2026, Volume 0, Issue 0, 1-8, e-ISSN 1938-3800

Family Medicine

ORIGINAL ARTICLE

Comparing Evaluations of Family Medicine Residents by Gender and Ethnicity

Zachary J. Kosak, MD[a]; Jenny Y. Wang, MD[a]; Janae K. Heath, MD, MSCE[b]; C. Jessica Dine, MD, MSHP[b]

AUTHOR AFFILIATIONS:

[a]Department of Family Medicine and Community Health, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

[b]Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA

CORRESPONDING AUTHOR:

Zachary J. Kosak, Department of Family Medicine and Community Health, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, zachary.kosak@pennmedicine.upenn.edu

HOW TO CITE: Kosak ZJ, Wang JY, Heath JK, et al. Comparing Evaluations of Family Medicine Residents by Gender and Ethnicity. *Fam Med*. 2026;0(0):1-8.
doi: 10.22454/FamMed.2026.0229

FIRST PUBLISHED: November 2, 2026

KEYWORDS: assessment, ethnicity, family practice, gender bias, medical education, racial groups

© Society of Teachers of Family Medicine

ABSTRACT

Background and Objectives: Faculty evaluations of resident physicians have previously identified qualitative and quantitative differences based on gender and ethnicity, but largely outside of the specialty of family medicine. This study sought to determine whether trends in evaluation biases from other specialties were also present in faculty evaluations of family medicine residents.

Methods: We performed qualitative and quantitative analysis on faculty evaluations of residents from academic years 2019–2020 through 2023–2024 from a single urban academic family medicine residency program site. We performed quantitative analyses to evaluate Accreditation Council for Graduate Medical Education (ACGME) milestone ratings, and we performed quantitative and qualitative text-based analyses of the comments using LIWC-22 and N-gram.

Results: Quantitative analyses demonstrated no significant differences in ACGME milestone ratings based on gender or underrepresented in medicine status. Qualitative analysis of evaluations showed some differences among ethnic groups. Asian residents were noted to have less feminine language while Asian and Hispanic residents were noted to have more communion language. The only gender/ethnicity interaction noted was that Asian women had less communion language present than their Asian men counterparts.

Conclusions: While no significant quantitative differences existed in resident evaluations based on gender and ethnicity, some qualitative differences were identified based on ethnicity. As family medicine moves into its competency-based medical education era, monitoring faculty evaluations for biases will be critical to ensure equitable determination of clinical competence.

INTRODUCTION

Ensuring accurate, objective, and unbiased evaluation and assessment of resident trainees is a critical element within graduate medical education programs. These evaluations are often conducted in relationship to the standards set forth in the Accreditation Council for Graduate Medical Education (ACGME) milestones for each training program.[1] Because resident evaluation and assessment data are critical to determine promotion, evaluate clinical competence, structure remediation, and determine whether resident physicians have met graduation targets, ensuring accuracy of assessments free from biases is paramount. To date, however, little is known about how trainee demographics impact evaluation within family medicine and whether biases of evaluation are present in this cohort.

In other specialties, qualitative and quantitative differences in evaluations have been well-described. Anesthesia residents who identify as women, for example, are less likely to be described with agentic qualities such as independence, competence, and efficacy early in their training when compared to men counterparts.[2] Similarly, literature has described differences in quantitative and qualitative evaluations of emergency medicine residents, highlighting that women are less likely to receive consistent feedback when struggling and are more likely to receive negative

feedback on traditionally masculine traits such as assertiveness and autonomy.[3,4] Quantitatively, literature from general surgery training programs has demonstrated lower milestone assessments across nearly all subcompetencies among women residents, although both men and women residents achieved scores consistent with competence for transition to independent practice prior to graduating residency.[5] In internal medicine training, numeric ratings of women residents have demonstrated a peak-and-plateau pattern after postgraduate year (PGY) 2, whereas men residents continued to show quantitative evaluation improvements.[6] Similar findings have been noted based on self-identified race, with data suggesting that residents identifying as ethnic minorities and women receive lower quantitative ratings on ACGME milestone domains when compared to their White men counterparts.[7,8]

