Differences in Narrative Language in Evaluations of Medical Students by Gender and Under-represented Minority Status

Alexandra E. Rojek, AB[1], Raman Khanna, MD, MAS[2], Joanne W. L. Yim, PhD[3], Rebekah Gardner, MD[4], Sarah Lisker, BA[1,5], Karen E. Hauer, MD, PhD[1], Catherine Lucey, MD[1], and Urmimala Sarkar, MD, MPH[1,5]

[1]University of California, San Francisco School of Medicine, San Francisco, CA, USA; [2]Division of Hospital Medicine, University of California, San Francisco, School of Medicine, San Francisco, CA, USA; [3]Health Informatics, UCSF Health, University of California, San Francisco, San Francisco, CA, USA; [4]Warren Alpert Medical School of Brown University, Providence, RI, USA; [5]UCSF Center for Vulnerable Populations, San Francisco, CA, USA.

BACKGROUND: In varied educational settings, narrative evaluations have revealed systematic and deleterious differences in language describing women and those underrepresented in their fields. In medicine, limited qualitative studies show differences in narrative language by gender and under-represented minority (URM) status.

OBJECTIVE: To identify and enumerate text descriptors in a database of medical student evaluations using natural language processing, and identify differences by gender and URM status in descriptions.

DESIGN: An observational study of core clerkship evaluations of third-year medical students, including data on student gender, URM status, clerkship grade, and specialty.

PARTICIPANTS: A total of 87,922 clerkship evaluations from core clinical rotations at two medical schools in different geographic areas.

MAIN MEASURES: We employed natural language processing to identify differences in the text of evaluations for women compared to men and for URM compared to non-URM students.

KEY RESULTS: We found that of the ten most common words, such as "energetic" and "dependable," none differed by gender or URM status. Of the 37 words that differed by gender, 62% represented personal attributes, such as "lovely" appearing more frequently in evaluations of women ($p < 0.001$), while 19% represented competency-related behaviors, such as "scientific" appearing more frequently in evaluations of men ($p < 0.001$). Of the 53 words that differed by URM status, 30% represented personal attributes, such as "pleasant" appearing more frequently in evaluations of URM students ($p < 0.001$), and 28% represented competency-related behaviors, such as "knowledgeable" appearing more frequently in evaluations of non-URM students ($p < 0.001$).

CONCLUSIONS: Many words and phrases reflected students' personal attributes rather than competency-related behaviors, suggesting a gap in implementing competency-based evaluation of students.

We observed a significant difference in narrative evaluations associated with gender and URM status, even among students receiving the same grade. This finding raises concern for implicit bias in narrative evaluation, consistent with prior studies, and suggests opportunities for improvement.

KEY WORDS: medical education; medical education—assessment/evaluation; medical student and residency education.

J Gen Intern Med 34(5):684–91
DOI: 10.1007/s11606-019-04889-9

INTRODUCTION

Core clerkships are a key foundation of medical education for students, and the assessments that are associated with these clerkships are informed by narrative evaluations completed by supervising physicians during these clerkships. These evaluations form the basis of clerkship grades, with the narrative language from evaluations being quoted in Medical Student Performance Evaluation (MSPE) letters and recommendation letters, that are a core component of residency applications.[1] Inherently, however, narrative language is open to bias and the consequences that can arise from it.[2]

The National Academies of Science, Engineering, and Medicine found that in academic settings, subjective evaluation criteria are often infiltrated with bias that disadvantages women.[3] The Association of American Medical Colleges reported that recruitment, evaluation, and promotion processes involve implicit and unconscious bias, inhibiting the development of a diverse medical workforce.[4] Research using manual and programmatic approaches to linguistic analyses, such as qualitative coding and automated text analysis, respectively, suggests that narrative evaluations can introduce gender-based stereotypes, including the perception of women as emotional and sensitive[5–11] that can be detrimental to the advancement of the individual being evaluated.[12] Furthermore, the consequences of subjective assessments may be even more damaging to racial and ethnic minorities that are underrepresented in these fields.[13] For example, underrepresented groups in medicine may be even more "othered," or differentiated in a manner that excludes, marginalizes, or subordinates.[14–18] This

An earlier version of this work was presented in Denver, Colorado, at the Society of General Internal Medicine's annual meeting in April 2018.

Electronic supplementary material *The online version of this article (https://doi.org/10.1007/s11606-019-04889-9) contains supplementary material, which is available to authorized users.*

Published online April 16, 2019

phenomenon may be due to an insufficiently diverse physician workforce,[19] as well as subject to the reported tendency of supervisors to focus on social and cultural factors rather than competency-related factors.[13]

In 1999, the Accreditation Council for Graduate Medical Education (ACGME) and the American Board of Medical Specialties endorsed competency-based evaluations in an effort to move towards assessment of specific competence domains based on behavior rather than personal attributes. The ACGME later introduced milestones as a standardized method to operationalize the assessment of students' progress towards achieving these competencies.[20] As a result, American medical schools focus on competency-based assessment.

