanvadia S, Radha Saseendrakumar B, Guo J, Daniel M, Lander L, Baxter SL. Evaluation of bias in medical student clinical clerkship evaluations using natural language processing. *Acad Med.* 2022;97(suppl 11):154. doi:10.1097/ACM.0000000000004807
 17. Rojek AE, Khanna R, Yim JW, 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
 18. 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
 19. Heath JK, Weissman GE, Clancy CB, Shou H, Farrar JT, Dine CJ. Assessment of gender-based linguistic differences in physician trainee evaluations of medical faculty using automated text mining. *JAMA Netw Open.* 2019;2(5):e193520. doi:10.1001/jamanetworkopen.2019.3520
 20. Engel-Rebitzer E, Kogan JR, Heath JK. Gender-based differences in language used by students to describe their noteworthy characteristics in medical student performance evaluations. *Acad Med.* 2023;98(7):844-850. doi:10.1097/ACM.0000000000005141
 21. Eagly AH. The his and hers of prosocial behavior: an examination of the social psychology of gender. *Am Psychol.* 2009;64(8):644-658. doi:10.1037/0003-066X.64.8.644
 22. Burgoon JK. Expectancy Violations Theory. In: Berger CR, Roloff ME, Wilson SR, Dillard JP, Caughlin J, Solomon D, eds. *The International Encyclopedia of Interpersonal Communication.* 1st ed. Wiley; 2015:1-9.
 23. Carnes M, Morrissey C, Geller SE. Women's health and women's leadership in academic medicine: hitting the same glass ceiling? *J Womens Health (Larchmt).* 2008;17(9):1453-1462. doi:10.1089/jwh.2007.0688
 24. Srinivasan M, Li STT, Meyers FJ, et al. "Teaching as a competency": competencies for medical educators. *Acad Med.* 2011;86(10):1211-1220. doi:10.1097/ACM.0b013e31822c5b9a
 25. Sutkin G, Wagner E, Harris I, Schiffer R. What makes a good clinical teacher in medicine? A review of the literature. *Acad Med.* 2008;83(5):452-466. doi:10.1097/ACM.0b013e31816bee61
 26. Hickman L, Thapa S, Tay L, Cao M, Srinivasan P. Text preprocessing for text mining in organizational research: review and recommendations. *Organ Res Method.* 2022;25(1):114-146. doi:10.1177/1094428120971683
 27. 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
 28. Ross DA, Boatright D, Nunez-Smith M, Jordan A, Chekroud 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
 29. Turrentine FE, Dreisbach CN, St Ivany AR, Hanks JB, Schroen AT. Influence of gender on surgical residency applicants' recommendation letters. *J Am Coll Surg.* 2019;228(4):356-365.e3. doi:10.1016/j.jamcollsurg.2018.12.020
 30. Costa PT, Terracciano A, McCrae RR. Gender differences in personality traits across cultures: robust

- and surprising findings. *J Pers Soc Psychol.* 2001;81(2): 322-331. doi:10.1037/0022-3514.81.2.322
31. Schmitt DP, Realo A, Voracek M, Allik J. Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *J Pers Soc Psychol.* 2008;94(1):168-182. doi:10.1037/0022-3514.94.1.168
 32. Association of American Medical Colleges. Table A-7.2: Applicants, First-Time Applicants, Acceptees, and Matriculants to U.S. MD-Granting Medical Schools by Sex, 2012-2013 through 2021-2022. Accessed March 20, 2025. <https://www.aamc.org/media/9576/download?attachment>
 33. Sulistio MS, Khera A, Squiers K, et al. Effects of gender in resident evaluations and certifying examination pass rates. *BMC Med Educ.* 2019;19(1):10. doi:10.1186/s12909-018-1440-7
 34. Ong TQ, Kopp JP, Jones AT, Malangoni MA. Is there gender bias on the American Board of Surgery General Surgery Certifying Examination? *J Surg Res.* 2019;237:131-135. doi:10.1016/j.jss.2018.06.014
 35. Driscoll SW, Robinson LR, Raddatz MM, Kinney CL. Is there evidence of gender bias in the oral examination for initial certification by the American Board of Physical Medicine & Rehabilitation? *Am J Phys Med Rehabil.* 2019;98(6):512-515. doi:10.1097/PHM.0000000000001126
 36. Passi V, Johnson S, Peile E, Wright S, Hafferty F, Johnson N. Doctor role modelling in medical education: BEME guide no. 27. *Med Teach.* 2013;35(9):e1422-e1436. doi:10.3109/0142159X.2013.806982



Jessica Hane, MD, is an Assistant Professor, Departments of Internal Medicine and Pediatrics, University of Minnesota Medical School, Minneapolis, Minnesota, USA; **Vivien Lee, BA**, is a Graduate Student, Department of Psychology, University of Minnesota-Twin Cities, Minneapolis, Minnesota, USA; **You Zhou, MA**, is a Graduate Student, Department of Psychology, University of Minnesota-Twin Cities, Minneapolis, Minnesota, USA; **Taj Mustapha, MD**, is an Associate Professor, Departments of Internal Medicine and Pediatrics, University of Minnesota Medical School, Minneapolis, Minnesota, USA; **Susan M. Culican, MD, PhD**, is Senior Associate Dean, Graduate Medical Education, University of Minnesota Medical School, Minneapolis, Minnesota, USA; **G. Nic Rider, PhD**, is an Associate Professor, Institute for Sexual and Gender Health, Department of Family Medicine and Community Health, University of Minnesota Medical School, Minneapolis, Minnesota, USA; **Paul R. Sackett, PhD**, is a Professor, Department of Psychology, University of Minnesota-Twin Cities, Minneapolis, Minnesota, USA; and **Michael J. Cullen, PhD**, is Senior Director of Assessment, Evaluation, and Research, Graduate Medical Education, University of Minnesota Medical School, Minneapolis, Minnesota, USA.

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

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

The authors would like to thank the program directors, program coordinators, and clinical competency committees from the 18 clinical departments that employed their measure of clinical teaching effectiveness during the study.

Corresponding author: Michael J. Cullen, PhD, University of Minnesota Medical School, Minneapolis, Minnesota, USA, cull0061@umn.edu, X@umngme

Received August 2, 2024; revisions received November 14, 2024, and February 8, 2025; accepted February 19, 2025.