Minimizing Bias in Narrative Assessment

Amber T. Pincavage, MD

Associate Professor
Clerkship Director, Internal Medicine
University of Chicago Pritzker School of Medicine

Disclosures

- I have no financial disclosures

Learning Objectives

At the conclusion of this activity, participants will be able to:

- Summarize the evidence for racial/ethnic and gender bias in narrative assessment
- Identify strategies to minimize bias in narrative assessment
- Apply these strategies to sample narratives

Educational Case

One of the other educators is working on their assessment methods. They ask for your feedback and also ask: “Is there any evidence of racial/ethnic or gender bias in educational assessment?”--How do you answer?

A. Only in AOA selection
B. Yes, in Clerkship grades only
C. Yes, in Clerkship evaluation narrative language only
D. Yes, in Clerkship grades, evaluations, AOA selection, MSPE letters, and Resident performance evaluations

What is Implicit Bias?

- Attitudes of stereotypes that affect our understanding, actions, and decisions in an unconscious manner

Merriam-Webster Dictionary

Bias in Assessment Evidence

TEACHING AND LEARNING IN MEDICINE
2019, VOL. 31, NO. 5, 487–496
https://doi.org/10.1080/10401334.2019.1597724

Routledge
Taylor & Francis Group

GROUNDWORK

Racial/Ethnic Disparities in Clinical Grading in Medical School

Daniel Low[a], Samantha W. Pollack[b], Zachary C. Liao[c], Ramoncita Maestas[d], Larry E. Kirven[e], Anne M. Eacker[f], and Leo S. Morales[g]

[a]Swedish Cherry Hill Family Medicine Residency, University of Washington School of Medicine, Seattle, Washington, WA, USA; [b]Department of Family Medicine, University of Washington School of Medicine, Seattle, Washington, USA; [c]Jackson Memorial Hospital, Internal Medicine Residency, University of Miami, Miami, FL, USA; [d]Student Affairs, University of Washington School of Medicine, Seattle, Washington, USA; [e]Wyoming WWAMI Program, University of Washington School of Medicine, Seattle, Washington, USA; [f]Kaiser Permanente School of Medicine, Pasadena, California, USA; [g]Center for Health Equity, Diversity, and Inclusion, University of Washington School of Medicine, Seattle, Washington, USA

ABSTRACT

Phenomenon: Performance during the clinical phase of medical school is associated with membership in the Alpha Omega Alpha Honor Medical Society, competitiveness for highly selective residency specialties, and career advancement. Although race/ethnicity has been found to be associated with clinical grades during medical school, it remains unclear whether other factors such as performance on standardized tests account for racial/ethnic differences in clinical grades. Identifying the root causes of grading disparities during the clinical phase of medical school is important because of its long-term impacts on the career advancement of students of color. *Approach:* To evaluate the association between race/ethnicity and clinical grading, we examined Medical Student Performance Evaluation (MSPE) summary words (Outstanding, Excellent, Very Good, Good) and 3rd-year clerkship grades among medical students at the University of Washington School of Medicine. The analysis

KEYWORDS

race; ethnicity; bias; grading; medical school; clerkships

Disparities in Grades & MSPE

Association between MSPE summary words and 3rd year clerkship grades and URM and non-URM status using ordinal logistic regression models

Low et al, TLM. 2019; 31(5):487-96.

Disparities in Grades & MSPE

- White or female students with higher final clerkship grades
- Grading disparities favored White students over either URM or non-URM minority students in 4 out of 6 clerkships
  - AOR ranged from 0.49 to 1.05
- URM status trended toward lower likelihood of higher category MSPE word
  - AOR 0.67, p= .11
- Non-URM minority students were significantly less likely to receive a higher category word than White students
  - AOR 0.53, p= .001
- Men less likely to receive a higher MSPE summary word than women
  - AOR=0.46 p>0.001

Low et al, TLM. 2019; 31(5):487-96.

AOA and GHHS Selection

Research Report

All Other Things Being Equal: Exploring Racial and Gender Disparities in Medical School Honor Society Induction

Thilan P. Wijesekera, MD, Margeum Kim, MS, Edward Z. Moore, PhD, Olav Sorenson, PhD, and David A. Ross, MD, PhD

Abstract

Purpose
A large body of literature has demonstrated racial and gender disparities in the physician workforce, but limited data are available regarding the potential origins of these disparities. To that end, the authors evaluated the effects of race and gender on Alpha Omega Alpha Honor Medical Society (AOA) and Gold Humanism Honor Society (GHHS) induction.

