Analysis for Racial and Gender Disparities in the Stanford School of Medicine Clerkship Evaluation System

Tracy Rydel MD, Irina Russell PhD, Haley Hedlin PhD, Daniel Bernstein MD, Neil Gesundheit MD
Stanford University School of Medicine
tarydel@stanford.edu

Objectives

Analyze the current clerkship grading system at Stanford Medicine to verify equity. Specifically, estimate whether gender or race/ethnicity are associated with the assignment of Pass With Distinction (PWD) scores when accounting for covariates.
Find ways to improve RIME[1] and professionalism frameworks to ensure excellence and equity in assessment for future trainees.

Background

Stanford School of Medicine completed an analysis of our assessment system, the Criterion-Based Evaluation System (CBES), from 2011-2019, investigating for gender- and race-based disparities.

Since January 2022, the USMLE Step 1 examination is now pass/fail[2]. In January 2021, Step 2 Clinical Skills was eliminated[3]. With these two changes, accurate clerkship assessments will be playing an increased role in residency applications. Concurrently, several medical schools have presented their findings on disparities and bias within their systems of medical student evaluation[4,5,6]. Medical schools nationwide must re-examine their methods of assessment of medical students.

Materials and Methods

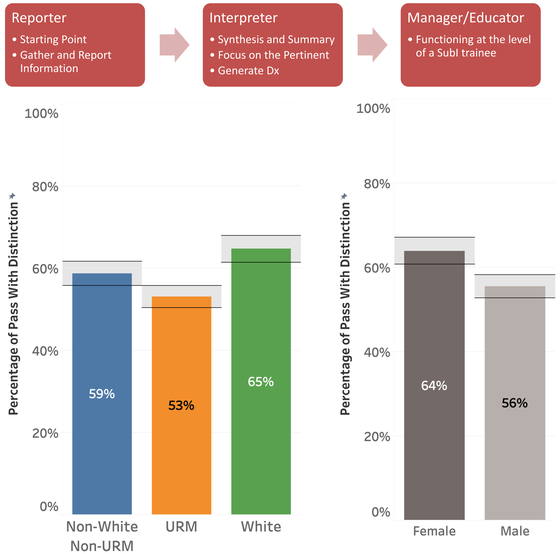
Stanford CBES awarded Pass with Distinction (PWD), Pass, or Fail in 3 domains: clinical skills, professionalism, and medical knowledge.
Our data analysis (N=736) assessed relative risks of receiving PWD in these domains based on student gender and race/ethnicity (White, Underrepresented in Medicine[7], and non-Underrepresented in Medicine). A multivariable analysis controlled for age at time of clerkship, specific clerkship, years since matriculation, rotation order, and Step 1 score.

	Overall	Female	Male
n	736	397	339
Non-Underrepresented in Medicine (Non-URM)	326 (44.3)	192 (48.4)	134 (39.5)
Underrepresented in Medicine (URM)	131 (17.8)	57 (14.4)	74 (21.8)
White	279 (37.9)	148 (37.3)	131 (38.6)

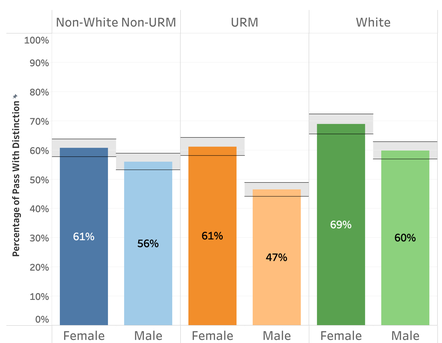
Results

We found statistically significant gender differences in clinical skills and professionalism domains. There were differences based on race/ethnicity in all 3 domains, with clinical skills and medical knowledge domains showing statistical significance and the professionalism domain approaching statistical significance. There was no interaction between gender and race/ethnicity in patient care or professionalism domains.

