

## UC Irvine

### Western Journal of Emergency Medicine: Integrating Emergency Care with Population Health

#### Title

Inequities in the National Clinical Assessment Tool for Medical Students in the Emergency Department

#### Permalink

<https://escholarship.org/uc/item/2900s961>

#### Journal

Western Journal of Emergency Medicine: Integrating Emergency Care with Population Health, 26(5)

#### ISSN

1936-900X

#### Authors

Amin, Bushra Z.  
Dine, C. Jessica  
Tabakin, Erica R.  
[et al.](#)

#### Publication Date

2025-10-04

#### DOI

10.5811/westjem.43506

#### Supplemental Material

<https://escholarship.org/uc/item/2900s961#supplemental>

#### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

# Inequities in the National Clinical Assessment Tool for Medical Students in the Emergency Department

Bushra Z. Amin, MD\*

C. Jessica Dine, MD, MSHP\*†§

Erica R. Tabakin, MD\*\*

Michael Trotter, MD\*\*

Janae K. Heath, MD, MSCE\*\*

\*Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania

†Hospital of the University of Pennsylvania, Department of Medicine, Philadelphia, Pennsylvania

\*\*Hospital of the University of Pennsylvania, Department of Emergency Medicine, Philadelphia, Pennsylvania

§Leonard Davis Institute of Economics at the University of Pennsylvania, Philadelphia, Pennsylvania

Section Editor: Jeffrey Druck, MD

Submission history: Submitted February 20, 2025; Revision received June 3, 2025; Accepted June 12, 2025

Electronically published October 3, 2025

Full text available through open access at [http://escholarship.org/uc/uciem\\_westjem](http://escholarship.org/uc/uciem_westjem)

DOI 10.5811/westjem.43506

**Introduction:** The National Clinical Assessment Tool for Emergency Medicine (NCAT-EM) was designed to standardize medical student assessments during emergency medicine clinical rotations. While multiple assessment tools implemented in medical education have been prone to inequities, it remains unknown how student and rater demographics impact NCAT-EM scores. In this study we examined how a student's gender and status as under-represented in medicine (URM) affected NCAT-EM scores.

**Methods:** This was a retrospective cohort study of all NCAT-EM assessments of clerkship medical students at a single institution in 2022. We performed mixed-effect ordinal logistic regression analyses to determine the association between the seven NCAT-EM domains (history/physical, prioritized differential, formulation of plans, observation/monitoring, emergency management, communication, and global assessment) and student gender, as well as the NCAT-EM domains and students' URM status (specifically in domains of race and ethnicity). We adjusted our analyses for the site of rotation, time, the rater's role (attending or resident), and rater demographics (gender, URM status). We then evaluated the interaction in gender concordance and URM-status concordance on outcomes.

**Results:** A total of 1,881 NCAT-EM assessment forms were submitted on 142 students completed by 266 raters. There were no significant associations between student gender and NCAT-EM ratings across the seven domains. We found an association between URM students and lower scores in multiple NCAT-EM domains, including global assessment (odds ratio [OR] 0.50, CI 0.25-0.99,  $P = .01$ ); history/physical (OR 0.38, CI 0.19-0.77,  $P = .01$ ); and prioritized differential (OR 0.47, CI 0.26-0.88,  $P = .02$ ). This effect was moderated by a significant positive interaction effect with URM concordance between raters and students in the prioritized differential and observation/monitoring domains.

**Conclusion:** This is the first study to highlight differences in both gender and status as under-represented in medicine within the nationally implemented NCAT-EM assessment tool. Women students were overall rated similarly across the NCAT-EM domains compared to men, with no association of gender on ratings. However, students' URM status was associated with lower scores in multiple NCAT-EM domains. This finding was mitigated by URM concordance between faculty and resident raters. Our findings support the need for additional studies to understand bias and inequities in the application of the NCAT-EM tool nationally. [West J Emerg Med. 2025;26(5)1250–1259.]

## INTRODUCTION

A longstanding challenge in medical education has been to accurately assess medical students on clinical rotations with assessment tools that have strong validity and reliability evidence.<sup>1,2</sup> The fairness and accuracy of clinical assessment of medical students is critical as it informs clinical grades, Medical Student Performance Evaluation (MSPE), or “dean’s letters”, and—for those applying to emergency medicine (EM) residency—the Standardized Letter of Evaluation (SLOE).<sup>3</sup> As EM residency program directors have consistently ranked SLOEs and EM rotation grades as some of the most important criteria when offering interviews and ranking applicants,<sup>4</sup> ensuring fairness and accuracy in these assessments plays a paramount role in achieving equity among EM applicants.