The influence of residency evaluations on the postgraduation careers of physicians cannot be understated. Literature has suggested that residents who have action taken for performance concerns have lower graduation rates than peers.[9] If discrepancies exist among residents based on gender or ethnicity, they could have a substantial impact on the career trajectories of residents from ethnic minority backgrounds. While prior work has focused on other specialties, little is known whether similar trends exist in family medicine residency training.[2,3,6,10] This possible trend is particularly important in family medicine, where a plurality of physicians in residency are women.[11] Prior work has identified that women residents may be more likely to receive more feedback on communication skills than men counterparts, often in relation to opportunities for improvement for the resident.[5] With the emphasis on continuity relationships and outpatient visits in family medicine, women residents may be disproportionately impacted if this trend occurs in family medicine.

To better understand whether similar differences exist in family medicine residency training, we evaluated demographic differences in quantitative and qualitative evaluations of residents based on identified gender and ethnicity. We hypothesized that quantitative milestone ratings and qualitative language use—including agentic, communal, feminine, and masculine language—would differ based on resident gender and ethnicity, consistent with findings from other specialties. We explored whether other linguistic features, including hedging language, differed across demographic groups.

METHODS

Study Design and Data Source

We conducted a retrospective observational study of milestone evaluations from a single urban academic family medicine residency program. Data included all submitted evaluations and milestone-level assessments completed between June 2020 and November 2024. Resident evaluations are submitted and stored via a centralized residency management system (MedHub). During the study period, the residency program increased its size from 27 to 36 residents across three postgraduate years. Evaluations were completed by supervising faculty from family medicine and other specialties participating in required residency rotations, including inpatient, outpatient, and obstetrics experiences, as well as following outpatient clinic precepting sessions. These evaluations were compiled biannually at clinical competency committee (CCC) meetings where consensus ACGME milestone evaluations were generated. The unit of analysis was the resident for each academic year, with residents contributing repeated observations across multiple academic years as applicable. Resident demographic variables included self-described gender and ethnic status; descriptive statistics summarized resident demographics, distribution across academic years and PGY levels, and evaluation counts. The demographic makeup of the residents is outlined in Table 1.

Quantitative Analysis of ACGME Milestone Differences

For analyzing quantitative differences in milestone ratings, the primary outcomes were milestone scores for six core competencies: interpersonal and communication skills (ICS), medical knowledge (MK), practice-based learning and improvement (PBLI), patient care (PC), professionalism (PROF), and systems-based practice (SBP). Milestone scores were treated as continuous variables. Covariates were selected a priori based on conceptual relevance to milestone assessment: PGY (modeled as a continuous variable to represent developmental progression), gender, ethnicity, evaluation exposure (defined as the number of evaluations completed within the same academic year), and academic year (included as a categorical fixed effect to account for cohort-level and system-level differences across calendar years). To account for differences in residents' duration in the study, we used the Kruskal-Wallis test to compare the mean

TABLE 1. Resident Evaluations and Demographics by Academic Year

	2019–2020	2020–2021	2021–2022	2022–2023	2023–2024	Totals
Total	28	25	29	32	36	
Women residents	22	21	24	26	27	57
Men residents	6	4	5	9	8	16
Black, Asian, or Hispanic residents	5	9	13	12	12	23
Mean number of evaluations per resident (SD)	7.93 (2.32)	7.80 (1.16)	6.62 (1.61)	7.44 (1.05)	4.44 (1.50)	16.60 (6.62)

Abbreviation: SD, standard deviation

annual number of evaluations received per resident across ethnic groups.

We conducted unadjusted comparisons of milestone scores by gender and ethnicity using two-sample *t* tests, both pooled across years and stratified by academic year and PGY. Primary analyses used linear mixed-effects regression models fit separately for each competency. Models included fixed effects for PGY, gender, ethnicity, evaluation exposure within the academic year, and academic year, with a random intercept for resident to account for repeated observations. Likelihood ratio tests compared mixed-effects models to corresponding linear models without random effects.

