Original Research Paper

A cross sectional 15-year analysis of the effect of student and faculty demographics on student assessment

Asia Bright[1], Matthew Decaro[2], Heath Goodrum[3], Deukwoo Kwon[4], Mohammad Rahbar[5], Mark Hormann[6], Elmer Bernstam[7], and Jennifer Swails[8]

[1]PhD, Assistant Professor of Psychiatry and Behavioral Sciences and Director of Professionalism, University of Texas Health Science Center at Houston, Houston, United States
[2]Programmer-analyst, McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, United States
[3]MS, Data scientist, Health Transformations Institute, University of Texas Health Science Center at Houston. Houston, United States
[4]PhD, Associate professor, Division of Clinical and Translational Sciences, Department of Internal Medicine, McGovern Medical School, the University of Texas Health Science Center at Houston, Houston, United States
[5]PhD, Professor, Division of Clinical and Translational Sciences, Department of Internal Medicine, McGovern Medical School, and Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, the University of Texas Health Science Center at Houston, Houston, United States
[6]MD, MSE, MS, Reynolds and Reynolds professor of clinical informatics, associate dean for research, McWilliams School of Biomedical Informatics at Houston and professor, Department of Internal Medicine, McGovern Medical School, the University of Texas Health Science Center at Houston, Houston, United States
[7]MD, MSE, MS, Reynolds and Reynolds professor of clinical informatics, associate dean for research, McWilliams School of Biomedical Informatics at Houston and professor, Department of Internal Medicine, McGovern Medical School, the University of Texas Health Science Center at Houston, Houston, United States
[8]MD, Professor and Vice Chair for Education, Department of Medicine, Emory University School of Medicine, Atlanta, United States

Abstract

Background: Medical school assessments impact residency placement and professional progression. Recently, standardized tests found to be highly impacted by demographics have been de-prioritized in an effort to focus on faculty-assessed competency. However, small studies suggest that faculty assessment may also be affected by demographic factors. The current study examines: 1) how student race and ethnicity are associated with standardized medical knowledge assessments; and 2) whether demographics are associated with faculty evaluations. **Method:** Using a 15-year data set (2007–2022) from a large academic medical center in the United States, generalized mixed regression models were conducted to determine whether student race/ethnicity was a predictor of: 1) failing National Board of Medical Examiners (NBME) score (<60%); 2) exceptionally good NBME score (>90%); or 3) low performance evaluations (LPE, failing evaluations). Covariables included age, gender, and rotation characteristics such as academic month, academic year, and subject. **Results:** Compared to White students, other racial groups had higher odds of failing the NBME (Asian OR= 2.24 [1.69, 2.93], $p < 0.001$; Hispanic students OR=1.59 [1.14, 2.23]; $p = 0.003$; Black OR=4.64 [3.45, 6.24]; $p < 0.001$). Compared to White students, all other races/ethnicities had lower odds to score 90 or above. Asian students had higher odds to receive an LPE from an Asian faculty member (OR 1.75 [1.03, 2.97]; $p = 0.04$). Black students had higher odds to receive an LPE from a White faculty member (OR 2.43 [1.46, 4.04]; $p < 0.001$). Compared to White students, Black, Asian, and Hispanic students had higher odds of failing the NBME than they were to receive an LPE. **Conclusions:** Demographic differences affected both faculty evaluations and exam scores, and were not mitigated by racial concordance between learners and evaluators. More work is needed to develop strategies that promote an equitable learning environment.

Keywords: student performance, faculty, race/ethnicity, nested regression, standardized test scores

Date submitted: 29-September-2025

Email: Jennifer Swails (jswails@emory.edu)

Citation: Bright A, Decaro M, Goodrum H, Kwon D, Rahbar M, Hormann M, Bernstam E, and Swails J. A cross sectional 15-year analysis of the effect of student and faculty demographics on student assessment. Educ Health 2026;39:38-48

Online access: www.educationforhealthjournal.org
DOI: 10.62694/efh.2026.476

Published by The Network: Towards Unity for Health

Background

Clerkship grades affect medical student competitiveness for selective residency placements, their psychological wellbeing, and also professional progression. In clinical clerkships, students can be assessed using several measures including nationally standardized examinations (e.g., National Board of Medical Examiners subject examinations or NBME), local examinations (e.g. Observed Clinical Skills Exams) and clinical performance assessments by faculty and residents. Recently, there has been increased attention on the impact of demographic factors that may inappropriately influence outcomes of these assessments,[1,2] especially national standardized examinations. Studies indicate that individuals from underrepresented minority (URM) groups (including those of Black, Hispanic, Native American, or native-Alaskan origin, or any combination thereof) have a higher likelihood of receiving lower grades across various clerkships, such as internal medicine, surgery, obstetrics and gynecology, pediatrics, neurology, and psychiatry.[3] In a separate study analyzing longitudinal data spanning five years, White students in their third-year clerkship achieved the highest percentage of Honors across all clerkships (ranging from 34% to 46%), whereas URM students achieved percentages ranging from 16% to 40%. Moreover, White students received more "Outstanding" ratings (71%) in their medical student performance evaluations (MSPE) than URM students (3%).[4]