In an effort to improve the fairness and accuracy of EM medical student assessments (and thus that of SLOEs and EM rotation grades), the National Clinical Assessment Tool for Emergency Medicine (NCAT-EM) was developed in 2016 via consensus as a standardized assessment tool.<sup>5</sup> This tool allowed for post-shift assessment of students by faculty or residents across six clinical performance domains: history and physical exam skills; prioritized differential diagnosis; ability to formulate a plan; observation, monitoring, and follow-up; emergency recognition and management; and patient- and team-centered communication. This tool has begun to replace institution-specific tools in numerous EM rotations across the United States,<sup>2,5,6</sup> and was the first nationally standardized, specialty-specific, entrustable professional activities-based assessment tool for medical students.<sup>5</sup>

Prior work evaluating early implementation of the NCAT-EM suggests it is achieving some of its stated goals. Specifically, Hiller et al noted high internal consistency in scores within a given institution,<sup>6</sup> suggesting that this tool supports reliable comparison of students within an institution during the residency application process. However, this work showcased some gaps in the validity evidence, specifically noting site-specific variation in ratings, perhaps suggesting limitations in the response process of the validity (how the raters differentially use the tool) or the generalizability of the tool.<sup>1</sup> Additionally, this prior work was predominantly limited to medical students in their fourth year (with a high percentage interested in EM residency), limiting generalizability. Additionally, although they found site-specific differences based on student and rater gender, they were unable to examine the association of race or ethnicity on NCAT-EM scores (despite known racial disparities in other standardized EM assessments,<sup>7,8</sup> such as the SLOE<sup>9–11</sup>). Additional work to investigate the various domains of validity of this tool would add to the literature.

Unfortunately, such disparities in assessment have been observed throughout medical education,<sup>12</sup> potentially contributing to the known leadership disparities and pay-based disparities based on gender and race.<sup>13–18</sup> Studies have identified gender- and race-based differences in language used

### *Population Health Research Capsule*

What do we already know about this issue?  
*Clinical assessment tools in medical education often show racial and sex disparities in scoring and narrative feedback.*

What was the research question?  
*Do student and rater demographics affect National Clinical Assessment Tool for Emergency Medicine (NCAT-EM) scores in emergency medicine clerkships?*

What was the major finding of the study?  
*Under-represented in medicine (URM) students had lower global scores (OR 0.50, 95% CI 0.25–0.99,  $P = .048$ ), which was mitigated by URM rater concordance.*

How does this improve population health?  
*Identifying disparities in clerkship assessments supports equitable evaluation, critical for building a diverse physician workforce.*

in the MSPE<sup>19,20</sup> and language used in clerkship evaluations.<sup>7,19,21</sup> Gender- and race-based differences have also been observed through clinical grades (with lower clinical grades for non-White students, even after adjusting for variables such as scores on Step 1 of the US Medical Licensing Exam)<sup>7,22</sup> and overall recommendations on SLOEs.<sup>9–11</sup> These differences likely represent inequities, especially when the observed variations are not explained by the student performance but instead by other factors such as the clinical learning environment or the evaluator.<sup>23,24</sup> It is not known whether such differences persist despite the implementation of a nationally standardized tool such as the NCAT-EM.

To address this gap, our goal in this study was to analyze the association between student and rater demographics and NCAT-EM scores of clerkship students rotating through various emergency department sites at a single institution. Given prior evidence suggesting that concordance in demographics may impact ultimate evaluation,<sup>25,26</sup> we similarly assessed how concordance in student and rater demographics was associated with NCAT-EM scores.

## METHODS

### Setting and Participants

We performed a retrospective, single-center cohort study

of all electronically completed NCAT-EM assessments of clerkship medical students at the University of Pennsylvania. Individualized NCAT-EM forms were made available electronically through a Qualtrics (Qualtrics International Inc, Provo, UT) QR code. The dataset included all submitted NCAT-EM assessments from January–December 2022, as assessments completed during this year were unaffected by the COVID-19 pandemic. Given our interest in the impact of demographics on NCAT-EM scores, we excluded assessments for which demographic information (for either the student or the rater) was unavailable.

The Perelman School of Medicine curriculum includes 1.5 years of pre-clerkship content, followed by a one-year clerkship year, consisting of eight core clerkships (emergency medicine, family medicine, internal medicine, neurology, obstetrics and gynecology, pediatrics, psychiatry, and surgery) graded on an honors/high pass/pass/fail basis. During the clerkship year, students additionally complete an additional month of otolaryngology, orthopedic surgery, anesthesia, and ophthalmology (one week each, graded on a pass/fail basis). The EM clerkship is a four-week core clerkship completed at either one or two of six affiliated clinical sites.

The NCAT-EM has been used for clinical assessment of EM clerkship students in this institution since 2018. The NCAT-EM consists of six clinical performance domains rated on a four-point entrustability scale, a global assessment domain, a professionalism section, and mandatory free-text comments for strengths and suggestions for improvement (Supplemental Table 1). Students are required to present a QR code linking the NCAT-EM to attendings or EM residents during every shift, which is completed at the rater's convenience on an online platform.