We aim to characterize narrative language differences among medical student evaluations using natural language processing techniques. Our primary measure is to understand whether students are described differently by gender as well as under-represented minority (URM) status using metrics commonly employed in natural language processing.

METHODS

Design

This study was approved by the University of California, San Francisco Institutional Review Board (15-18271) and deemed exempt by the Brown University Institutional Review Board. This is a secondary data analysis of narrative evaluations (text) from two medical schools. We applied natural language processing to elucidate differences by gender and URM status.

Data Sources

We included data from all third-year core clerkship evaluations from two medical schools affiliated with public and private academic institutions in large urban settings, with associated information about student demographics, clerkship specialty, and grade. Data were collected from 2006 to 2015 at school 1 (as identified in Table 1), and from 2011 to 2016 in school 2, to exclude years in which major grading practice changes were implemented. At both of these schools, grading choices in each clerkship were three mutually exclusive choices: non-pass, pass, or honors, with no intermediate options. Only complete cases containing student gender, URM status, clerkship specialty, and grade received were used in analyses, with a total of 87,922 evaluations meeting these criteria. Students self-identified their ethnicity, and the medical schools determined which racial/ethnic categories were URM. We used this institutional definition of URM status as Black or African American, Hispanic or Latino, and American Indian or Alaska Native. All other self-identified ethnicities were categorized as non-URM.

Both schools included in this study fully incorporate ACGME recommendations for Core Competencies for medical student training,[20] and these recommendations had been implemented before the study period. Additionally, grades for required core clerkships were determined by a combination of clinical ratings with standardized exam scores, where the National Board of Medical Examiners (NBME) exam accounted for no more than 25% of a grade. At each school, no more than 30% of students received honors in a clerkship. Sample evaluation forms from each institution are available in Appendix Figures 1–2 online. Faculty at each institution were similar in composition: in 2015, school 1 had 48% female faculty, while school 2 had 45%. At school 1, 8% of faculty were URM, 88% were non-URM, and 4% were unknown; at school 2, 4% of faculty were URM, 84% were non-URM, and 12% were unknown.

Data were collected with unique identifiers for each narrative evaluation, without linkages between multiple evaluations for a single student across clerkships and time. A breakdown of evaluation composition by grade, gender, URM status, and specialty is shown in Table 1.

Table 1 Dataset Characteristics

Characteristic	Evaluations, $N = 87{,}922$, (%)	Evaluations, school 1 (%)	Evaluations, school 2 (%)
Student gender			
Male	38,952 (44)	30,431 (43)	8521 (46)
Female	48,970 (55)	39,074 (56)	9896 (53)
Student minority status			
Non-URM	65,974 (75)	51,933 (74)	14,041 (76)
URM	21,948 (25)	17,572 (25)	4376 (23)
Clerkship grade			
Honors	28,883 (32)	21,905 (31)	6978 (37)
Pass	58,748 (66)	47,332 (68)	11,416 (62)
Non-pass	291 (0.3)	268 (0.4)	23 (0.1)
Clerkship specialty			
Internal medicine	18,731 (21)	13,271 (19)	5460 (29)
Family medicine	8560 (9)	7139 (10)	1421 (7)
Surgery	11,049 (12)	8338 (12)	2711 (14)
Pediatrics	17,929 (20)	13,686 (19)	4243 (23)
Neurology	6366 (7)	5877 (8)	489 (2)
Psychiatry	9041 (10)	7712 (11)	1329 (7)
Ob/Gyn	9995 (11)	7231 (10)	2764 (15)
Anesthesia	6251 (7)	6251 (9)	0 (0)

URM, under-represented minority; Ob/Gyn, obstetrics/gynecology

Table 2 Grade Distribution by Gender, URM Status and Specialty

Clerkship	Evaluations of women with honors grades (%)	Evaluations of men with honors grades (%)	*p* value
Internal medicine	3503 (33)	2790 (33)	0.75
Family medicine	1581 (33)	1024 (26)	< 0.001
Surgery	1829 (30)	1627 (32)	0.01
Pediatrics	3505 (35)	2182 (27)	< 0.001
Neurology	1227 (34)	872 (30)	< 0.001
Psychiatry	1714 (34)	1121 (27)	< 0.001
Ob/Gyn	2353 (42)	1457 (32)	< 0.001
Anesthesia	1164 (31)	934 (36)	< 0.001
	Evaluations of URM students with honors grades (%)	Evaluations of non-URM students with honors grades (%)	*p* value
Internal medicine	792 (17)	5501 (38)	< 0.001
Family medicine	471 (22)	2134 (33)	< 0.001
Surgery	414 (15)	3042 (36)	< 0.001
Pediatrics	788 (17)	4899 (36)	< 0.001
Neurology	243 (15)	1856 (38)	< 0.001
Psychiatry	348 (15)	2487 (36)	< 0.001
Ob/Gyn	657 (24)	3153 (43)	< 0.001
Anesthesia	401 (26)	1697 (35)	< 0.001