In line with the Liaison Committee on Medical Education (LCME) requirement that medical school "includes a comprehensive, fair, and uniform system of formative and summative medical student assessment,"[5] measures influenced by demographics have been removed from grading schemes. At some institutions, this has led to a shift away from nationally standardized examinations toward locally designed and administered skill assessments. Educators hoped this would focus learning on the clinical environment and show reduced bias when paired with implicit bias training as well as carefully anchored grading rubrics. However, studies from single institutions and clerkships demonstrated that the local standardized assessments may also show demographic variability.[6,7] Increasingly, medical schools moved toward a pass-fail grade system. Although there are data that students are less stressed and may adopt a growth mindset with this change,[8] they may also struggle to distinguish themselves for residency and feel pressure to be involved in a "shadow economy" of extracurricular projects (research and otherwise) that takes them away from clinical learning.[8]

Given the large-scale changes happening within medical education, the relationship between demographics and rotation evaluations warrants thorough, longitudinal exploration across clerkship specialties and faculty-student pairings, in the context of an institution engaged in implicit bias training and competency-based medical education. In the healthcare literature, numerous studies demonstrate that race concordance between physicians and patients leads to favorable clinical outcomes, including increased utilization of necessary health services, reduced delays in seeking care, and lower total healthcare expenditures.[9–11] A study examining the effects of faculty–resident gender and under-represented minority (URM) status concordance on faculty teaching found that URM status concordant pairs receive the highest ratings.[12] It is not clear how demographic concordance might impact student evaluations during clerkships.

In the current study, we examined the demographic predictors of standardized and subjective components of clinical clerkship grades. More specifically, we investigated whether student race/ethnicity is associated with likelihood of : (1) failing the NBME, or (2) exceptionally passing the NBME (score $\geq$90), or (3) whether student and faculty race/ethnicity predicts receiving a low performance evaluation (LPE). Based on the available published work, we hypothesized that (1) there is a strong correlation between student and faculty race/ethnicity and performance on both standardized and subjective measures; and (2) faculty members would have a greater likelihood of giving an LPE to a student who is not of the same race/ethnicity.[13–15] To test these hypotheses, we analyzed a large data set with 15 years of medical student data from a large academic medical center with engagement in faculty development (both for implicit bias and competency-based assessment) in the southern United States.

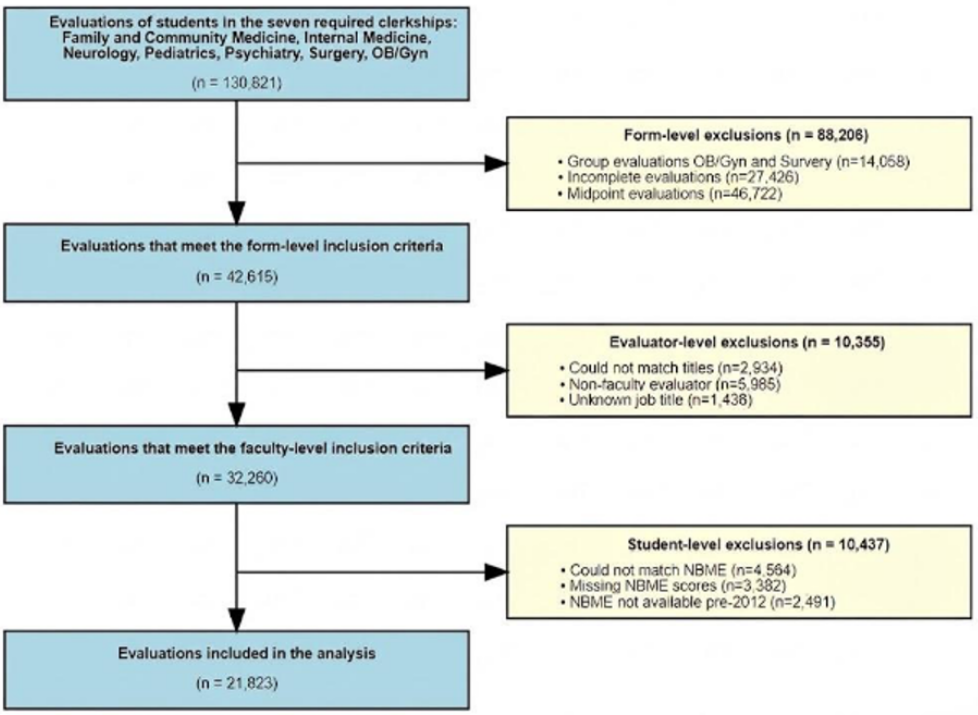
Methods

We analyzed data from five required third-year medical student clerkships (family and community medicine, internal medicine, neurology, pediatrics, and psychiatry) over a 15-year period (August 2007–June 2022) from a single institution. This study has been approved by the Committee for the Protection of Human Subjects (The University of Texas Health Science Center at Houston Institutional Review Board) under protocol HSC-SBMI-19-0385.