During the EM clerkship, students were assigned to work 14 eight-hour shifts (or total hourly equivalent for sites with 10- or 12-hour shifts) for the duration of the clerkship. Students were required to present the QR code to at least one rater per shift (either an attending or a supervising resident ranging from postgraduate year [PGY] 2-4). The dataset did not include discrete PGY-level data for resident raters. Prior to and during this study, the clerkship directors performed annual education focused on the NCAT-EM tool and the process of assessment, which consisted of an introduction to the tool, a brief overview of the scale, and a review of medical student evaluation processes. This information session included both faculty and residents at all sites (for all individuals who would be working with students).

### Data Collection and Analysis

Data collected in addition to completed NCAT-EM forms included student factors (gender and under-represented in medicine [URM] status) and rater factors (gender, URM status, and role, either resident or faculty). We extracted gender and race data for students from admissions demographics based on self-identification. For faculty,

self-reported gender and race were obtained through the university's faculty affairs database. We defined URM status for both faculty and students using the Association of American Medical Colleges (AAMC) definitions. The URM status was specifically chosen as a binary variable (as opposed to race and ethnicity data), to improve power in our statistical analysis. Importantly, the definition of URM can broadly include groups that are minoritized, such as first-generation, low-income students, or students with disabilities, although for this work we used the AAMC definition of URM based on race and ethnicity. The dataset also included the quarter of year in which the student was completing the clerkship (block 1, 2, 3, or 4), and the clinical site where they were rotating. All data were deidentified prior to data analysis.

We performed univariate ordinal logistic regression analyses to determine the association between the global assessment on the NCAT-EM tool (bottom third, middle third, top third, or top 10%), with student gender, student URM status, faculty gender, faculty URM status, clerkship site, and rotation block. We then performed mixed-effect ordinal logistic regression analyses to determine the association between NCAT-EM scores and student gender, clustered by student, after adjusting for site of rotation, time, role of rater, student URM status, and rater demographics (gender, URM status). To assess the association with URM status, we performed mixed-effect ordinal logistic regression analyses to determine the association between NCAT-EM scores and student URM status, clustered by student, after adjusting for site of rotation, block, role of rater, student gender, and rater demographics. Given the hypothesis that concordance in rater gender and student gender and URM status might be associated with NCAT-EM scores, we also assessed the interaction between student gender and rater gender, and student URM status and rater URM status.

For each analysis, the model was clustered on student (random effects) and rater (random effects). This model was used intentionally to adjust for the non-independent nature of students and/or raters throughout the dataset, as this model provides adjusted standard errors accounting for student and/or rater clustering (random effects) throughout the dataset. Based on prior factor analyses showing each domain in the NCAT-EM assessed unique domains, we repeated the above analysis for each of the six clinical performance domains of the assessment. (See Supplemental Table 1 for the NCAT-EM domains.)

While our primary analysis included URM (as per AAMC definition) as a binary variable, we aimed to further understand our findings in the context of URM categories, recognizing that URM individuals who spanned different identities might have had different experiences with assessment. Thus, we performed a sensitivity analysis using racial and ethnic groups within the AAMC definition of URM (African-American/Black, Hispanic/Latino, Native American, including American Indian, Alaska Native, and Native

Hawaiian, Pacific Islander, and mainland Puerto Rican). We then performed a second sensitivity analysis comparing individuals identifying as Black compared to other individuals (noting the large proportion of individuals within the cohort identifying as URM were Black and the distinct experiences of this population<sup>27,28</sup>).

We completed statistical analysis using STATA v18.0 (StataCorp, LLC, College Station, TX). Statistical significance was determined using a *P*-value of .05 (not adjusting for multiple comparisons given the exploratory nature of the analysis, to reduce the risk of type 2 error). This study was deemed exempt by the University of Pennsylvania Institutional Review Board.

## RESULTS

Over the course of 2022, 1,881 complete NCAT-EM assessment forms were submitted on 142 distinct students (consisting of 74 women [52%] and 68 men [48%], including 34 [24%] who identified as URM) completed by 266 different raters. We excluded 122 NCAT-EM forms prior to analysis, due to incomplete demographic information for the rater. The median number of completed forms per student was 13 [IQR 11–15], which was similar between genders (13.4 for men vs 13.1 for women, *P* = .59). There were fewer NCAT-EM assessments completed on those who identified as URM within the sample (with a mean of 11.8 vs 13.8 assessments for non-URM students, *P* = .01). Most assessments were completed by raters who identified as men (*n* = 1,070, 60%), and 11% (*n* = 195) were completed by raters who identified as URM. The racial demographics of faculty identified as URM (as per AAMC definitions) was 119 (61%) Black, 55 (28%) Hispanic or Latino, and 21 (11%) Pacific Islander. The racial demographics of students identified as URM (as per AAMC definitions) was 224 (56%) Black, 38 (10%) Hispanic or Latino, and 138 (35%) spanning multiple groups. Complete demographic information of completed NCAT-EM forms are included in Table 1.