De-identification

The narrative text of evaluations was de-identified in a two-step process. First, a database of names from publicly available US Census and Social Security name records was compiled,[21–23] and the text of evaluations was matched against these names, with a second filter of parts-of-speech processing to identify proper nouns. All names identified in this process were replaced with generic fillers. A subset of the narrative evaluations was manually verified for complete de-identification.

Parsing

We used an open software trained English language parser available from Google for parsing, which uses SyntaxNet,[24, 25] an open-source neural network framework for TensorFlow machine learning. This was applied to the narrative evaluations both to assist in de-identification as well as attribution of parts-of-speech tagging and text parsing, which formed the basis of the dataset used below in the primary analyses.

Analysis

First, we compared the distribution of grades, dichotomized to honors versus pass, across gender and URM status, as well as clerkship specialty, using Pearson's chi-squared tests after applying the Benjamini-Hochberg procedure for multiple testing correction. Second, we examined the length of evaluations, quantifying differences in distribution with the Wilcoxon-Mann-Whitney test.

Next, we generated a list of frequently used descriptors, defining descriptors as adjectives. The ten most frequent terms did not differ by gender or URM status when stratified by grade (Appendix Tables 2a and 2b online). To accurately characterize word frequency, we employed a widely used natural language processing method known as term frequency-inverse document frequency (TF-IDF),[26] which is a measure of the frequency of a term, adjusted for how rarely it is used. Here, we defined term-frequency as the frequency of a given word in an evaluation, and inverse document frequency as the inverse of the total frequency of the word's usage across all evaluations. We then averaged the TF-IDF value for a given word by gender and URM status. Examining TF-IDF by gender and URM status allowed us to infer the significance of a word, and whether this word was used with similar weight and meaning across evaluations. For example, the word "excellent" has a highly positive connotation. However, because it appeared so frequently across all evaluations, it corresponds to a lower TF-IDF score, thus one particular usage of the word "excellent" does not confer much meaning. In contrast, the word "energetic" appeared in fewer evaluations overall, so has a higher TF-IDF score, making each usage of "energetic" carry more weight.

TF-IDF has been shown in other work,[27] primarily in the field of information retrieval, to be a superior method compared to absolute term frequency as it has the ability to weight the frequency of terms in a manner that relates their "importance." As we suspected, the TF-IDF values of the most commonly used words determined by overall frequency were low, suggesting that their wide usage reflects a range of meanings (Appendix Table 3 online). We ranked the descriptors that were used in more than 1% of evaluations (a common threshold in large text datasets) by TF-IDF score, by gender, and by URM status.

Finally, we reported which descriptors evaluators used differently between groups by gender and URM status, using Pearson's chi-squared tests with Benjamini-Hochberg corrections. We surveyed this study's co-authors, who represent experts in medical education as well as clinical faculty, about the descriptors found to be used differently with statistical significance by gender and URM status, asking whether the descriptors were reflective of "personal attributes" versus "competency-related" terms, or neither of the above. We then categorized each word based on majority vote and present this categorization in Table 4.

Table 3 Important and Unique Descriptors, Among Commonly Used Words

Men (TF-IDF)	Women (TF-IDF)	Non-URM (TF-IDF)	URM (TF-IDF)
Energetic (0.72)	Friendly (0.64)	Energetic (0.64)	Friendly (0.76)
Friendly (0.68)	Energetic (0.62)	Friendly (0.61)	Energetic (0.71)
Fine (0.55)	Dependable (0.58)	Fine (0.56)	Dependable (0.56)
Competent (0.53)	Fine (0.56)	Knowledgeable (0.53)	Fine (0.53)
Smart (0.53)	Knowledgeable (0.53)	Dependable (0.52)	Competent (0.53)
Knowledgeable (0.52)	Personable (0.51)	Competent (0.50)	Personable (0.52)
Technical (0.48)	Technical (0.49)	Smart (0.49)	Technical (0.51)
Dependable (0.46)	Competent (0.48)	Technical (0.48)	Knowledgeable (0.50)
Personable (0.45)	Attentive (0.48)	Personable (0.47)	Smart (0.49)
Attentive (0.44)	Smart (0.46)	Attentive (0.46)	Attentive (0.47)