Patient Care Assessment – RIME framework

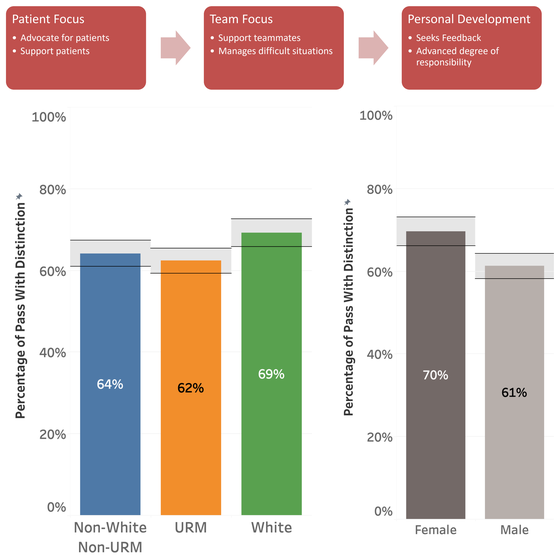


- **Patient care**: differences of ~7-12% in race/ethnicity (Relative Risk (95% CI) p = 0.010); ~8% difference in gender (RR (95% CI) p < 0.001)

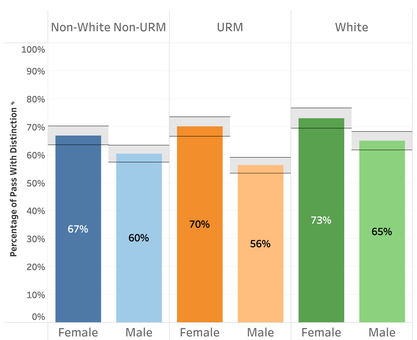


- **Patient care**: Differences appear to be additive, but no interaction between gender and race/ethnicity was shown (RR, 95% CI, p = 0.235)

Professionalism Assessment – Internal Tool



- **Professionalism**: modest differences of ~6% in race/ethnicity (Relative Risk (95% CI) p = 0.053); ~10% difference in gender (RR (95% CI) p < 0.001)



- **Professionalism**: Differences appear to be additive, but no interaction between gender and race/ethnicity was shown (RR, 95% CI, p = 0.325)

Discussion

These data, collected during a 9-year period in over 700 students, illustrate that using a validated instrument, anchors to evaluate professionalism result in modest (6-12%,) differences between URM, non-URM/non-White, and White students in the assignment of honors grades. We expect that refinement of assessment tools and improved rater training has the potential to reduce or eliminate racially-based grading disparities in future. Both domains showed differences favoring female gendered students by ~13-15%. We plan to also further investigate root causes of these gender differences.

We found no interaction between race/ethnicity and gender but found an additive effect when the variables (gender and race/ethnicity) are combined. Going forward, we plan to study whether disparities shown increased or decreased over the 9-year study period.

Future Directions

Residency program directors, undergraduate medical education faculty, and students are grappling with optimizing systems for assessments. At Stanford, we have developed working groups involving clerkship faculty, evaluation team members, advising deans, program directors, and students to determine next steps in assessing learning and maximizing success in the residency match. We believe criterion-based systems of assessment with well-established rubrics are strong but can be further improved with deeper investigation. We are in the process of refining the RIME framework in accordance with current findings of variability of PWD awarding over time; we aim to determine if faculty development and behavioral anchors can bolster this validated instrument and result in fewer disparities.

References

1. Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations: *Academic Medicine*. 1999;74(11):1203-1207. doi:10.1097/00001888-199911000-00012
2. United States Medical Licensing Examination. Change to pass/fail score reporting for Step 1. https://www.usmle.org/incus/#decision. Retrieved Jun 12 2021.
3. United States Medical Licensing Examination. Work to relaunch USMLE Step 2 CS discontinued. https://www.usmle.org/announcements/?ContentId=309. Retrieved Jun 12 2021.
4. Colson ER et al. Washington University School of Medicine in St. Louis case study: a process for understanding and addressing bias in clerkship grading. Acad Med. 2020; 95(12): S131-S135.
5. Low D et al. (2019) Racial/Ethnic Disparities in Clinical Grading in Medical School, Teaching and Learning in Medicine, 31:5, 487-496.
6. Teherani A et al. 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.
7. https://www.aamc.org/what-we-do/equity-diversity-inclusion/underrepresented-in-medicine. Last accessed April 5, 2023.