Qualitative Text Based Analysis Using LIWC-22 and N-Gram Analysis

We used LIWC-22 (linguistic inquiry and word count; Pennebaker Conglomerates Inc) to analyze narrative comments for each submitted evaluation during the study period.[12,13] This tool allows for quantitative analysis of written language—providing metrics like word count while also quantifying the presence of specific words within previously developed dictionaries. Based on prior literature analyzing gender differences in narrative evaluation using LIWC-22,[14] we assessed for differences in the use of "masculine" terms, "feminine" terms, "agentic" terms (terms that connotate an individual's ability), and "communal" terms (terms that highlight an individual's desire to relate and cooperate) using pre-developed LIWC-22 dictionaries.[15-18] Using the LIWC-22 score for each domain, we evaluated summary statistics of the domains by resident gender, resident ethnicity, and PGY.

In addition to LIWC-22 analysis, we analyzed word use through n-gram analysis,[19] a natural language processing method that converts free-text comments into discrete words and phrases. N-gram analysis allows creation of a dataset indicating the presence or absence of single words (unigrams, eg, "clinician" or "brilliant"). Using previously published research on hedging language in evaluation,[20] we identified and categorized the unigrams into hedging language to explore potential differences by resident gender and ethnicity (Table 2). Using this category as a binary variable (present vs absent hedging language), we evaluated summary statistics of the hedging language by resident gender, ethnicity, and PGY. Similar to work that has been done in other settings, we evaluated specific terms of confidence (including words such as "confident," "confidence," or "confidently").[21]

We then performed mixed-effects linear regression analyses to assess the association of the LIWC-22 categories (agentic, communal, masculine, feminine, word count) with resident gender and ethnicity, adjusted for resident PGY, clustered on rater (random effects). We also performed mixed-effects linear regression analyses to assess the association of the hedging language with resident gender and ethnicity, adjusted for PGY, clustered on rater (random effects). For all analyses, statistical significance was determined using a *P* value of .008 (using a Bonferroni correction for multiple comparisons). All statistical analyses and additional text analysis were completed using Stata 19.5 (StataCorp) and LIWC-22.

This project was reviewed and determined to qualify as quality improvement by the University of Pennsylvania's Institutional Review Board.

RESULTS

The dataset included 152 resident–academic year observations from 73 residents (79% women) and 3,184 total evaluations completed by 67 unique faculty. Residents received a mean of 16.6 (SD 6.6) faculty evaluations during the study period, with the number varying by academic year (Table 1). Of the rotation evaluations, 74% were in reference to women and 34% were of residents classified as ethnic minorities (Black, Asian, and/or Hispanic). The mean number of evaluations received per resident per academic year did not differ significantly by ethnicity (Kruskal-Wallis P = 0.1). Residents were evenly distributed across PGY-1 through PGY-3, with few PGY-4 observations. The evaluations of the cohort were predominantly of Caucasian individuals (36% total evaluations, n = 1,151), followed by Black (29%, n = 923), Asian (26%, n = 835), and Hispanic (9%, n = 275). When pooled across academic years, mean milestone scores did not differ significantly by gender or ethnicity for any of the six competencies. Stratified analyses by academic year and PGY identified some statistically significant differences in individual years; however, these findings were inconsistent in direction and based on small subgroup sizes, limiting interpretability. As such, adjusted mixed-effects models were prioritized for inference.

Mixed-Effects Regression Analyses of Milestone Ratings

Across all six competencies, PGY was the strongest and most consistent predictor of milestone scores. Each increase in PGY was associated with an approximately 0.8–0.9-point increase in milestone level (all P<0.001), consistent with expected developmental progression. The number of evaluations completed within the same academic year was positively associated with milestone scores for several competencies. Significant associations were observed for ICS, MK, and PROF, with marginal associations for PC and SBP, while no significant association was observed for PBLI. After adjustment for PGY, evaluation exposure, and academic year, gender and ethnicity were not independently associated with milestone scores for any competency. Academic year fixed effects were generally small and not statistically significant, indicating minimal systematic shifts in milestone scoring across calendar years after accounting for PGY and evaluation exposure. All mixed-effects models demonstrated significant resident-level random intercept variance, indicating meaningful between-resident differences in baseline milestone performance. Likelihood ratio tests confirmed that mixed-effects models provided a significantly better fit than models without random effects (all P<0.001).