Among commonly used words (defined as appearing in > 1% of evaluations), importance was measured by term frequency-inverse document frequency, which is a metric of weighting term usage in an evaluation relative to usage in all evaluations; values closest to zero indicate that terms are used near equally across all evaluations and are deemed less unique

All analyses were performed with scripts written in R version 3.3.0 (2016-05-03). We considered two-sided $p < 0.05$ to be significant, after correcting with the Benjamini-Hochberg procedure, a multiple testing correction using false discovery rate estimation.[28, 29]

RESULTS

Grade Distribution

Overall, 32% of evaluations among all students were associated with honors grades, with 66% of evaluations associated with passing grades, and the remainder receiving non-pass grades (Table 1). Women received more honors than men and were more likely to receive honors in pediatrics, obstetrics/gynecology, neurology, and psychiatry; men were more likely to receive honors in surgery and anesthesia (Table 2). A comparison of non-URM and URM students showed that evaluations of URM students were associated with fewer honors grades than evaluations of non-URM students. When stratifying by clerkship specialty, URM students received fewer honors grades across all specialties. These distributions were comparable at each school included in our dataset (data not shown).

Table 4 Categorization of descriptors by personal attribute vs. competency

a. Personal attribute descriptors		
Active	Enthusiastic	Poised
Affable	Fabulous	Polite
Assertive	Humble	Relaxed
Bright	Intelligent	Reliable
Caring	Interesting	Respectful
Cheerful	Lovely	Sharp
Clear	Mature	Social
Considerate	Modest	Sophisticated
Delightful	Motivated	Talented
Earnest	Nice	Thoughtful
Easy-going	Open	Warm
Energetic	Pleasant	Wonderful
b. Competency-related descriptors		
Advanced	Impressive	
Basic	Integral	
Clinical	Knowledgeable	
Compassionate	Medical	
Complex	Relevant	
Comprehensive	Scientific	
Conscientious	Smart	
Efficient	Superior	
Empathic	Thorough	
Excellent		

The descriptors found in Figures 1 and 2 were categorized by a survey of the co-authors into "personal attribute descriptors" versus "competency-related descriptors" versus neither category. Those that a majority felt belonged in either of the first two groups are presented above, with all remaining descriptors deemed to belong in neither group

Evaluation Length

We looked at evaluation length by gender and URM status, stratified by grade (Appendix Table 1 online). We found that the distributions of evaluation length between different groups were similar and, although statistically significant in some instances, did not represent meaningful differences.

Common Descriptors by Gender and URM Status

Among descriptors that are used in more than 1% of evaluations, we examined the highest ranking words as measured by TF-IDF by gender and URM status (Table 3). Here, we found that the top ten ranked words were comparable across gender and URM status, suggesting that this measure does not provide sufficient granularity of analysis to assess meaningful differences in narrative evaluations.

Differential Usage of Descriptors by Statistical Significance

We found that among all evaluations, there were 37 words that differed by usage between men and women. Sixty-two percent (23/37) of these descriptors represented personal attributes, and of these, 57% (13/23) were used more in evaluations of women. In these evaluations of women, we saw that personal attribute descriptors such as "pleasant" were associated with pass grades, while "energetic," "cheerful," and "lovely" were neutral in their grade association. Additionally, personal attribute descriptors such as "wonderful" and "fabulous" that were used more frequently in evaluations of women were also associated with honors grades. In evaluations of men, personal attribute descriptors such as "respectful" or "considerate" were neutral in their association with grade, while "good" was seen more with pass grades, and "humble" was seen more with honors grades.

Of the 37 descriptors we found that differed by gender, only 19% (7/37) of these were words that we assigned as competency-related descriptors, and of these, 57% (4/7) were used more in evaluations of women. The descriptors "efficient," "comprehensive," and "compassionate" were used more often in evaluations of women and were also associated with honors grades; evaluations of men described as "relevant" were also associated with honors grades.

These descriptors that were associated with significantly different usage between men and women are shown in Figure 1 by their distribution along the *x*-axis, and the association any given word has with honors or pass grades is indicated by its distribution along the *y*-axis. In addition, words represented in Figure 1 (and Figure 2, described below) were found to be of high importance as measured by TF-IDF, with even higher values than the common words reported in Table 3 (data not shown).

Among all evaluations, there are 53 descriptors that differed by their usage between evaluations of URM and non-URM students. Thirty percent (16/53) of descriptors represented personal attributes, and of these, 81% (13/16) were used more often to describe non-URM students. The descriptors "pleasant," "open," and "nice" were used to describe URM students and were associated with passing grades. Many personal attribute descriptors used to describe non-URM students, such as "enthusiastic," "sharp," or "bright," were neutral in their association with grade, while "mature" and "sophisticated" were more frequently associated with honors grades.