Method
In this retrospective cohort study, the authors examined data from 11,781 Electronic Residency Application Service applications from 133 U.S. MD-granting medical schools to 12 residency programs in the 2014–2015 application cycle and to all 15 residency programs in the 2015–2016 cycle at Yale-New Haven Hospital. They estimated the odds of induction into AOA and GHHS using logistic regression models, adjusting for Step 1 score, research publications, citizenship status, training interruptions, and year of application. They used gender- and race-matched samples to account for differences in clerkship grades and to test for bias.

Results
Women were more likely than men to be inducted into GHHS (odds ratio 1.84, $P < .001$) but did not differ in their likelihood of being inducted into AOA. Black medical students were less likely to be inducted into AOA (odds ratio 0.37, $P < .05$) but not into GHHS.

Conclusions
These findings demonstrate significant differences between groups in AOA and GHHS induction. Given the importance of honor society induction in residency applications and beyond, these differences must be explored further.

Wijesekera et al, Acad Med. 2019; 94(4):562-569.

PLOS ONE

RESEARCH ARTICLE

Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations

David A. Ross[1]*, Dowin Boatright[2], Marcella Nunez-Smith[3,4], Ayana Jordan[1], Adam Chekroud[5], Edward Z. Moore[6]

1 Department of Psychiatry, Yale University School of Medicine, New Haven, CT, United States of America, **2** Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT, United States of America, **3** Department of General Internal Medicine, Yale University School of Medicine, New Haven, CT, United States of America, **4** Department of Epidemiology, Yale School of Public Health, New Haven, CT, United States of America, **5** Department of Psychology, Yale University, New Haven, CT, United States of America, **6** Department of Engineering, Central Connecticut State University, New Britain, CT, United States of America

* david.a.ross@yale.edu

OPEN ACCESS

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

Editor: Jeffrey A. Gold, Oregon Health and Science University, UNITED STATES

Received: March 30, 2017

Accepted: July 5, 2017

Published: August 9, 2017

Abstract

Purpose

The transition from medical school to residency is a critical step in the careers of physicians. Because of the standardized application process–wherein schools submit summative Medical Student Performance Evaluations (MSPE's)–it also represents a unique opportunity to assess the possible prevalence of racial and gender disparities, as shown elsewhere in medicine.

Method

The authors conducted textual analysis of MSPE's from 6,000 US students applying to 16 residency programs at a single institution in 2014–15. They used custom software to extract demographic data and keyword frequency from each MSPE. The main outcome measure was the proportion of applicants described using 24 pre-determined words from four thematic categories ("standout traits", "ability", "grindstone habits", and "compassion").

GME Evidence

Gender Bias in EM

- Analysis of 2,765 performance evaluations in EM no gender bias in year 1, however, in year 3, men were perceived as outperforming women.
- In 3rd year but not the 1st, women received more harsh criticism and less supportive feedback than men for medical errors of similar severity
- Although male and female residents received similar evaluations at the beginning of residency, the rate of milestone attainment throughout training was higher for male than female residents across all EM sub-competencies in 8 EM programs

Brewer A, et al. American Sociological Review. 2020; 85(2):247-270.
Dayal A, et al. JAMA Intern Med. 2017 May 1;177(5):747.

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 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

Narrative Evaluation Differences

Question: Are there differences in medical student evaluation narrative language based on gender and URM status?

Design: Retrospective cohort study

Setting and population: University of California San Francisco SOM and Warren Alpert Medical School of Brown University third-year medical students

Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

Outcomes: natural language processing differences in text of de-identified core clinical rotation evaluations for:

1. women compared to men
2. URM compared to non-URM

Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

Results: Demographics

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

Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

Results: Grade Distribution

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

Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

Results: categorization of descriptors

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	

Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

Results: Top ten words and Eval length

Table 3 Important and Unique Descriptors, Among Commonly Used Words

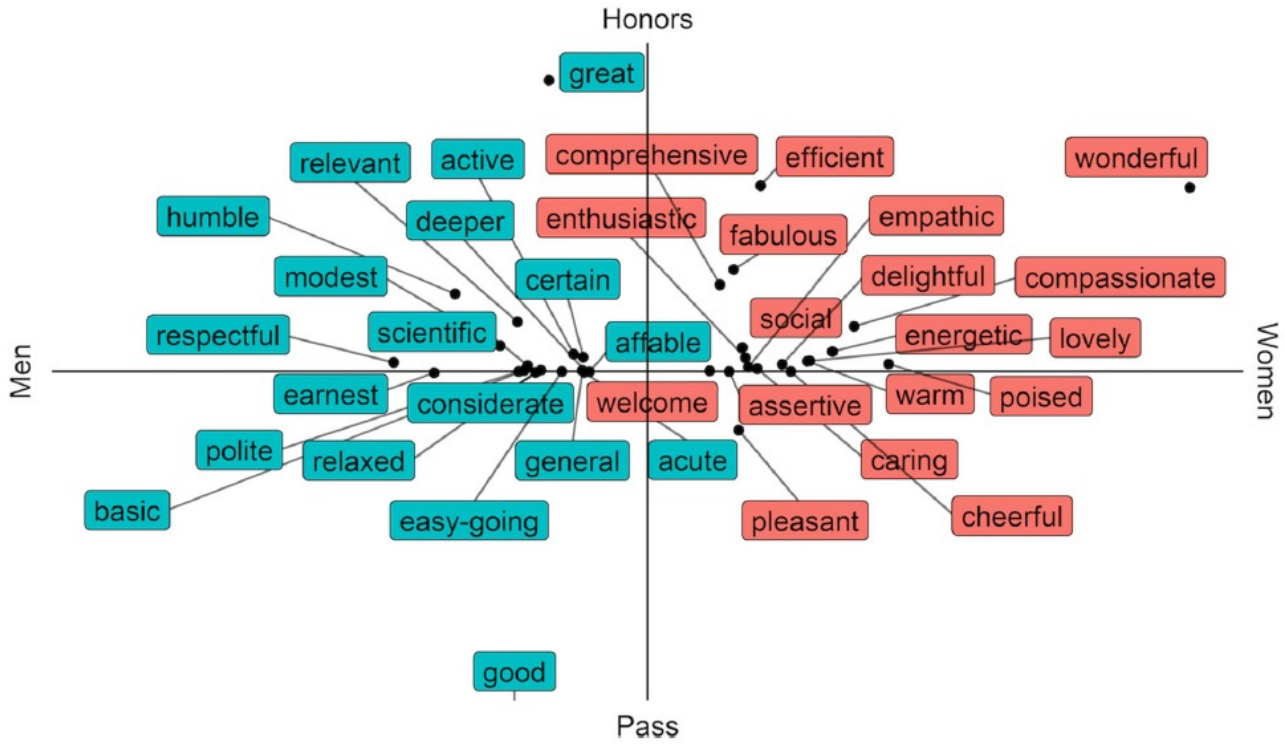
Men (TF-IDF)	Women (TF-IDF)	Non-URM (TF-IDF)	URM (TF-IDF)
Energetic (0.72)	Friendly (0.64)	Energetic (0.64)	Friendly (0.76)
Friendly (0.68)	Energetic (0.62)	Friendly (0.61)	Energetic (0.71)
Fine (0.55)	Dependable (0.58)	Fine (0.56)	Dependable (0.56)
Competent (0.53)	Fine (0.56)	Knowledgeable (0.53)	Fine (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

Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

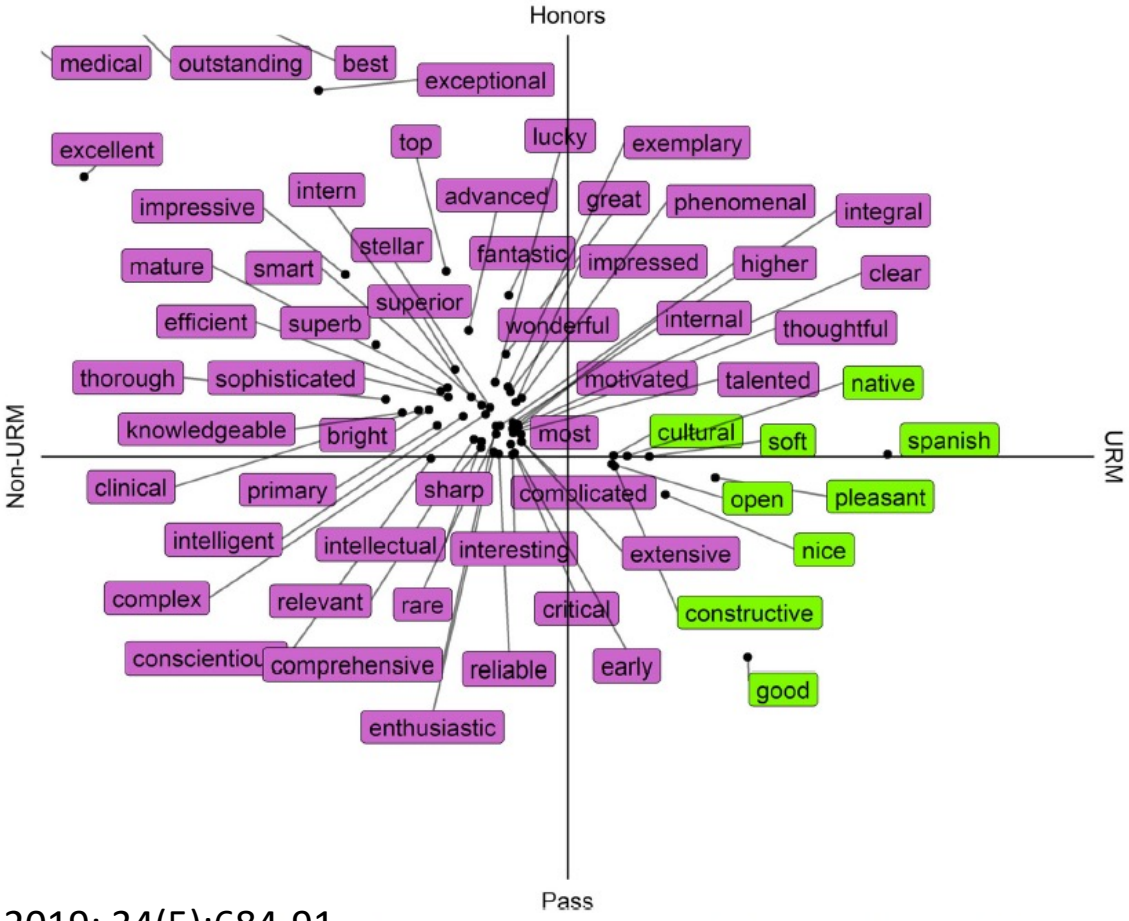
Results: Descriptors that differ by gender



Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

Results: Descriptors that differ by URM status



Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

Limitations:

- H/P/F grading system
- Two institutions represented
- Unable to link evaluations of individual students across clerkships
- Unable to assess interaction of evaluator demographics
- Unable to look at intersectionality
- Only looked at single words out of context
- Categorization of descriptors as personal attributes vs. competency based was subjective
- Did not look at MSPE summaries or residency LOR's

Rojek et al, JGIM. 2019; 34(5):684-91.

Narrative Evaluation Differences

Conclusions:

There were significant differences in the usage of particular words between genders and by URM status. These were often words that described personal attributes as opposed to competency-related behaviors.

Rojek et al, JGIM. 2019; 34(5):684-91.

Strategies to Minimize Narrative Bias

Strategies to Minimize Narrative Bias

- Intentional Narrative language
- Intentional competency inclusion
- Group decision making
- Blinded editing of narratives
- Implicit bias training
- Systematic approaches

Intentional Narrative language

Personal-attribute descriptor examples		
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

Competency-related descriptor examples	
Advanced	Impressive
Basic	Integral
Clinical	Knowledgeable
Compassionate	Medical
Complex	Relevant
Comprehensive	Scientific
Conscientious	Smart
Efficient	Superior
Empathic	Thorough
Excellent	

Rojek AE, 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): 684-91

Intentional Language

- Use of titles is also important
  - Study of video recordings of Grand Rounds at 2 institutions
    - Women nearly always used the title “doctor” to introduce speakers (96%)
    - Men who made introductions used it 66% of the time:
      - When men introduced men, they used formal titles 73% of the time
      - When men introduced women this dropped to 49%

Files JA, Mayer AP, Ko MG, et al. *Speaker introductions at internal medicine grand rounds: forms of address reveal gender bias. J Womens Health (Larchmt)2017;26:4139.doi:10.1089/jwh.2016.6044*