Figure 1 details the selection process from the initial dataset of available evaluations (130,821) to the final set of evaluations to be used in the analysis (21,823). Two required clerkships (obstetrics/gynecology and surgery) and summary evaluations from the family medicine clerkship that used group/consensus evaluations were excluded because individual faculty characteristics (such as faculty rank) previously found to be associated with evaluation outcomes were not assessable for these rotations.[16] Rotation-specific NBME subject examination scores were compared to faculty Clinical Performance Evaluations (fCPE) for the same rotation.

fCPE data included all final evaluations completed by individual faculty members. Midpoint evaluations were excluded due to inconsistent grading criteria. For example, some faculty members documented sub-standard midpoint but not final performance for most students. Next, we applied evaluator-level exclusions. The evaluation must have been performed by faculty with one of the following titles: assistant, associate, or full professor. Evaluations completed by non-faculty (e.g., residents), employees whose titles were not known, or who had other titles such as instructor (very rare), and staff physicians (non-faculty) were excluded. Lastly, we applied a set of student-level exclusions. fCPE data were matched to NBME scores by student name and time (shelf-exam within six months of rotation). Pairs of students who had the same first and last name were excluded from the analysis because their evaluations could not be linked to their NBME scores. Additionally, the student must have had a full set of fCPE outcomes available (in earlier years, oral presentation scores were not available), which are combined into a single measure (cognitive score) for the final model.

Figure 1: CONSORT diagram that describes the inclusion/exclusion criteria for evaluations to be used in the analysis in three categories: form-level exclusion, evaluator-level exclusion, and student-level exclusion. The final number of evaluations available for the analysis was 21,823.



Student variables

Student demographics (age, gender, ethnicity and race), clerkship specialty (e.g., neurology, internal medicine), rotation month and year were all included in the statistical models. Unfortunately, race and ethnicity were conflated in our data. For example, our institution defines Hispanic as both a race and an ethnicity, and a student can identify racially as White and ethnically Hispanic or non-Hispanic. The possible race categories were (in decreasing order of frequency): White, Asian, Hispanic, and Black. Other racial categories, such as American Indian and students that identify as multiple races, were excluded from the analysis due to a lack of available data.

Faculty variables

Faculty characteristics included age, gender, ethnicity/race, academic rank, and experience. We previously found that faculty characteristics including rank and gender correlated with the willingness to document poor student performance.[16] In addition, we analyzed faculty/student concordance with respect to race and gender. Due to the limited amount of faculty data available, the racial categories Hispanic and Black were combined into a single category, "Other" to ensure convergence of the statistical model. We distinguished faculty rank from experience.[16] Experience was calculated as the number of evaluations completed by a faculty member at the time of a particular evaluation. For example, a faculty member submitting their 19th evaluation of a student would be assigned an experience value of 19 for that evaluation, and a value of 25 for the 25th evaluation. We grouped experience into three categories: low (<50), intermediate (>=50 and <100), and high (>=100).

Standardized Medical Knowledge Assessments (NBME scores)

We downloaded all available NBME scores for the study period. For most subjects, NBME scores were available for the entire 15-year period (2007–2022), however for family medicine the test scores (0–100) were only available for 10 years (2012–2022). Typically, students take the exam within a few weeks of completing the relevant rotation. Students who failed were required to re-take the exam. However, we considered only the student's first attempt with one NBME score per student for each rotation. NBME outcomes were used to define two outcomes representing negative and positive outliers, failing and exceptional performance respectively. NBME scores less than 60 were defined as failing; approximately in line with NBME cutoffs that varied very slightly (1–2 points) over the years. We defined an NBME score of 90 or above as exceptional performance because this corresponded to approximately the top 10% of scores.

Faculty Clinical Performance Assessments (fCPE)

For each rotation, faculty completed evaluations for students that they had supervised. Each student received multiple evaluations during the third-year clerkships, and most faculty evaluated multiple students during the study period. Students could request specific faculty, but most assignments were random. In contrast to NBME scores, each student had a variable number of fCPEs per rotation, depending on the rotation length, as well as overlap between student and faculty schedules.

Evaluation forms varied over the 15-year study period. Final evaluation had different numbers of question items per form. Evaluation questions were either binary (e.g., "Do you have any serious concerns about this student's ethical or professional behavior?" —Yes/No) or used 3 to 5-point Likert scales. All evaluation questions were based on behavioral anchors, which were available to both faculty and students prior to the rotation. All faculty involved in student evaluation received yearly in-person training on the use of behavioral anchors throughout the study period. Education regarding implicit bias was also required as a part of yearly online modules.

Because the distribution of faculty evaluations was skewed toward the maximum with very low variability, there was little difference between students. Therefore, analyzing percentile scores derived from the sum of individual component scores would not yield meaningful results.

We defined a "low performance evaluation" or LPE. Intuitively, LPEs represent evaluations where the evaluator determined that the student did not meet expectations, i.e., a failing or near-failing

evaluation. We manually reviewed the evaluation forms completed each year to determine what answers defined LPE as a binary outcome for that form.