Of the 53 descriptors that differed by URM status, only 28% (15/53) of these were competency-related descriptors, and 100% of these (15/15) were used more in evaluations of non-URM students. The competency-related descriptors "outstanding," "impressive," and "advanced" were more frequently associated with honors, while "superior," "conscientious," and "integral" were neutral in their association with grade. Of note, all of the descriptors (either personal attribute or competency-related) that were used more frequently in evaluations of non-URM students had either neutral associations with grade or were associated with honors grades. These descriptors that were associated with significantly different usage between URM and non-URM students are shown in Figure 2 by their distribution along the *x*-axis, along with the association any given word had with honors or pass grades as indicated by its distribution along the *y*-axis.

DISCUSSION

This novel application of natural language processing to what we believe is the largest sample of medical student evaluations analyzed to-date reveals how students are described differently by gender and URM status. We found that across student evaluations, common, important words were used with similar frequency across gender and URM status. However, our analysis revealed significant differences in the usage of particular words between genders, as well as by URM status. While

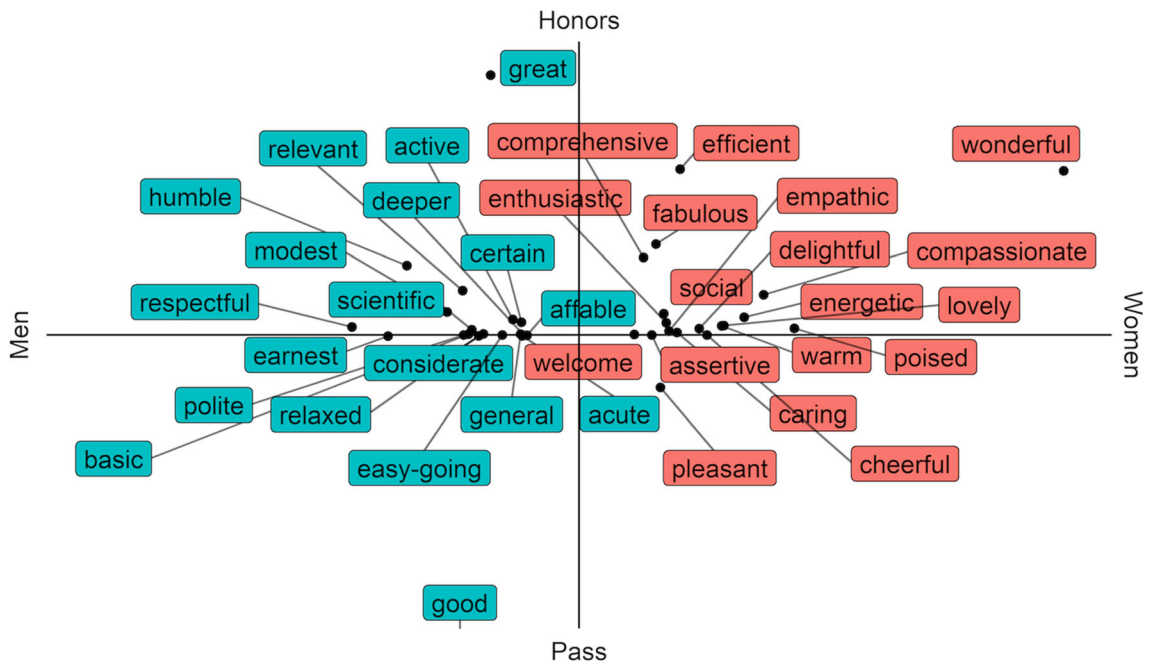

Fig. 1 Descriptors with statistically significant differences in usage by gender All words were assessed for differential usage between groups of interest, with statistical significance defined as $p < 0.05$. Location of a word point on the men-women axis indicates its preferential use in either gender. Distance from the *y*-axis also indicates increased difference from expected word distribution, noting however that all words shown are statistically significant in their usage by gender. Placement along the pass-honors axis indicates association of a given word with usage in either more honors- or pass-graded evaluations. Red-highlighted words identify words that are used more in evaluations of women, while blue-highlighted words identify words that are used more in evaluations of men. The categorization of these terms by "personal attribute" versus "competency-related" descriptors can be found in Table 4.

terms deemed important by the TF-IDF metric were reflective of personal attributes and competence, and were comparable among genders and URM status, the words with statistically significant differences in usage between these groups indicate inclusion of personal attributes more so than competencies, as defined in Table 4. Although there were both competency-related and personal attribute descriptors that are used differentially between gender and URM statuses, there is a dominance of personal attribute descriptors in the words we found to be used differently between these groups that we believe is important in signaling how student performance is assessed.