Competency related language

UME

- Medical knowledge
- Clinical Skills
- Communica
- Clinical
- Profes
- System
- Practice-ba
  and improvement

GME

- Medical Knowledge
- Patient Care
- rsonal and
- ation skills
- ism
- sed learning
- ovement
- ystems-based practice

Selection of competencies included in narrative may also introduce bias

- Can use EPA's as well

Competency descriptors

UME

- Reporter=Good/Very good
- Interpreter=Excellent
- Manager=Outstanding
- Educator=Exceptional

- Ready for GME training

GME

- Critical Deficiency
- Basic Competence (average intern)
- Advanced Competence (avg PGY-2)
- Ready for Unsupervised practice
- Aspirational

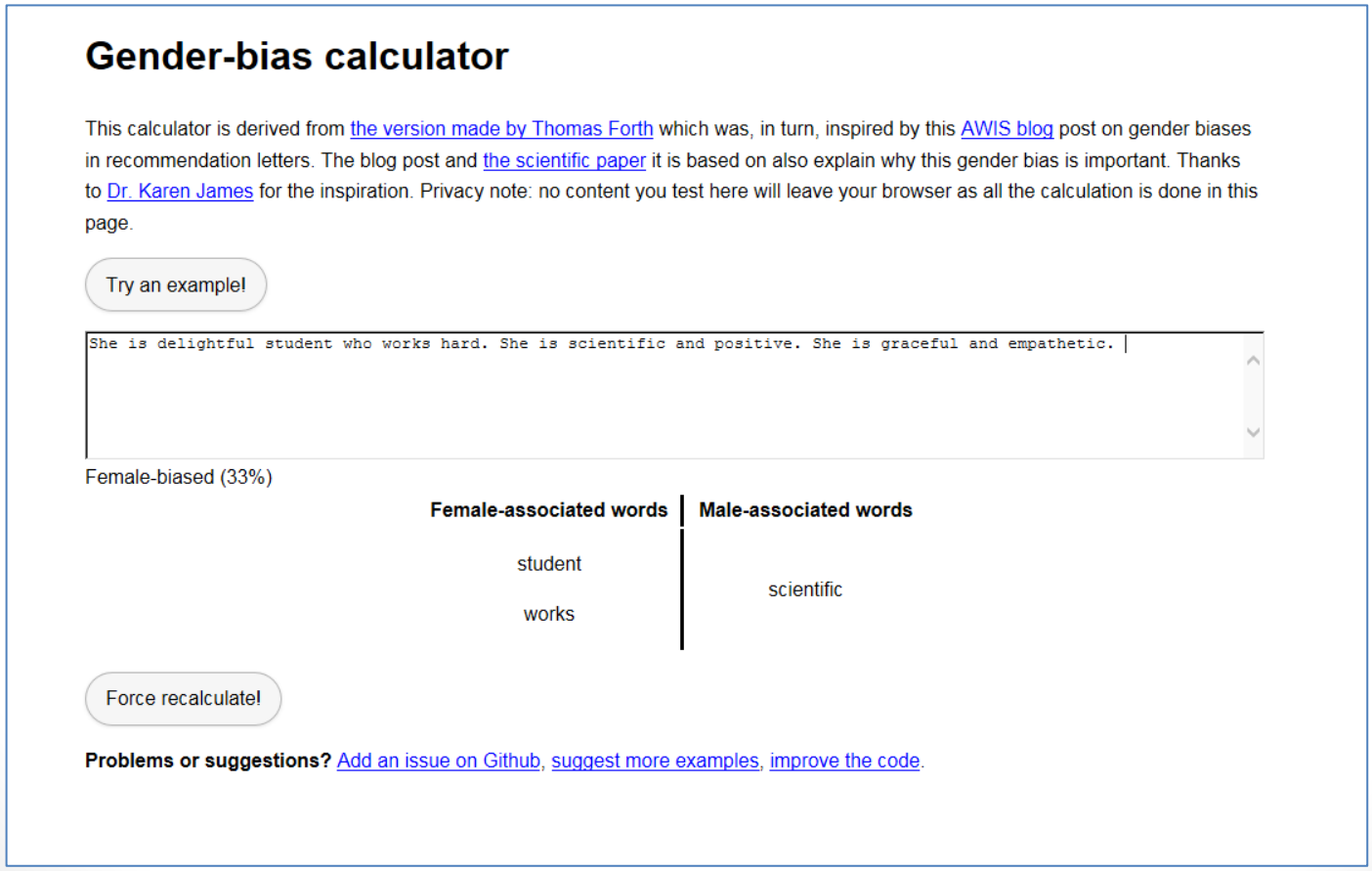
How do you still create unique narratives?