Distribution of scores for each of the six clinical performance domains on the NCAT-EM as well as the global assessment domain (see Supplemental Table 1) are summarized in Table 2. Global assessment scores were skewed leftward (consistent with prior national data<sup>5</sup>), with 38 ratings (2.1) ratings representing the lower third in the global assessment, 506 (28%) in the middle third, 878 (49%) in the upper third, and 387 (21%) in the top 10% (“exceptional”).

The results of the univariate ordinal logistic regression are shown in Table 3. In the univariate analysis, there was a significant association based on rater role, with faculty raters being less likely to rate students in the higher entrustment scores compared to resident raters for all domains (*P* < .001 for all domains); thus, this was included in the multivariate analysis. There was also an association between rotation block and NCAT-EM scores, and site of the rotation and NCAT-EM scores; thus, these variables were included in the final

**Table 1.** Student characteristics, rater characteristics, clinical site and block of completed NCAT-EM assessment forms.

| Characteristic           | Completed NCAT-EM forms (n, %) |
|--------------------------|--------------------------------|
| Student gender           | 1881 (100.0%)                  |
| Men                      | 913 (48.5%)                    |
| Women                    | 968 (51.5%)                    |
| Student URM status       | 1881 (100.0%)                  |
| URM                      | 400 (21.3%)                    |
| African American/Black   | 224 (56%)                      |
| Hispanic/Latino          | 38 (10%)                       |
| Multiple AAMC URM Groups | 138 (35%)                      |
| Non-URM                  | 1481 (78.7%)                   |
| Rater Role               | 1775 (100.0%)                  |
| Resident                 | 484 (27.3%)                    |
| Faculty                  | 1291 (72.7%)                   |
| Rater Gender             | 1783 (100.0%)                  |
| Men                      | 1070 (60.0%)                   |
| Women                    | 713 (40.0%)                    |
| Rater URM Status         | 1759 (100.0%)                  |
| URM                      | 195 (11.1%)                    |
| African American/Black   | 119 (61%)                      |
| Hispanic/Latino          | 55 (28%)                       |
| Pacific Islanders        | 21 (11%)                       |
| Non-URM                  | 1564 (88.9%)                   |
| Clinical Site            | 1880 (100.0%)                  |
| Quaternary care site     | 694 (36.9%)                    |
| Community site A         | 662 (35.2%)                    |
| Community site B         | 187 (9.9%)                     |
| Community site C         | 129 (6.9%)                     |
| Pediatric site           | 121 (6.4%)                     |
| Veteran's Affairs        | 87 (4.6%)                      |
| Block                    | 1881 (100.0%)                  |
| 1                        | 569 (30.2%)                    |
| 2                        | 569 (30.2%)                    |
| 3                        | 305 (16.2%)                    |
| 4                        | 438 (23.3%)                    |

All results are expressed as number of NCAT assessment forms completed within each category, followed by percent of assessments for which data are available.

NCAT-EM, National Clinical Assessment Tool for Emergency Medicine; URM, under-represented in medicine; AAMC, Association of American Medical Colleges.

regression model.

The results of the multivariate ordinal logistic regression, clustered by student and rater, are shown in Table 4, using a significant threshold of *P* = .05 (rather than adjusting for

**Table 2.** Overview of ratings for National Clinical Assessment Tool for Emergency Medicine clinical performance domains and global assessment.

| NCAT-EM Domain                                                   | n (%)        |
|------------------------------------------------------------------|--------------|
| Focused history and physical exam skills                         | 1,879 (100%) |
| Unable to assess                                                 | 25 (1%)      |
| Pre-entrustable                                                  | 118 (6%)     |
| Mostly entrustable                                               | 742 (40%)    |
| Fully entrustable/Milestone 1                                    | 754 (40%)    |
| Outstanding/Milestone 2                                          | 240 (13%)    |
| Ability to generate a prioritized differential diagnosis         | 1,866 (100%) |
| Unable to assess                                                 | 31 (2%)      |
| Pre-entrustable                                                  | 152 (8%)     |
| Mostly entrustable                                               | 818 (44%)    |
| Fully entrustable/Milestone 1                                    | 664 (36%)    |
| Outstanding/Milestone 2                                          | 201 (11%)    |
| Ability to formulate plan (diagnostic, therapeutic, disposition) | 1,861 (100%) |
| Unable to assess                                                 | 34 (2%)      |
| Pre-entrustable                                                  | 163 (9%)     |
| Mostly entrustable                                               | 894 (48%)    |
| Fully entrustable/Milestone 1                                    | 592 (32%)    |
| Outstanding/Milestone 2                                          | 178 (10%)    |
| Observation, monitoring, and follow-up                           | 1,861 (100%) |
| Unable to assess                                                 | 42 (2%)      |
| Pre-entrustable                                                  | 120 (6%)     |
| Mostly entrustable                                               | 683 (37%)    |
| Fully entrustable/Milestone 1                                    | 747 (40%)    |
| Outstanding/Milestone 2                                          | 263 (14%)    |
| Emergency recognition and management                             | 1,851 (100%) |
| Unable to assess                                                 | 385 (21%)    |
| Pre-entrustable                                                  | 98 (5%)      |
| Mostly entrustable                                               | 673 (36%)    |
| Fully entrustable/Milestone 1                                    | 523 (28%)    |
| Outstanding/Milestone 2                                          | 172 (9%)     |
| Patient- and team-centered communication                         | 1,848 (100%) |
| Unable to assess                                                 | 40 (2%)      |
| Pre-entrustable                                                  | 65 (4%)      |
| Mostly entrustable                                               | 605 (33%)    |
| Fully entrustable/Milestone 1                                    | 794 (43%)    |
| Outstanding/Milestone 2                                          | 344 (19%)    |
| Global assessment                                                | 1,809 (100%) |
| Lower third                                                      | 38 (2%)      |
| Middle third                                                     | 506 (28%)    |
| Top third                                                        | 878 (49%)    |
| Exceptional (top 10%)                                            | 387 (21%)    |