Our study is consistent with previous work examining differences in grading between genders and URM status groups in limited settings. Lee et al. showed that URM students receive lower grades across clerkships,[30] and other work has shown conflicting effects of student demographics on clerkship grades, including student age.[31] Whether other objective measures of academic performance, such as prior standardized test performance or undergraduate GPAs, contribute to clerkship evaluation grades has also been debated.[30, 32] However, previous work has been limited in scope, both with respect to clerkship specialties and small sample sizes. The breadth of our data allows for identification of infrequent instances of differential descriptors that are concerning when considered in the context of the entire population of medical students.

In prior studies of narrative evaluations, investigators examined differential usage of a pre-determined set of words. Women have been shown to be more likely than men to be associated with words like "compassion," "enthusiasm," and "sensitivity," while other studies have shown that the presence of "standout" words, such as "exceptional" or "outstanding," predicted evaluations of men but not of women.[8] Additional research suggests that similar patterns extend beyond the realm of student evaluations.[5, 33] A strength of natural language processing is that we did not have to pre-specify words that might differ; instead, we were able to extract any differing words without the introduction of additional analytic bias.

Despite the intent of clerkship assessment to address competencies by observing behaviors, the differences we found

Fig. 2 Descriptors with statistically significant differences in usage by URM status All words were assessed for differential usage between groups of interest, with statistical significance defined as $p < 0.05$. Location of a word point on the Non-URM-URM axis indicates its preferential use by URM status. Distance from the *y*-axis also indicates increased difference from expected word distribution, noting however that all words shown are statistically significant in their usage by URM status. Placement along the pass-honors axis indicates association of a given word with usage in either more honors- or pass-graded evaluations. Green-highlighted words identify words that are used more in evaluations of URM students, while purple-highlighted words identify words that are used more in evaluations of non-URM students. The categorization of these terms by "personal attribute" versus "competency-related" descriptors can be found in Table 4.

among URM and non-URM assessments were more reflective of perceived personal attributes and traits. In prior work, Ross et al. found that Black residency applicants were more likely to be described as "competent", whereas White applicants more frequently received standout and ability descriptors, like "exceptional".[34] Examining our findings of the variation in descriptors used for URM and non-URM students in the context of the literature is pertinent considering the discrimination and disparities faced by racial and ethnic minorities as trainees and healthcare providers. Disparities in clerkship grades,[30] membership in honors societies,[35] and promotion[36, 37] are well-documented. Research on performance assessments suggests that small differences in assessment can result in larger differences in grades and awards received—a phenomenon referred to as the "amplification cascade." Teherani et al. illustrate the presence of this phenomenon among URM and non-URM students in undergraduate medical education.[38] The amplification cascade holds major implications for residency selection and ultimately career progression that can disproportionately affect students from underrepresented groups.

Our study has limitations. First, although our sample size is large, we analyzed evaluations from two medical schools. Second, baseline measures of academic performance or subsequent markers of career success were unavailable, which limited our ability to extrapolate the effect of these differences in text beyond the grade received in the clerkship. Third, due to data limitations, we were unable to link evaluations of individual students across clerkships to assess patterns in grading behaviors and biases, although all students rotating in core clerkships in the study years were included in the dataset. Fourth, we were unable to assess any interaction between evaluator demographics and narrative language differences, as seen in other studies.[8] Fifth, we did not have access to Dean's Letters, also known as the MSPE, which are a compilation of comments from individual clerkships and are the direct link between student evaluations and residency applications. Finally, we did not examine if narrative differences were more pronounced among members of both groups, such as women who also identify as URM.

Despite efforts to standardize medical student evaluations, the differences in narrative language suggest directions for improvement of medical education assessment. At a minimum, our findings raise questions about the inclusion of verbatim commentary from these assessments in MSPE letters used in residency applications, as is the accepted national standard.[1] Similarly, our work demonstrates that the competency-based evaluation framework[37] ostensibly in use for evaluating medical students remains incompletely implemented. Finally, behavioral science research has uncovered best practices for reducing bias in evaluations, including comparative evaluations,[39] structured criteria,[40, 41] task demonstrations, blinded evaluations,[42] and evaluator training featuring de-biasing techniques.[43] In the future, it may be possible for language processing tools to provide real-time and data-driven feedback to evaluators to address unconscious bias. Perhaps it is time to rethink narrative clerkship evaluations to better serve all students.