We defined an LPE as one or more of the following:[16]

1. An overall failing grade on the final evaluation—when an overall question existed in the evaluation form (defined as value of 1 or 2, on a 5-point Likert scale)
2. One or more scores below the center score on any Likert scale for performance (for example, a value of 1 or 2, on a 5-point Likert scale) on the final evaluation
3. Flagged for unethical behavior on the midpoint or final evaluation

Analysis

We reported descriptive statistics for covariates: mean and standard deviation for continuous variables; count and percentage for categorical variables. Due to the nature of the correlated data (multiple rotations for each student), we used a generalized linear mixed effects model to estimate the association between student outcomes and study covariates. Due to the random assignment of evaluators to students, evaluators are modeled as a random effect, whereas other predictors used are treated as fixed effects. We reported adjusted odds ratios (ORs), corresponding 95% confidence intervals and p-values.

The binary outcome variables for the analyses were:

1) LPE vs. not for each evaluation
2) NBME fail (<60)
3) Exceptionally good NBME score (≥90)

Independent variables that had significance in the univariate analyses were included in the multivariable models (Table 1). In addition, we evaluated interaction effects between gender, race/ethnicity, age and faculty rank as potential confounders or effect modifiers. Results are presented as Odds Ratio (OR) along with 95% confidence intervals (CI). Statistical modeling was conducted using SAS version 9.4, and analysis of both the data and models (such as checking for multi-collinearities and the assumption of linearity between independent variables and the log-odds) were conducted using Python version 3.9. Statistical tests were two-sided and significance was set at $\alpha=0.05$

Results

The final dataset included 21,823 evaluations completed by 537 faculty evaluators on 2,766 students between August 2007 and June 2022 with a total of 326 (1.5%) LPEs. Table 1 shows the student and faculty characteristics for completed evaluations.

The student race/ethnicity breakdown changed significantly over time ($p < 0.001$) with an increase in Asian, Hispanic and Black representation (Figure 2). There were a total of 10,981 NBME scores, of which 154 (1.4%) scores were below 60 (failing), and 1,278 (11.6%) scores were 90 or above (exceptionally good).

Variables that were significant in a univariate model were included in a multivariable model (Table 2). In the first analysis (Table 2a), LPE was the outcome; student and faculty characteristics, NBME scores, rotation (subject) and year were the predictors. Assistant and associate professors gave LPEs at a rate of 1.4%, full professors gave LPEs at a rate of 1.9%. The rate of LPE decreased with faculty experience (low: 1.9%, middle: 1.3%, high 0.8%). White students received 50% of LPEs, Asian students received 20%, Hispanic students received 18% and Black students received 12%. Black students, but not Asian or Hispanic students, had higher odds of receiving an LPE compared to White students (OR=1.74 [1.16, 2.61]; $p = 0.003$). The rate of LPEs is decreasing over time for White, Asian, and Hispanic students. For Black students, there is no significant trend over time (-0.12% yearly, $p=0.307$). For White students, it decreases by 0.11% per year ($p=0.004$), for Asian students 0.22% ($p=0.001$), and for Hispanic students 0.26% ($p=0.003$) (Figure 3). White faculty gave 47% of LPEs, Asian faculty gave 28% and "Other" (Black or Hispanic) faculty gave 25%.

We found that White faculty had higher odds of giving LPEs to Black students (OR 2.43 [1.46, 4.04]; $p < 0.001$). In contrast, Asian faculty had

Table 1: Descriptive statistics of student and faculty characteristics in final evaluations. A checkmark indicates the inclusion of a variable in the model.

	Final evaluations only Mean (SD) or n (%)	Multivariate Model 1 (LPE)	Multivariate Model 2 (NBME <60)	Multivariate Model 3 (NBME >=90)
NBME Score	79 (8.85)	✓		
Subject (Rotation)		✓	✓	✓
Neurology	5,130 (23.5%)			
Internal Medicine	5,103 (23.4%)			
Psychiatry	4,146 (19.0%)			
Pediatrics	3,908 (17.9%)			
Family Medicine	3,536 (16.2%)			
Cognitive Score (fCPE)			✓	✓
Quartile 4	5,428 (24.9%)			
Quartile 3	5,469 (25.1%)			
Quartile 2	4,567 (20.9%)			
Quartile 1 (lowest)	6,359 (29.1%)			
Academic Month			✓	✓
Academic Year		✓		
Student Age	25 (2.50)		✓	✓
Student Gender		✓	✓	
Male	12,540 (57%)			
Female	9,283 (43%)			
Student Race		✓	✓	✓
White	12,632 (58%)			
Asian	4,478 (21%)			
Hispanic	3,199 (15%)			
Black	1,514 (0.7%)			
Faculty Age	52 (12.6)			
Faculty Gender		✓		
Male	10,952 (50.1%)			
Female	10,871 (49.9%)			
Faculty Race				
White	10,109 (46.2%)			
Asian	6,524 (30.0%)			
Other	5,190 (23.8%)			
Faculty Rank			✓	
Assistant	14,880 (68.2%)			
Associate	3,343 (15.3%)			
Full Professor	3,600 (16.5%)			
Faculty Experience		✓		
Low	9,279 (42.5%)			
Middle	5,096 (23.4%)			
High	7,448 (34.1%)			

Figure 2: The distribution of students by race over time. Over a twelve-year period, the relative fraction of Asian, Hispanic, and Black students increased by 1.2% (p < 0.001), 0.4% (p = 0.007), and 0.3% (p < 0.001) each year, respectively. Total enrollment remained stable.