TABLE 2. Words or Phrases Indicative of Language in Evaluations

Agentic words	Communal words	Feminine words	Hedging words	Masculine words
able, accomplish*, accurac*, accurate*, achiev*, acquir*, actualiz*, adaptab*, adept*, ambition*, ambitious*, aptitude*, aptly, aptness, aspiration*, aspire*, aspiring, assert*, attain*, authoritative*, autonomous*, autonomy, capab*, careful*, choice, choices, clever*, compet*, completion, confident, confidently, conquer*, conscientious*, contemplat*, contend*, contest*, decid*, decision*, decisive*, defeat*, deliberat*, dependable, determin*, difficult*, do, doable, doing, eager*, earn, earned, earning, earns, easiness, easy, effective*, efficien*, effort*, empowered, enact*, endeavor*, establish, established, establishes, establishing, exact*, expert*, fail*, fluen*, freedom*, freely, goal, goal-oriented, goals, importan*, independ*, individualist, insight*, intent*, intuition, intuitive*, keen*, know*, liberties, liberty, logic*, loner*, made, make, makes, making, mastered, masterful*, mastering, mastery, motivat*, need, needed, needing, needs, objectiv*, obtain*, opportun*, overcame, overcome, overcomes, overcoming, persever*, persist*, persistent, pioneer*, practic*, pragmat*, prevail*, pride, prideful*, priorit*, proactive*, productive*, productivity, proficien*, prosper*, proud*, purpose*, pursu*, rational*, realiz*, rebel*, recog*, reliab*, reputation*, resilien*, resolute*, resolution, resolv*, responsib*, reward*, risk*, savv*, score, scored, scores, scoring, self, self-*, should*, significant*	accept*, accommodat*, accompan*, accordaffab*, affection*, affiliat*, affinity, agree*, aid, aided, aiding, allegian*, alliance*, allienally, altruis*, amenab*, amiab*, amicab*, amigo*, apolog*, appreciat*, assist*, benevolen*, buddies, buddy, care, cared, cares, caring*, ceremon*, charit*, chat, chats, chatted, chatting, civic, civil*, closeness, collab*, colleague*, collective*, common good, common, ground*, commun*, companion*, compassion*, compromis*, concert*, confer*, congrat*, consen*, considerate*, contribut*, conversat*, converse, conversed, converses, conversing, cooperat*, counsel*, courteous*, crew, democr*, dialogue*, discuss*, educat*, empath*, equitab*, familial, families, family, familys', family's, fellow*, festiv*, forgave, forgiv*, frat*, friend*, generos*, generous*, grateful*, gregarious*, group*, guidanc*, harmon*, help*, honest*, hospitab*, hospitality, human*, impartial*, interpersonal*, intima*, justice, justly, kin, kindly, kindness, kinship, law, lawful*, laws, learn*, love*, loving*, loyal*, magnanimous*, marriag*, matrimon*, member*, middle, ground, mingl*, mutual*, negotiat*, neighbor*, nurtur*, offer*, oneness, open-mind*, our, ours, pal, pals, participat*, partied, parties, partner*, party*, philanth*, pluralis*, polite*, public, publicly, recipr*, recommend*, reconcil*, relationship*, request*, respect*, ritual*, roommate*, sacrific*, selfless*, servic*, share, shared, shares, sharing,	affectionate, child, cheer, commit, communal, compassion, connect*, considerate, cooperat*,depend*, emotiona*, empath*, feminine, flatterable, gentle, honest, interpersonal, interdependen*, interpersona*, kind, kinship, loyal, modesty, nag, nurtur*, pleasant, polite, quiet, respon*, sensitiv*, submissive, support*, sympath*, tender, together, trust, understand, warm, whin*, depend*, emotiona*, empath*, feminine, flatterable, gentle, honest, interpersonal, interdependen*, interpersona*, kind, kinship, loyal, modesty, nag, nurtur*, pleasant, polite, quiet, respon*, sensitiv*, submissive, support*, sympath*, tender, together, trust, understand, warm, whin*, yield	about, almost, apparently, appear, appeared, appearing, appears approximately, arguably, around, broadly, essentially, generally, implied, implies, indicated, indicates, kind of, likely, maybe, often, ostensibly, perhaps, plausably, possibly, presumably, probably, relatively, roughly, seemed, seemingly, seems, sometimes, sort of, suggest, suggested, suggests, supposedly, tend, tended, tends typically, unlikely, usually, virtually	active , adventurous, aggress*, ambitio*, analy*, assert, athlet*, autonom*, boast, challeng*, compet*, confident, courag*, decide, decisive, decision, determin*, dominant, domina*, force, greedy, headstrong, hierarch*, hostil, impulsive, independen*, individual, intellect*, lead, logic, masculine, objective, opinion, outspoken, persist, principle, reckless, stubborn, superior, self-confiden*, self-sufficien*, self-relian*