All results are expressed as number of NCAT-EM assessment forms completed within each category, followed by percentage of assessments for which data are available.

NCAT-EM, National Clinical Assessment Tool for Emergency Medicine.

multiple comparison due to the exploratory nature of the study). As there was no significant interaction between the student gender and rater gender in each analysis, this interaction term was excluded from the final regression model. The final regression model included the rater gender and student gender, rater and student URM status (and the interaction between them), rater role (faculty vs resident), clinical site, and rotation block. The results of the mixed regression identified no significant associations between student gender and NCAT-EM scores for each NCAT-EM domain (Table 4). There was a significant interaction effect between student gender and rater gender in the domain of history and physical exam (OR 0.31, CI 0.11-0.83,  $P = .02$ ). There were no other significant interaction effects in the remainder of the domains.

Student URM status was associated with lower scores for the global assessment, (OR 0.50, CI 0.25-0.99,  $P = .05$ ), history/physical exam domain (OR 0.38, CI 0.19-0.77,  $P = .01$ ) and the prioritized differential diagnosis domain (OR 0.47, CI 0.26-0.88,  $P = .02$ ) after multivariate adjustment, as shown in Table 4. These findings were moderated by a significant positive interaction effect between student and rater URM status in the observation/monitoring domain (OR 4.55, CI 1.21-17.1,  $P = .03$ ), suggesting that concordance in URM status between raters and students lessened (and in some instances, reversed) the negative effect of URM status on NCAT-EM scores. More specifically, when assessing the significant domains for a given student, the adjusted ORs for URM-concordant dyads were adjusted OR 0.83 for the global assessment (suggesting the difference persisted); adjusted OR 1.06 for history/physical domains; and adjusted OR 1.76 for observation/monitoring (suggesting reversal of the score, and URM-concordant dyads had higher odds of receiving a higher score than the reference cohort).

The sensitivity analysis using racial and ethnic groups within the URM status (Black, Hispanic/Latino, Native Americans, including American Indians, Alaska Natives, and Native Hawaiians, Pacific Islanders, and mainland Puerto Ricans demonstrated a poorer fit to the data—based on likelihood ratio tests and Bayesian information criterion and Akaike's information criterion comparisons—than the original model. As this approach risked underestimating the true effect, it was not included in the results. In our sensitivity analysis comparing individuals who were Black to other individuals in the cohort, we found significant associations with NCAT-EM ratings in the domains of history/physical domain (OR 0.32, CI 0.14-0.76,  $P = .01$ ); prioritized differential domain (OR 0.36, CI 0.17-0.79,  $P = .01$ ); the ability to formulate a plan domain (OR 0.46, CI 0.59-1.48,  $P = .05$ ), with significant interaction effects noted in the majority of domains. (See Table 5 for full details.)

## DISCUSSION

Our study demonstrates important associations between both rater and student demographics and NCAT-EM scores

**Table 3.** Univariate associations between ratings for National Clinical Assessment Tool for Emergency Medicine domains and student and rater characteristics.