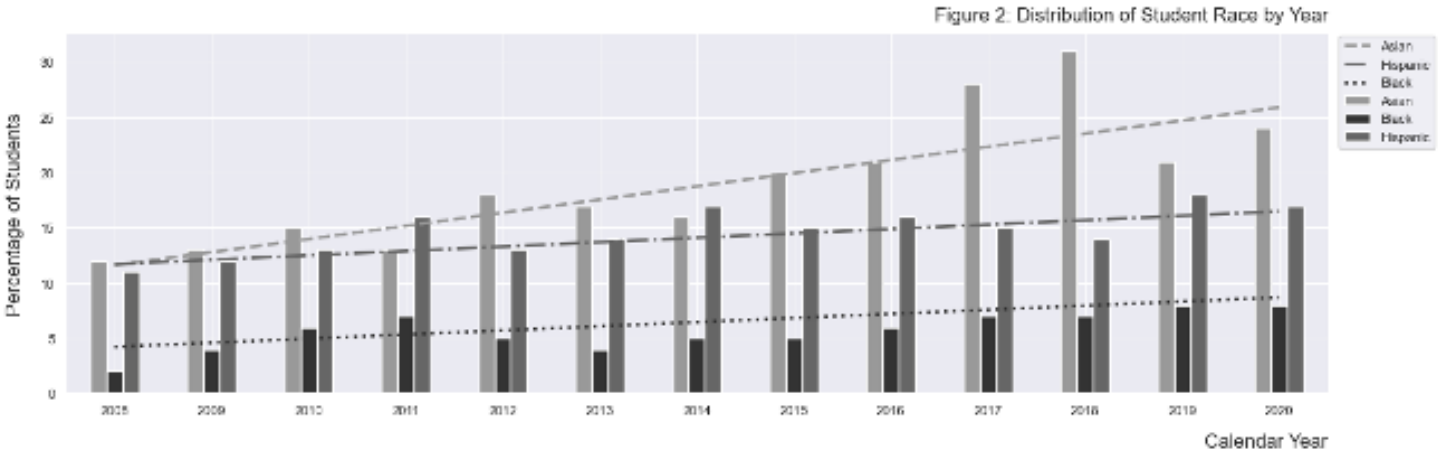
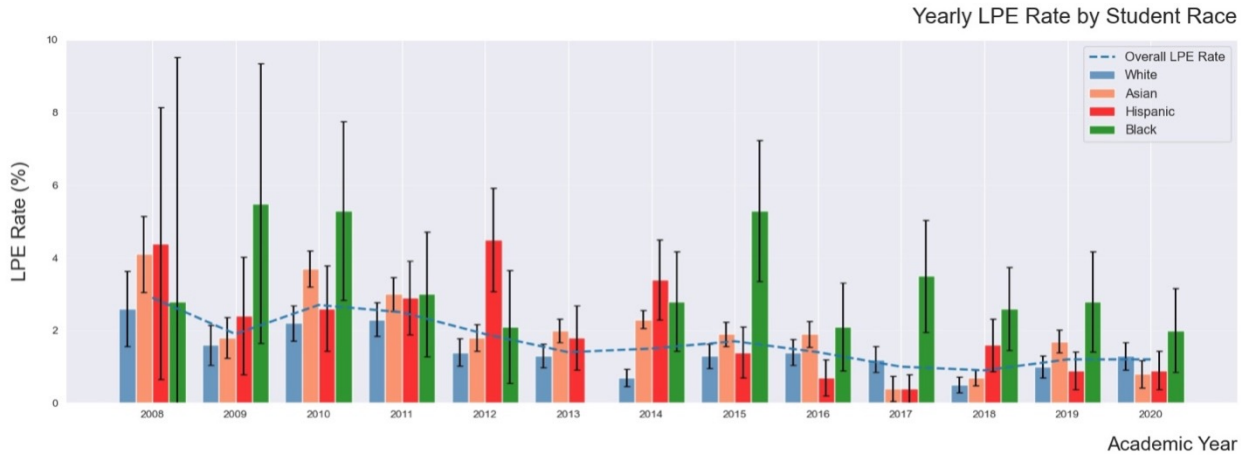


Figure 3: Distribution of LPE by Race over time. Over a 12-year period, despite reduction in LPE overall, demographic variability persisted. (White fitted slope: -0.115, p = 0.004; Asian fitted slope: -0.22, p = 0.001; Hispanic, fitted slope: -0.264, p = 0.003; Black, fitted slope: -0.123, p = 0.307)



higher odds of documenting LPE for Asian students, compared to students of other races/ethnicities.