(*Continued*)

https://doi.org/10.22454/FamMed.2026.0229

TABLE 2. Words or Phrases Indicative of Language in Evaluations (Continued)

Agentic words	Communal words	Feminine words	Hedging words	Masculine words
skill, skilled, skillful*, skills*, smart, smartly, steadfast*, strive*, striving*, struggl*, stubborn*, succeed*, success*, sure, take, takes, taking, tenac*, think*, thought, took, tried, tries, triumph*, trying, unaided, unyielding*, vanquish*, victor*, will, willing*, willpower, win, winner*, winning*, wins, wit, wits, witting*, won, won't, you, you'*, your, yours, yourself,	sincer*, socia*, societ*, solidarity, soror*, squad*, suggest*, support*, sympath*, talk*, taught, teach*, team*, thank*, together*, tradition*, treatise*, treaty, tribe*, trust*, truth*, unanim*, unif*, unite*, uniting, unity, unselfish*, us, volunt*, we, welcom*, welfare inclusiv*			

Note: The asterisk indicates the acceptance of all letters, hyphens, or numbers following its appearance.

Mixed-Effects Regression Analyses of Qualitative Comment Data

In analysis of the qualitative comments submitted in each evaluation, we found no significant difference in the use of masculine language (0.45 vs 0.41, $P = 0.55$), feminine language (0.85 vs 0.60, $P = 0.05$), agency (4.63 vs 4.60, $P = 0.89$), or communion (4.70 vs 4.69, $P = 0.97$) between evaluations of men and women. We found no difference in word count in evaluations between men and women residents (45.5 vs 46.3, $P = 0.75$). We found no difference in use of hedging language between men and women ($P = 0.59$), nor any difference in use of words describing confidence between men and women ($P = 0.34$).

In the univariable analyses examining the main effects of ethnicity, several differences in language use were observed (Table 3). We found significant word count differences based on ethnicity, where Hispanic residents had fewer words per evaluation than Caucasian residents (36.5 vs 47.3, $P = 0.008$). We found no significant differences in use of masculine language or agency language by ethnicity. We found significantly less use of feminine language in evaluations of Asian residents ($\beta = -0.31$, $P = 0.03$) compared to Caucasian residents. We also found differences in communion language, with more language of communion in evaluation of Asian residents ($\beta = 1.10$, $P = 0.001$) and Hispanic residents ($\beta = 1.72$, $P<0.001$).

These main effects were then evaluated in adjusted mixed effects models (Table 4). We found no significant differences by gender or ethnicity in use of masculine terms, agency language, hedging language, or confidence language. However, the adjusted analyses continued to demonstrate significantly less feminine language in evaluations of Asian residents ($\beta = -0.27$, $P = 0.03$) after adjustment for resident gender and clustering by resident and faculty member. Similarly, more communal language was observed in evaluations of Asian residents ($\beta = 1.14$, $P = 0.008$) and Hispanic residents ($\beta = 1.74$, $P = 0.007$).