Acknowledgements: *The authors would like to thank Roy Cherian, Cassidy Clarity, Gato Gourley, Bonnie Hellevig, Mark Lovett, Kate Radcliffe, and Alvin Rajkomar.*

Corresponding Author: *Urmimala Sarkar, MD, MPH; Health Informatics, UCSF Health, University of California, San Francisco, San Francisco, CA, USA (e-mail: urmimala.sarkar@ucsf.edu).*

Funding Information *Dr. Sarkar is supported by the National Cancer Institute (K24CA212294).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. Association of American Medical Colleges. Recommendations for Revising the Medical Student Performance Evaluation (MSPE). May 2017. https://www.aamc.org/download/470400/data/mspe-recommendations.pdf. Accessed December 11, 2018.
2. **Biernat M**, **Tocci MJ**, **Williams JC**. The Language of Performance Evaluations: Gender-Based Shifts in Content and Consistency of Judgment. *Social Psychological and Personality Science*. 2012;3(2):186–192. https://doi.org/10.1177/1948550611415693
3. **Manning CD**, **Raghavan P**, Schütze H. *Introduction to Information Retrieval*. Cambridge Unviersity Press; 2009.
4. **Corrice A.** Unconscious Bias in Faculty and Leadership Recruitment: A Literature Review. *Association of American Medical Colleges*. 2009;9(2).
5. **Trix F**, **Psenka C**. Exploring the Color of Glass: Letters of Recommendation for Female and Male Medical Faculty. *Discourse & Society*. 2003;14(2):191–220. https://doi.org/10.1177/0957926503014002277
6. **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. https://doi.org/10.1177/0163278710375097
7. **Galvin SL**, **Parlier AB**, **Martino E**, **Scott KR**, **Buys E**. Gender Bias in Nurse Evaluations of Residents in Obstetrics and Gynecology. *Obstet Gynecol*. 2015;126 Suppl 4:7S–12S. https://doi.org/10.1097/AOG.0000000000001044
8. **Isaac C**, **Chertoff J**, **Lee B**, **Carnes M**. Do students' and authors' genders affect evaluations? A linguistic analysis of Medical Student Performance Evaluations. *Acad Med*. 2011;86(1):59–66. https://doi.org/10.1097/ACM.0b013e318200561d
9. **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. https://doi.org/10.1007/s11199-007-9291-4
10. **Magua W**, **Zhu X**, **Bhattacharya A**, et al. Are Female Applicants Disadvantaged in National Institutes of Health Peer Review? Combining Algorithmic Text Mining and Qualitative Methods to Detect Evaluative Differences in R01 Reviewers' Critiques. *Journal of Women's Health*. 26(5):560-570. https://doi.org/10.1089/jwh.2016.6021
11. **Kaatz A**, **Magua W**, **Zimmerman DR**, **Carnes M**. A quantitative linguistic analysis of National Institutes of Health R01 application critiques from investigators at one institution. *Acad Med*. 2015;90(1):69–75. https://doi.org/10.1097/ACM.0000000000000442
12. **Moss-Racusin CA**, **Dovidio JF**, **Brescoll VL**, **Graham MJ**, **Handelsman J**. Science faculty's subtle gender biases favor male students. *PNAS*. 2012;109(41):16474–16479. https://doi.org/10.1073/pnas.1211286109
13. **Wilson KY**. An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations* 63(12):1903-1933. https://doi.org/10.1177/0018726710369396