In the second analysis (Table 2b), we estimated the probability of failing the subject NBME (score <60). Student characteristics, rotation (subject), month and the fCPE were the predictors. Roughly two percent of NBME scores were below 60. Compared to White students, Asian students had more than twice the odds of failing the NBME (OR 2.24 [1.69, 2.93], p < 0.001), Hispanic students also had higher odds of failing (OR=1.59 [1.14, 2.23]; p = 0.003), and Black students had more than four times higher odds of failing (OR=4.64 [3.45, 6.24]; p < 0.001). The distribution of NBME scores by race is shown in Supplemental Figure 3.

Finally, we estimated the probability of NBME success (score≥90, Table 2c); student characteristics, rotation (subject), month and fCPEs were the predictors. Compared to White students, all other races/ethnicities had lower odds of scoring 90 or above; all results were statistically significant with p<0.001 (table 2c). This model did not include interaction effects due to statistical non-significance. For each race category, the relationship between cognitive score (quartile) and NBME score were similar (supplement Figure 2).

Discussion

In our sample, White students received 50% of all LPEs, followed by Asian (20%), Hispanic (18%), and Black students (12%). Consistent with past studies, Black students had significantly higher

Table 2: Results of the multivariate models to predict NBME fail (2a), NBME success (2b) and LPE (2c).

Table 2a	**OR of NBME failure [95% CI]**	
Student Race		<0.001*
White	Ref	
Asian	2.24 [1.69, 2.93]	< 0.001
Hispanic	1.59 [1.14, 2.23]	0.003
Black	4.64 [3.45, 6.24]	< 0.001
Table 2b	**OR of NBME success [95% CI]**	
Student Race		<0.001*
White	Ref	
Asian	0.48 [0.42, 0.54]	< 0.001
Hispanic	0.39 [0.33, 0.45]	< 0.001
Black	0.16 [0.12, 0.22]	< 0.001
Table 2c	**OR of LPE [95% CI]**	**p-value**
Student Race		0.06*
White	Ref	
Asian	1.09 [0.78, 1.52]	0.83
Hispanic	1.19 [0.83, 1.72]	0.40
Black	1.74 [1.16, 2.61]	0.003
Student Race * Faculty Race	Ref = White	0.18*
Faculty (White), compare students		
Asian	0.81 [0.49, 1.35]	0.43
Hispanic	1.54 [0.99, 2.41]	0.06
Black	2.43 [1.46, 4.04]	< 0.001
Faculty (Asian), compare students		
Asian	1.75 [1.03, 2.97]	0.04
Hispanic	1.60 [0.89, 2.89]	0.12
Black	1.56 [0.73, 3.35]	0.25
Faculty (Other), compare students		
Asian	0.90 [0.50, 1.63]	0.73
Hispanic	0.69 [0.33, 1.46]	0.33
Black	1.38 [0.65, 2.94]	0.41

*indicates p value for overall significance of the variable (e.g., student race) in the multivariable model, as opposed to comparisons of subcategories within the variable (e.g., Asian to White, etc.)
When students are separated by race, there were no significant changes over time in NBME scores within each group from 2008 through 2020 (supplement Figure 1).

odds of receiving an LPE compared with White students (OR=1.74 [1.16, 2.61]; p = 0.003), supporting prior evidence that Black learners experience disproportionate negative workplace-based evaluations.[3,4] Prior research has also shown that structural inequities, differential access to learning opportunities, and implicit bias can contribute to these disparities, and our findings align with this body of work.[6] In contrast, Asian and Hispanic students did not show increased odds of receiving an LPE relative to White students. This finding occurs in the context of mixed evidence in the existing literature and provides a more nuanced understanding of how demographic patterns vary by racial and ethnic group.[12] The observation that LPE rates for White, Asian, and Hispanic students declined over time also supports emerging reports that faculty development efforts and evolving evaluation practices may be reducing the overall frequency of low ratings, though not uniformly across groups.[17]

Our findings related to NBME performance similarly support longstanding trends: Asian, Hispanic, and especially Black students had higher odds of failing compared with White students, with Black students exhibiting more than four times the odds of failure (OR=4.64 [3.45, 6.24]; p < 0.001). This pattern has emerged extensively in past research of racial/ethnic disparities on standardized examinations and is often attributed to structural educational inequities, test design factors, and stereotype threat.[14] All non-White groups also had significantly lower odds of scoring 90 or above, supporting evidence that these disparities extend across the full performance spectrum.

What is unique about this study is the ability to compare demographic patterns in both NBME scores and faculty-assigned LPEs within the same large, longitudinal sample, allowing direct evaluation of how standardized and workplace-based assessments may diverge or reinforce each other. While prior studies have typically examined these assessment types separately or with smaller datasets, our findings show that demographic differences are more pronounced in NBME performance than in LPEs, yet the two forms of assessment remain strongly associated across groups. This integrated analysis provides new insight into how inequities propagate across multiple components of clerkship grading and highlights implications for grading policies that rely heavily on national exam scores.