We next examined whether the association between ethnicity and qualitative language differed by resident gender (ie, gender-by-ethnicity interaction effects). We observed no significant interactions for masculine, feminine, agency, hedging, or confidence language. However, we identified a significant interaction for communal language such that Asian women received fewer communal terms than Asian men ($\beta = -1.65$, $P = 0.03$).

DISCUSSION

This study was the first to analyze qualitative and quantitative evaluations of family medicine residents, looking for differences in faculty-provided milestone ratings as well as in language used in narrative comments with submitted faculty evaluations. Our findings suggest that, while quantitative evaluations do not demonstrate differences in

TABLE 3. Adjusted Coefficient for Use of Language Types by Resident Ethnicity in Resident Evaluations (95% CI)

	Masculine language β (*P*)	Feminine language β (*P*)	Agency language β (*P*)	Communion language β (*P*)
Asian	0.07 (−0.09–0.23)	−0.30 (−0.59–0.03)[a]	0.01 (−0.51–0.53)	1.10(0.47–1.74)[b]
Black	0.07 (−0.09–0.23)	−0.20 (−0.46–0.07)	0.41 (−0.10–0.92)	0.38 (−0.24–1.00)
Hispanic	0.12 (−0.11–0.47)	0.17 (−0.23–0.58)	0.13 (−0.64–0.90)	1.72(0.78–2.66)[b]

[a]$P<.05$
[b]$P<.01$
Abbreviation: CI, confidence interval

TABLE 4. Multivariate Linear Regression of Communal Term Language Use by Resident Ethnicity in Resident Evaluations

	Adjusted coefficient (95% CI)	*P* value
Asian	1.13(0.32–1.95)	.006[a]
Black	0.42 (−0.35–1.19)	.283
Hispanic	1.73(0.55–2.92)	.004[a]

[a]*P*<.008
Abbreviation: CI, confidence interval

milestone ratings, qualitative evaluations of residents may have some word use differences that are associated with gender and/or ethnicity.

Interestingly, we found no significant differences in quantitative ratings of residents and few qualitative differences across gender or ethnicity. These results are discrepant from prior work in other specialties that have highlighted differences.[3–8] Faculty development bringing awareness to and mitigation of evaluation biases may have reduced discrepancies in quantitative and qualitative evaluation data based on ethnicity or gender. Possibly family medicine faculty training is robust at providing structured and objective evaluative feedback for residents, thus mitigating biases seen in prior studies. Attempts to mitigate bias through faculty development on evaluation best practices (both through professional development by national or regional organizations and the institution itself) also contribute to avoiding biases in qualitative or quantitative evaluation metrics. Additionally, individual institutional CCC structures that could identify discrepant ratings may contribute to these findings. We would be surprised to find, however, that any individual institutional structure captures and mitigates subtle differences in evaluations that would appear over larger sample sizes.

As one exception, we found two differences within data from Asian residents. More communal language was noted in evaluations of Asian women when compared to Asian men. This finding suggests that observed differences in communal language vary based on gender (ie, Asian women have less communal language in evaluations than Asian men in this domain, which is not observed in other races). Additionally, we found that feminine terms were less used in evaluations of Asian residents. Interestingly, this finding did not result in quantitative milestone ratings of Asian residents (as previously reported in internal medicine residents).[22] These differences are notable because Asian women in academic medicine have identified stereotyping, feeling othered, and institutional culture as barriers to entry into academic medicine leadership.[23]

Possibly our findings among Asian residents are a demonstration of these previously identified themes and representative of barriers to leadership opportunity and growth for Asian women residents. These differences are interesting to consider in the context of the model minority myth (MMM), which downplays structural inequities that Asians face in the United States given a perceived high degree of educational and societal success.[24] While the MMM does not fully apply to the gender differences in communal language given that these differences occur within Asian residents alone, the decrease in feminine terminology used in qualitative evaluations of Asian residents is compatible with the stereotypes perpetuated by the MMM. The presence of qualitative differences without quantitative milestone rating discrepancies may be reflective of faculty knowledge of standard progression through milestone ratings throughout residency, while discrepancies in terminology used in qualitative evaluations may indicate a true difference in the evaluation of Asian residents. Also notable is that faculty who do not participate in the CCC process complete resident evaluations, allowing for the possibility that CCC discussion helps to blunt any effects seen by biased comments or evaluations. Finally, some literature has suggested that faculty evaluators may be more reliant on individual writing styles, and style differences may be more impactful on differences in evaluations rather than an experience with an individual resident.[25]