| Variable                  | Global Assessment       |        | History/Physical        |        | Prioritized Differential |        | Ability to Formulate Plan |        | Observation/Monitoring  |        | Emergency Management    |        | Communication           |        |
|---------------------------|-------------------------|--------|-------------------------|--------|--------------------------|--------|---------------------------|--------|-------------------------|--------|-------------------------|--------|-------------------------|--------|
|                           | OR<br>(95%<br>CI)       | P      | OR<br>(95%<br>CI)       | P      | OR<br>(95%<br>CI)        | P      | OR<br>(95%<br>CI)         | P      | OR<br>(95%<br>CI)       | P      | OR<br>(95%<br>CI)       | P      | OR<br>(95%<br>CI)       | P      |
| <b>Student gender</b>     |                         |        |                         |        |                          |        |                           |        |                         |        |                         |        |                         |        |
| Men                       | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)             | -      | 1.0<br>(ref)              | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      |
| Women                     | 0.91<br>(0.77,<br>1.08) | .30    | 0.85<br>(0.72,<br>1.01) | .06    | 0.91<br>(0.77,<br>1.08)  | .30    | 0.90<br>(0.76,<br>1.07)   | .24    | 0.97<br>(0.83,<br>1.15) | .75    | 1.05<br>(0.88,<br>1.26) | .59    | 0.97<br>(0.82,<br>1.15) | .73    |
| <b>Student URM-status</b> |                         |        |                         |        |                          |        |                           |        |                         |        |                         |        |                         |        |
| Non-URM                   | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)             | -      | 1.0<br>(ref)              | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      |
| URM                       | 0.91<br>(0.73,<br>1.12) | .37    | 0.90<br>(0.73,<br>1.11) | .23    | 0.92<br>(0.75,<br>1.14)  | .45    | 0.67<br>(0.37-<br>1.22)   | .19    | 1.00<br>(0.81,<br>1.23) | .99    | .99<br>(0.81,<br>1.21)  | .93    | 1.11<br>(0.90,<br>1.36) | .33    |
| <b>Rater role</b>         |                         |        |                         |        |                          |        |                           |        |                         |        |                         |        |                         |        |
| Resident                  | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)             | -      | 1.0<br>(ref)              | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      |
| Faculty                   | 0.46<br>(0.38,<br>0.57) | < .001 | 0.34<br>(0.28,<br>0.42) | < .001 | 0.29<br>(0.24,<br>0.36)  | < .001 | 0.38<br>(0.31,<br>0.46)   | < .001 | 0.32<br>(0.27,<br>0.40) | < .001 | 0.43<br>(0.35,<br>0.53) | < .001 | 0.19<br>(0.15,<br>0.24) | < .001 |
| <b>Rater gender</b>       |                         |        |                         |        |                          |        |                           |        |                         |        |                         |        |                         |        |
| Men                       | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)             | -      | 1.0<br>(ref)              | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      |
| Women                     | 1.05<br>(0.88,<br>1.26) | .60    | 1.26<br>(1.06,<br>1.51) | .01    | 1.21<br>(1.01,<br>1.44)  | .04    | 1.21<br>(1.02,<br>1.45)   | .03    | 1.24<br>(1.04,<br>1.48) | 0.02   | 0.86<br>(0.72,<br>1.02) | 0.09   | 1.04<br>(0.87,<br>1.24) | .65    |
| <b>Rater URM-status</b>   |                         |        |                         |        |                          |        |                           |        |                         |        |                         |        |                         |        |
| Non-URM                   | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)             | -      | 1.0<br>(ref)              | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      | 1.0<br>(ref)            | -      |
| URM                       | 0.67<br>(0.53,<br>0.93) | .02    | 1.14<br>(0.87,<br>1.50) | .33    | 1.23<br>(0.93,<br>1.61)  | .13    | 1.29<br>(0.97,<br>1.69)   | .08    | 1.37<br>(1.04,<br>1.79) | .02    | 0.95<br>(0.72,<br>1.26) | .74    | 1.08<br>(0.82,<br>1.43) | .58    |

OR, odds ratio; NCAT-EM, National Clinical Assessment Tool for Emergency Medicine; URM, under-represented in medicine.

within our cohort, with notable findings based on student URM status. Our multivariate analysis did not find any gender-related differences in NCAT-EM domains. However, the multivariate analysis showed that students identified as URM received lower NCAT-EM scores in several domains, including the history/physical exam domain and the prioritized differential diagnosis domain. This effect was mediated (and in some cases reversed) by concordance of URM status between raters and students in some of the domains, such that concordance in URM status between students and raters was associated with higher NCAT-EM scores.

The association between student URM status with lower NCAT-EM scores is consistent with prior literature

documenting longstanding racial disparities in clerkship grading.<sup>21,29,30</sup> Despite the NCAT-EM being noted to have excellent internal consistency based on prior studies,<sup>6</sup> this suggests that the use of the tool continues to be impacted by its differential use by raters and raises some concern about additional domains of validity with the tool.<sup>1</sup> It is important for us as medical educators to ensure the assessment tool widely used to guide clerkship grading does not introduce any construct irrelevance variance at all.<sup>31,32</sup>

In this study, the observed score differences by URM status may be the result of implicit bias of raters affecting both their global perception and perception of competency-related behaviors of students, lack of mentorship leading to

**Table 4.** Multivariate associations between ratings for National Clinical Assessment Tool for Emergency Medicine domains and student sex and under-represented in medicine (URM) status, after adjusting for rater sex and URM status, concordance of student-rater URM status, clinical site, and time.