Strengths and Limitations[18,19,20]

This study includes a large sample size of students and faculty over a 15-year study period as well as a relatively diverse student body, particularly with respect to Hispanic representation (15%). Our data also included standardized tests that are used by many US medical schools, increasing the likelihood that our results are generalizable to other institutions. Sophisticated hierarchical modelling allowed us to control multiple potential confounders including both evaluator (faculty) and student factors, subject (e.g., internal medicine vs neurology), academic year, and academic month. However, while the study provides a basic disaggregation of race/ethnicity (White, Black, Asian, Hispanic, Other), there are some categories that remain too broad to capture the diversity of racial and ethnic backgrounds; for example. the category of "Asian," which mixes underrepresented groups with overrepresented groups. We are limited to the racial categories provided by the registrar and the convergence of our model. Also, our limited number of Black faculty did not allow us to determine the effect of race concordance/discordance between faculty and student for this population. Causal inferences between NMBE, faculty evaluations and student and faculty race/ethnicity cannot be derived from this single-institution correlational study, and generalizability to other institutions may be limited.

Conclusion[20,21]

This is the largest published study evaluating assessment outcomes at one large, diverse, public medical school, and it shows persistent differences in outcomes across demographic groups. The inequities found in this study persisted despite initiatives to reduce implicit bias, create a positive learning environment, and improve assessments that were ongoing during the time period of this study, including an accreditation site visit and a pilot of Entrustable professional activities through the Association of American Medical Colleges.

Implications

Concerns regarding inequity in medical student assessment are not new. An equitable assessment system provides fair opportunities for advancement to all learners without influence from structural, personal, or demographic factors.[20] Fair assessment involves the instruments used, the learning environment, and the way that outcomes influence career advancement. For two decades, the healthcare education community has been working to improve the system by promoting competency-based medical education, faculty development, and improve assessment tools.[21] Previous literature has established that race-concordance between physician and patient improves outcomes among underrepresented groups. When demographic variability in medical student assessment doesn't correlate with clinical outcomes, it suggests an opportunity to align educational programs with patient needs, especially for students underrepresented in medicine. It is not clear if the current approach, in which grades have been removed[22] and proxies (especially publications) have become markers of excellence for competitive subspecialties,[23] will promote more equity or alignment with patient needs. Just as this work shows that shifting focus from national examinations to localized assessments did not remove bias, it is unknown if there is equitable access to scholarship across demographic and socioeconomic groups. As recommended in the Coalition for Physician Accountability UME-GME Recommendations,[23] it is essential to monitor the impact of assessment changes, including NBME, pass-fail grading, and the shift toward scholarship, on vulnerable student groups, acknowledging the complex system of interactions.[24] Educators have an opportunity to design and study innovative solutions of support and assessment to promote and recognize excellence in learners that is aligned with the patient care they will deliver.

List of Abbreviations

Diversity, equity, inclusion (DEI)
Association of American Medical Colleges (AAMC)
Competency-based educational (CBE)
Medical student performance evaluation (MSPE)
Underrepresented minority (URM)
Low performance evaluations (LPE)
National Board of Medical Examiners (NBME)
Odds ratios (ORs)
Faculty Clinical Performance Assessment (fCPE)

Statements & Declarations

Funding: This work was supported by NIH/NCATS grant number UL1 TR003167, UM1 TR004906, UM1 TR004539 and the Reynolds and Reynolds Professorship.

The authors have no relevant financial or non-financial interests to disclose.

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Matthew Decaro, Heath Goodrum, Deukwoo Kwon, and Mohammad H. Rahbar. The first draft of the manuscript was written by Asia Bright and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

This study has been approved by the Committee for the Protection of Human Subjects (The University of Texas Health Science Center at Houston Institutional Review Board) under protocol HSC-SBMI-19-0385.

References

1. Solomon SR, Atalay AJ, Osman NY. Diversity Is Not Enough: Advancing a Framework for Antiracism in Medical Education. *Academic Medicine*. 2021;96(11):1513-1517. https://doi.org/10.1097/ACM.0000000000004251

2. Onumah CM, Lai CJ, Levine D, Ismail N, Pincavage AT, Osman NY. Aiming for Equity in Clerkship Grading: Recommendations for Reducing the Effects of Structural and Individual Bias. *The American Journal of Medicine*. 2021;134(9):1175-1183.e4. https://doi.org/10.1016/j.amjmed.2021.06.001

3. Lee KB, Vaishnavi SN, Lau SKM, Andriole DA, Jeffe DB. "Making the grade:" noncognitive predictors of medical students' clinical clerkship grades. *Journal of the National Medical Association*. 2007;99(10):1138-1150. PMID: 17987918 PMCID: PMC2574397

4. Low D, Pollack SW, Liao ZC, Maestas R, Kirven LE, Eacker AM, Morales LS. Racial/Ethnic Disparities in Clinical Grading in Medical School. *Teaching and Learning in Medicine*. 2019;31(5):487-496. https://doi.org/10.1080/10401334.2019.1597724

5. Standards, Publications, & Notification Forms | LCME. Retrieved August 19, 2025 from https://lcme.org/publications/

6. Card D, McPherson L, Marka N, Langworthy B, Mustapha T, Violato C, Englander R, Stork Poeppelman R. Sex and racial bias in medical student EPA assessments: Findings and hypotheses for bias mitigation targets. *Medical Teacher*. 0(0):1-10. https://doi.org/10.1080/0142159X.2025.2502162