Our study had several limitations. First, our study was limited to a single site with specific institutional training regarding learner evaluation. As a result, the data may not reflect family medicine faculty evaluations on a larger scale. Additionally, our study period notably came just prior to the emphasis among family medicine professional organizations on implementation of competency-based medical education (CBME) through entrustability and robust frequent observations.[26] Possibly biases could be magnified or eliminated with a different approach to evaluation than less frequent and higher-stakes observations and evaluations.

Another limitation included a lack of nonbinary identifying residents in the study population. Our residency management software uses self-identified gender based on residency application data and may not capture all nuance in understanding resident gender identity. Further research including nonbinary identifying individuals and more nuanced data capture regarding gender identity may be useful to detect differences among nonbinary residents compared to women or men counterparts.

Although our analyses accounted for individual evaluator effects using mixed-effects models clustered by rater, we did not examine evaluator demographic characteristics or evaluator-resident gender or ethnicity concordance. We cannot exclude the possibility that some observed differences in language use were influenced by evaluator demographics rather than resident characteristics alone. Finally, we used a narrow definition of comments regarding resident confidence by using words similar to confidence itself. A broader definition could exist that would more fully capture comments referring to resident confidence, but that definition would need further research and validation.

CONCLUSIONS

Further research among multiple family medicine residency programs would help determine whether the effects we saw are broadly applicable across family medicine graduate medical education at-large and, if so, provide an avenue to

develop training for faculty to mitigate these differences. Additionally, research comparing pre-CBME implementation data with post-CBME implementation data could help to identify unintended deleterious effects of CBME evaluation principles on residents based on gender or ethnic characteristics. Finally, given the explosion of artificial intelligence (AI), one might plausibly expect residency faculty to interact with AI to summarize and generate narrative feedback. Further study regarding the impacts of AI in generating feedback for resident learners and the potential biases that could be introduced will be critical to ensure that fair and equitable evaluations inform resident progression and growth.