14. **P. Bourdieu**. *Distinction: A Social Critique of the Judgement of Taste*. Cambridge, MA: Harvard University Press; 1984.
15. **G. C. Spivak**. The Rani of Sirmur: An Essay in Reading the Archives. *History and Theory*. 1985;24(3):247–272.
16. **S. De Beauvoir**. *The Second Sex*. New York: Knopf; 1952.
17. **L. Weis**. Identity formation and the process of "othering": Unraveling sexual threads. *Educational Foundations*. 1995;9(1):17–33.
18. **W. I. U. Ahmad**. Making Black people sick: 'Race', ideology and health research. In: *'Race' and Health in Contemporary Britain*. Philadelphia: Open University Press; 1993:12–33.
19. Association of American Medical Colleges. Diversity in the Physician Workforce: Facts & Figures 2014. 2014. http://aamcdiversityfactsandfigures.org/. Accessed December 11, 2018.
20. **Holmboe E**, **Edgar L**, **Hamstra S**. *The Milestone Guidebook*. Accreditation Council for Graduate Medical Education; 2016.
21. US Census Bureau. Frequently Occurring Surnames from the 1990 Census. https://www.census.gov/topics/population/genealogy/data/1990_census.html. Published September 15, 2014. Accessed December 11, 2018.
22. US Census Bureau. (2000) Frequently Occurring Surnames from the Census. https://www.census.gov/topics/population/genealogy/data/2000_surnames.html. Published September 15, 2014. Accessed December 11, 2018.
23. Social Security Administration. Beyond the Top 1000 Names. Popular Baby Names. https://www.ssa.gov/OACT/babynames/limits.html. Published 2017. Accessed December 11, 2018.
24. **Presta**, **A.**, **Severyn**, **A.**, Golding, A., et al. *SyntaxNet: Neural Models of Syntax*. tensorflow; 2018. https://github.com/tensorflow/models. Accessed December 11, 2018.
25. **Andor D**, **Alberti C**, Weiss D, et al. *Globally Normalized Transition-Based Neural Networks.*; 2016. http://arxiv.org/abs/1603.06042. Accessed December 11, 2018.
26. **Vijila, S.** F., Nirmala, K. Dr. Quantification of Portrayal Concepts using tf-idf Weighting. *IJIST*. 2013;3(5). https://doi.org/10.5121/ijist.2013.3501
27. **Dertat**, **A.** (2011) How to Implement a Search Engine Part 3: Ranking tf-idf. http://www.ardendertat.com/2011/07/17/how-to-implement-a-search-engine-part-3-ranking-tf-idf/. Accessed December 11, 2018.
28. **Storey JD**, **Tibshirani R**. Statistical significance for genomewide studies. *PNAS*. 2003;100(16):9440–9445. https://doi.org/10.1073/pnas.1530509100
29. **Noble WS**. How does multiple testing correction work? *Nat Biotechnol*. 2009;27(12):1135–1137. https://doi.org/10.1038/nbt1209-1135
30. **Lee KB**, **Vaishnavi SN**, **Lau SKM**, **Andriole DA**, **Jeffe DB**. "Making the grade:" noncognitive predictors of medical students' clinical clerkship grades. *J Natl Med Assoc*. 2007;99(10):1138–1150.
31. **Colbert C**, **McNeal T**, **Lezama M**, et al. Factors associated with performance in an internal medicine clerkship. *Proc (Bayl Univ Med Cent)*. 2017;30(1):38–40.
32. **Ogunyemi D**, **De Taylor-Harris S**. NBME obstetrics and gynecology clerkship final examination scores: predictive value of standardized tests and demographic factors. *J Reprod Med*. 2004;49(14):978–982.
33. **Mueller AS**, **Jenkins TM**, **Osborne M**, **Dayal A**, **O'Connor DM**, **Arora VM**. Gender Differences in Attending Physicians' Feedback to Residents: A Qualitative Analysis. *Journal of Graduate Medical Education*. 2017;9(5):577–585. https://doi.org/10.4300/JGME-D-17-00126.1
34. **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. https://doi.org/10.1371/journal.pone.0181659
35. **Boatright D**, **Ross D**, **O'Connor P**, **Moore E**, **Nunez-Smith M**. Racial Disparities in Medical Student Membership in the Alpha Omega Alpha Honor Society. *JAMA Intern Med*. 2017;177(5):659–665. https://doi.org/10.1001/jamainternmed.2016.9623
36. **Jena AB**, **Khullar D**, **Ho O**, **Olenski AR**, **Blumenthal DM**. Sex Differences in Academic Rank in US Medical Schools in 2014. *JAMA*. 2015;314(11):1149–1158. https://doi.org/10.1001/jama.2015.10680
37. **Nunez-Smith M**, **Ciarleglio MM**, **Sandoval-Schaefer T**, et al. Institutional Variation in the Promotion of Racial/Ethnic Minority Faculty at US Medical Schools. *Am J Public Health*. 2012;102(5):852–858. https://doi.org/10.2105/AJPH.2011.300552
38. **A Teherani**, **K E Hauer**, **A Fernandez**, **T E King Jr**, **C Lucey**. How Small Differences in Assessed Clinical Performance Amplify to Large Differences in Grades and Awards: A Cascade With Serious Consequences for Students Underrepresented in Medicine. *Acad Med*. 2018;93(9):1286–1292.
39. Bohnet, I., van Geen, A., Bazerman, M. When Performance Trumps Gender Bias: Joint vs. Separate Evaluation. *Management Science*. 2015;621225–1234. https://doi.org/10.1287/mnsc.2015.2186
40. **Reskin BF**, **McBrier DB**. Why Not Ascription? Organizations' Employment of Male and Female Managers. *American Sociological Review*. 2000;65(2):210–233. https://doi.org/10.2307/2657438
41. **Gawande A.** *The Checklist Manifesto: How to Get Things Right*. New York: Picador; 2011.
42. **Goldin C**, **Rouse C**. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review*. 2000;90(4):715–741. https://doi.org/10.1257/aer.90.4.715
43. **Babcock L**, **Loewenstein G**. Explaining bargaining impasse: the role of self-serving biases. *Journal of Economic Perspectives*. 1997;11(1):109–126. https://doi.org/10.1257/jep.11.1.109