7. Brown C, Khavandi S, Sebastian A, Badger K, Westacott R, Reed MWR, Gurnell M, Sam AH. The influence of candidates' race on examiners' ratings in standardised assessments of clinical practice. *Medical Teacher*. 2025;47(3):492-497. https://doi.org/10.1080/0142159X.2024.2345266

8. Iyer AA, Hayes C, Chang BS, Farrell SE, Fladger A, Hauer K, Schwartzstein RM. Should Medical School Grading Be Tiered or Pass/Fail? A Scoping Review of Conceptual Arguments and Empirical Data. *Academic Medicine*. 2025;100(8):975-985. https://doi.org/10.1097/ACM.0000000000006085

9. Jetty A, Jabbarpour Y, Pollack J, Huerto R, Woo S, Petterson S. Patient-Physician Racial Concordance Associated with Improved Healthcare Use and Lower Healthcare Expenditures in Minority Populations. *Journal of Racial and Ethnic Health Disparities*. 2022;9(1):68-81. https://doi.org/10.1007/s40615-020-00930-4

10. Alsan M, Garrick O, Graziani GC. Does Diversity Matter for Health? Experimental Evidence from Oakland. *American Economic Review* 109 (12): 4071–4111**.** https://doi.org/10.1257/aer.20181446

11. Greenwood BN, Hardeman RR, Huang L, Sojourner A. Physician-patient racial concordance and disparities in birthing mortality for newborns. *Proceedings of the National Academy of Sciences U S A*. 2020;117(35):21194-21200. https://doi.org/10.1073/pnas.1913405117

12. McOwen KS, Bellini LM, Guerra CE, Shea JA. Evaluation of clinical faculty: gender and minority implications. *Academic Medicine*. 2007;82(10 Suppl):S94-96. https://doi.org/10.1097/ACM.0b013e3181405a10

13. Rubright JD, Jodoin M, Barone MA. Examining Demographics, Prior Academic Performance, and United States Medical Licensing Examination Scores. *Academic Medicine*. 2019;94(3):364-370. https://doi.org/10.1097/ACM.0000000000002366

14. Jones AC, Nichols AC, McNicholas CM, Stanford FC. Admissions Is Not Enough: The Racial Achievement Gap in Medical Education. *Academic Medicine*. 2021;96(2):176-181. https://doi.org/10.1097/ACM.0000000000003837

15. Lucey CR, Saguil A. The Consequences of Structural Racism on MCAT Scores and Medical School Admissions: The Past Is Prologue. *Academic Medicine*. 2020;95(3):351-356. https://doi.org/10.1097/ACM.0000000000002939

16. Swails JL, Gadgil MA, Goodrum H, Gupta R, Rahbar MH, Bernstam EV. Role of faculty characteristics in failing to fail in clinical clerkships. *Medical Education*. 2022;56(6):634-640. https://doi.org/10.1111/medu.14725

17. Lupton KL, O'Sullivan PS. How Medical Educators Can Foster Equity and Inclusion in Their Teaching: A Faculty Development Workshop Series. *Academic Medicine*. 2020;95(12S Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments):S71-S76. https://doi.org/10.1097/ACM.0000000000003687

18. Shankar M, Henderson K, Garcia R, et al. Presence 5 for Racial Justice Workshop: Fostering Dialogue Across Medical Education to Disrupt Anti-Black Racism in Clinical Encounters. *MedEdPORTAL*. 18:11227. https://doi.org/10.15766/mep_2374-8265.11227

19. Brooks KC. A Silent Curriculum. *Journal of the American Medical Association*. 2020;323(17):1690-1691. https://doi.org/10.1001/jama.2020.2879

20. Lucey CR, Hauer KE, Boatright D, Fernandez A. Medical Education's Wicked Problem: Achieving Equity in Assessment for Medical Learners. *Academic Medicine*. 2020;95(12S Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments):S98-S108. https://doi.org/10.1097/ACM.0000000000003717

21. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Medical Teacher*. 2010;32(8):676-682. https://doi.org/10.3109/0142159X.2010.500704

22. Agolia J, Green A, Spain DA, Choi J. Diminishing Objectivity in the Residency Application Process. *Journal of the American Medical Association*. 2025;333(12):1034-1035. https://doi.org/10.1001/jama.2024.28397

23. Warm E, Hirsh DA, Kinnear B, Besche HC. The Shadow Economy of Effort: Unintended Consequences of Pass/Fail Grading on Medical Students' Clinical Education and Patient Care Skills. *Academic Medicine*. 2025;100(4):419-424. https://doi.org/10.1097/ACM.0000000000005973

24. Swails JL, Angus S, Barone MA, et al. The Undergraduate to Graduate Medical Education Transition as a Systems Problem: A Root Cause Analysis. *Academic Medicine*. 2023;98(2):180-187. https://doi.org/10.1097/ACM.0000000000005065