| Variable                                                | Global assessment        |      | History/physical         |      | Prioritized differential |     | Ability to formulate plan |      | Observation/monitoring   |     | Emergency management    |     | Communication            |     |
|---------------------------------------------------------|--------------------------|------|--------------------------|------|--------------------------|-----|---------------------------|------|--------------------------|-----|-------------------------|-----|--------------------------|-----|
|                                                         | OR<br>(95%<br>CI)        | P    | OR<br>(95%<br>CI)        | P    | OR<br>(95%<br>CI)        | P   | OR<br>(95%<br>CI)         | P    | OR<br>(95%<br>CI)        | P   | OR<br>(95%<br>CI)       | P   | OR<br>(95%<br>CI)        | P   |
| <b>Student Sex</b>                                      |                          |      |                          |      |                          |     |                           |      |                          |     |                         |     |                          |     |
| Men                                                     | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -   | 1.0<br>(ref)              | -    | 1.0<br>(ref)             | -   | 1.0<br>(ref)            | -   | 1.0<br>(ref)             | -   |
| Women                                                   | 0.93<br>(0.54,<br>1.60)  | .78  | 0.84<br>(0.50,<br>1.42)  | .52  | 0.98<br>(0.61,<br>1.57)  | .93 | 0.95<br>(0.60,<br>1.51)   | .84  | 1.26<br>(0.84,<br>1.89)  | .26 | 0.87<br>(0.60,<br>1.27) | .48 | 1.21<br>(0.82,<br>1.78)  | .35 |
| <b>Interaction between rater and student gender</b>     |                          |      |                          |      |                          |     |                           |      |                          |     |                         |     |                          |     |
| Sex concordance between men                             | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -   | 1.0<br>(ref)              | -    | 1.0<br>(ref)             | -   | 1.0<br>(ref)            | -   | 1.0<br>(ref)             | -   |
| Sex concordance between women                           | 0.74<br>(0.49,<br>1.09)  | .13  | 0.59<br>(0.40,<br>0.85)  | .006 | 0.87<br>(0.60,<br>1.27)  | .48 | 0.76<br>(0.52,<br>1.11)   | 0.16 | 0.96<br>(0.66,<br>1.40)  | .83 | 0.90<br>(0.62,<br>1.29) | .56 | 1.02<br>(0.70,<br>1.28)  | .93 |
| <b>Student URM Status</b>                               |                          |      |                          |      |                          |     |                           |      |                          |     |                         |     |                          |     |
| Non-URM                                                 | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -   | 1.0<br>(ref)              | -    | 1.0<br>(ref)             | -   | 1.0<br>(ref)            | -   | 1.0<br>(ref)             | -   |
| URM                                                     | 0.50<br>(0.25,<br>0.99)  | .006 | 0.38<br>(0.19,<br>0.77)  | .007 | 0.47<br>(0.26,<br>0.88)  | .02 | 0.67<br>(0.37,<br>1.22)   | .19  | 0.62<br>(0.37,<br>1.06)  | .08 | 0.78<br>(0.60,<br>1.27) | .34 | 0.78<br>(0.47,<br>1.30)  | .35 |
| <b>Interaction between rater and student URM-status</b> |                          |      |                          |      |                          |     |                           |      |                          |     |                         |     |                          |     |
| Non-URM concordance                                     | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -   | 1.0<br>(ref)              | -    | 1.0<br>(ref)             | -   | 1.0<br>(ref)            | -   | 1.0<br>(ref)             | -   |
| URM concordance                                         | 2.92<br>(0.70,<br>12.23) | .14  | 3.97<br>(0.74,<br>21.30) | .11  | 4.38<br>(0.99,<br>19.35) | .05 | 3.47<br>(0.72,<br>16.64)  | .12  | 4.56<br>(1.21,<br>17.09) | .03 | 1.28<br>(0.33,<br>4.93) | .72 | 2.60<br>(0.65,<br>10.41) | .18 |

OR, odds ratio; NCAT-EM, National Clinical Assessment Tool for Emergency Medicine; URM, under-represented in medicine.

**Table 5.** Multivariate associations between ratings for National Clinical Assessment Tool for Emergency Medicine domains and Black vs non-Black race, after adjusting for rater sex and race (Black vs non-Black), concordance of student-rater race, clinical site, and time.