REFERENCES

1. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366(11):1051–1056. doi:10.1056/NEJMsr1200117
2. Arkin N, Lai C, Kiwakyou LM, et al. What's in a word? qualitative and quantitative analysis of leadership language in anesthesiology resident feedback. *J Grad Med Educ*. 2019;11(1):44–52. doi:10.4300/JGME-D-18-00377.1
3. Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM. Gender differences in attending physicians' feedback to residents: a qualitative analysis. *J Grad Med Educ*. 2017;9(5):577–585. doi:10.4300/JGME-D-17-00126.1
4. 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 Intern Med*. 2017;177(5):651–657. doi:10.1001/jamainternmed.2016.9616
5. Landau SI, Syvyk S, Wirtalla C, et al. Trainee sex and Accreditation Council for Graduate Medical Education milestone assessments during general surgery residency. *JAMA Surg*. 2021;156(10):925–931. doi:10.1001/jamasurg.2021.3005
6. Klein R, Ufere NN, Rao SR, et al; Gender Equity in Medicine Workgroup. Association of gender with learner assessment in graduate medical education. *JAMA Netw Open*. 2020;3(7):e2010888. doi:10.1001/jamanetworkopen.2020.10888
7. Lett E, Tran NK, Nweke N, et al. Intersectional disparities in emergency medicine residents' performance assessments by race, ethnicity, and sex. *JAMA Netw Open*. 2023;6(9):e2330847. doi:10.1001/jamanetworkopen.2023.30847
8. Amin BZ, Dine CJ, Tabakin ER, Trotter M, Heath JK. Inequities in the national clinical assessment tool for medical students in the emergency department. *West J Emerg Med*. 2025;26(5):1250–1259. doi:10.5811/westjem.43506
9. Turner JA, Fitzsimons MG, Pardo MC Jr, et al. Effect of performance deficiencies on graduation and board certification rates: a 10-yr multicenter study of anesthesiology residents. *Anesthesiology*. 2016;125(1):221–229. doi:10.1097/ALN.0000000000001142
10. Klein R, Julian KA, Snyder ED, et al; Gender Equity in Medicine Workgroup. Gender bias in resident assessment in graduate medical education: review of the literature. *J Gen Intern Med*. 2019;34(5):712–719. doi:10.1007/s11606-019-04884-0
11. Association of American Medical Colleges. Table B3. number of active residents by type of medical school, GME specialty, and gender: 2024-25 residents. Accessed March 26, 2026. https://www.aamc.org/data-reports/students-residents/data/report-residents/2025/table-b3-number-active-residents-type-medical-school-gme-specialty-and-gender
12. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW. The Development and Psychometric Properties of LIWC-22. The University of Texas at Austin; 2022. https://www.liwc.app/static/documents/LIWC-22%20Manual%20-%20Development%20and%20Psychometrics.pdf
13. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol. 2010;29(1):24–54. doi:10.1177/0261927X09351676
14. Heath JK, Kogan JR, Holmboe ES, Conforti L, Park YS, Dine CJ. Gender differences in work-based assessment scores and narrative comments after direct observation. J Gen Intern Med. 2024;39(10):1795–1802. doi:10.1007/s11606-024-08645-6
15. Madera JM, Hebl MR, Martin RC. Gender and letters of recommendation for academia: agentic and communal differences. *J Appl Psychol*. 2009;94(6):1591–1599. doi:10.1037/a0016539
16. Trix F, Psenka C. Exploring the color of glass: letters of recommendation for female and male medical faculty. Discourse Soc. 2003;14(2):191–220. doi:10.1177/0957926503014002277
17. Gaucher D, Friesen J, Kay AC. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *J Pers Soc Psychol*. 2011;101(1):109–128. doi:10.1037/a0022530
18. Pietraszkiewicz A, Formanowicz M, Gustafsson Sendén M, Boyd RL, Sikström S, Sczesny S. The big two dictionaries: capturing agency and communion in natural language. *Euro J Social Psych*. 2019;49(5):871–887. doi:10.1002/ejsp.2561
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. Ginsburg S, van der Vleuten C, Eva KW, Lingard L. Hedging to save face: a linguistic analysis of written comments on in-training evaluation reports. *Adv Health Sci Educ Theory Pract*. 2016;21(1):175–188. doi:10.1007/s10459-015-9622-0
21. Heath JK, Alvarado ME, Clancy CB, Barton TD, Kogan JR, Dine CJ. The context of "confidence": analyzing the term confidence in resident evaluations. *J Gen Intern Med*. 2022;37(9):2187–2193. doi:10.1007/s11606-022-07535-z
22. Boatright D, Anderson N, Kim JG, et al. Racial and ethnic differences in internal medicine residency assessments. *JAMA Netw Open*. 2022;5(12):e2247649. doi:10.1001/jamanetworkopen.2022.47649
23. Ho SJ, Jiang JW, Yu CT, Xu E, Yuen EY. Asian American female residents' perceptions of facilitators and barriers to leadership in medicine. *JAMA Netw Open*. 2025;8(5):e2512271. doi:10.1001/jamanetworkopen.2025.12271

24. Rajagopal SK, Durkee MI. Internalizing the model minority myth: dangers for Asian American mental health and attitudes towards other minorities. Soc Personal Psychol Compass. 2024;18(5):e12959. doi:10.1111/spc3.12959
25. Ginsburg S, Gingerich A, Kogan JR, Watling CJ, Eva KW. Idiosyncrasy in assessment comments: do faculty have distinct writing styles when completing in-training evaluation reports? *Acad Med*. 2020;95(Suppl 2):S81–S88. doi:10.1097/ACM.0000000000003643
26. Tulshian P, Montgomery L, McCrory K, et al. National recommendations for implementation of competency-based medical education in family medicine. *Fam Med*. 2025;57(4):253–260. doi:10.22454/FamMed.2025.866091