- Tell a story describing an encounter that illustrates the learner's behavior to show their personality

Student D went to talk to their patient after their procedure and review results although the team already dismissed them to study

Intentional Narrative Evaluations

- Consider using free online tool to look for gender bias in eval:
- **https://www.tomforth.co.uk/genderbias/**



Gender-bias calculator

This calculator is derived from the version made by Thomas Forth which was, in turn, inspired by this AWIS blog post on gender biases in recommendation letters. The blog post and the scientific paper it is based on also explain why this gender bias is important. Thanks to Dr. Karen James for the inspiration. Privacy note: no content you test here will leave your browser as all the calculation is done in this page.

Try an example!

She is delightful student who works hard. She is scientific and positive. She is graceful and empathetic.

Female-biased (33%)

Female-associated words	Male-associated words
student	scientific
works	

Force recalculate!

Problems or suggestions? Add an issue on Github, suggest more examples, improve the code.

Group Decision Making

- Synthesize multiple data points in a standardized and consistent manner
- Social decision scheme theory
  - Sharing and processing information-> better decisions
- Examples: CCC or Grading committee

Blinded Evaluations

- Can someone edit or review evaluations in a blinded fashion to create summative narratives?

Systematic approaches

- Consider changes to your evaluation form or evaluation system
  - Prompts about intentional narrative language
  - Prompts to consider implicit biases that may be present
  - Requiring all competencies be evaluated

Systematic Approaches

- "Student evaluations of teaching play an important role in the review of faculty. Your opinions influence the review of instructors that takes place every year. Iowa State University recognizes that student evaluations of teaching are often influenced by students' unconscious and unintentional biases about the race and gender of the instructor. Women and instructors of color are systematically rated lower in their teaching evaluations than white men, even when there are no actual differences in the instruction or in what students have learned. As you fill out the course evaluation please keep this in mind and make an effort to resist stereotypes about professors…"

Peterson DAM, Biederman LA, Andersen D, Ditonto TM, Roe K. *Mitigating gender bias in student evaluations of teaching. PLoS One 2019; 14:e0216241.doi:10.1371/journal.pone.0216241*

Application Exercise

How would you edit?

- A had solid medical knowledge and gave nice reads with a “great eye”. A exceeded all of the basic expectations. A always has a cheerful uplifting attitude, but was still never satisfied unless A could find some way to improve. A’s immediate responsiveness to feedback demonstrates both A’s unique abilities as a professional, as a radiologist, and also A’s humble grace as a learner. As is common in A’s culture, A is quiet and studious. A was reliable, respectful, pleasant, and mature.

How would you edit?

- B was a delight to work with, and B is also the most professional team member I have ever worked with. B always demonstrated compassion, honesty, poise and a high level of integrity. B was assertive and always the first to volunteer for additional reading responsibilities, and could always be trusted to carry out those responsibilities. B readily sought out constructive criticism and applied it to improve reading skills. B gave excellent presentations for the department that were a joy to attend. B is always a pleasure to work with and will make a fabulous resident.

How would you edit?

- C had outstanding medical knowledge and sharp reading skills. C had a detailed approach to reading images and is an exceptionally bright radiologist with a wealth of medical knowledge. C's teaching presentations to the department are always concise, organized, and sophisticated. C is efficient, accurate and hard-working. C demonstrated leadership by being adept at managing the chaos of call in a busy level one trauma center.

In Summary

- There is bias in narrative assessment based on URM status and gender
- Strategies to mitigate narrative bias include:
  - Intentional narrative:
    - Using competency-based as opposed to personal attribute language
  - Intentional competency inclusion:
    - Commenting on all UME and GME competencies in narratives
  - Group Decision-Making
  - Blinded Narratives

Tell stories to illustrate personality and behavior to make narratives unique!

Questions

- Contact info: apincava@bsd.uchicago.edu
- Upcoming CDIM workshop on Minimizing Bias in Assessment at 2021 Academic Internal Medicine Week