| Variable                                            | Global assessment        |      | History/physical         |      | Prioritized differential |      | Ability to formulate plan |      | Observation/monitoring     |       | Emergency management    |     | Communication              |      |
|-----------------------------------------------------|--------------------------|------|--------------------------|------|--------------------------|------|---------------------------|------|----------------------------|-------|-------------------------|-----|----------------------------|------|
|                                                     | OR<br>(95%<br>CI)        | P    | OR<br>(95%<br>CI)        | P    | OR<br>(95%<br>CI)        | P    | OR<br>(95%<br>CI)         | P    | OR<br>(95%<br>CI)          | P     | OR<br>(95%<br>CI)       | P   | OR<br>(95%<br>CI)          | P    |
| <b>Student Race</b>                                 |                          |      |                          |      |                          |      |                           |      |                            |       |                         |     |                            |      |
| Non-Black                                           | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -    | 1.0<br>(ref)             | -    | 1.0<br>(ref)              | -    | 1.0 (ref)                  | -     | 1.0<br>(ref)            | -   | 1.0 (ref)                  | -    |
| Black                                               | 0.67<br>(0.28,<br>1.62)  | .38  | 0.32<br>(0.14,<br>0.76)  | .01  | 0.36<br>(0.17,<br>0.79)  | .01  | 0.46<br>(0.59,<br>1.48)   | .05  | 0.54<br>(0.28,<br>1.04)    | .07   | 0.78<br>(0.42,<br>1.44) | .42 | 0.65<br>(0.34,<br>1.24)    | .20  |
| Interaction between Black Raters and Black Students | 25.8<br>(3.23,<br>205.1) | .002 | 39.4<br>(3.42,<br>452.5) | .003 | 38.9<br>(4.56,<br>331.4) | .001 | 32.9<br>(3.40,<br>317.43) | .003 | 29.37<br>(4.55,<br>189.46) | <.001 | 2.07<br>(0.31,<br>13.9) | .45 | 17.59<br>(2.29,<br>135.41) | .006 |

OR, odds ratio; NCAT-EM, National Clinical Assessment Tool for Emergency Medicine.

inequitable opportunities, or different lived experiences of URM students impacting their experience, and performance in the clinical environment (including stereotype threat, microaggressions, patient mistreatment, being tasked with being a racial ambassador, unrewarded labor, limited resources, and othering).<sup>17,21,33–35</sup> The complexity of this amalgam of factors that exacerbates disparities in clerkship grading has been described as the “social milieu of medical education,”<sup>35</sup> and may ultimately contribute to the inequities observed in other standardized assessments used in EM, such as the SLOE. Improved understanding of these disparities in the EM clerkship setting and further evaluation of the validity evidence of the NCAT-EM tool is critical to identify solutions to mitigate these issues.

Perhaps more interestingly, some of the findings of a differential score based on URM status were mitigated (and in some cases, reversed) by URM concordance between the student and rater, specifically within the prioritization of a differential, and the observation and monitoring domains. As a possible explanation, URM concordance may reduce implicit bias of the rater as well as the other effects of racism on the medical student, such as stereotype threat. Concordance between the rater and student can also enhance performance of the student through the role-model effect.<sup>33</sup> This aligns with prior studies that have shown the importance of racial concordance in multiple domains, including patient care, professional development,<sup>17</sup> and medical education assessment,<sup>36,37</sup> further highlighting the critical nature of supporting equity initiatives to advanced diversity across EM faculty and residents.

It is not clear why this phenomenon would be present for only two of the NCAT-EM domains, although it could represent something unique about those domains, including that they may capture more direct interaction between students and raters (such as prioritizing differentials) and, thus, concordance would be more heavily impacted. Regardless, this further suggests that additional robust validity studies of the NCAT-EM tool are needed. Additionally, while we noted various impacts of URM concordance on the NCAT-EM scores, the impact on the overall disparities identified in our study may be negligible, and further work is needed to examine this phenomenon across a larger sample of more diverse learners and raters.

Another interesting observation within our study was that URM students had fewer submitted NCAT-EM forms in the full cohort, which persisted after adjusting for minor site differences. The structure of the NCAT-EM within our institution requires learners to seek out designated feedback and collect assessments via a QR code. This difference could indicate differences in self-promotion behavior,<sup>38</sup> which may uniquely disadvantage URM students. Specifically, there is a complex interplay between evaluations, biases, and the associated impact on confidence, self-esteem, and motivation. In URM students, negative evaluations, even if biased, may

reinforce stereotype threat—defined as a fear of confirming negative stereotypes about their group<sup>17</sup> and ultimately hinder professional growth. Understanding this, as well as other unique barriers to seeking evaluation by URM students as observed in this study, should be further evaluated.

We also found no difference in NCAT-EM scores based on gender, with no significant difference in scores between men and women in the cohort. This absence of gender associations across the NCAT-EM performance domains was surprising, and in contrast to prior work analyzing the NCAT-EM in medical students. Specifically, in a study by Hiller and colleagues, there were student gender-based differences in composite NCAT-EM scores at 4 of the 13 sites included in their study.<sup>6</sup> However, this prior work was conducted with limited demographic data and an overrepresentation of male students, with predominantly students in their final year of medical school. It is possible that gender disparities across diverse assessment domains become apparent at later stages in training (as has been observed in residency assessments).<sup>39,40</sup>

As NCAT-EM scores inform clinical grades, and subsequently the SLOE and MSPE, it is critical to mitigate disparities in use of the NCAT-EM tool as found in our study. The NCAT-EM has features of prior recommendations to reduce grading inequity, including workplace-based assessment, criterion-based rubrics, and competency-based grading. Our data show that rating disparities are still present, despite high internal consistency metrics of the NCAT-EM. This argues that the tool itself does not contribute to disparities, but real-world use of the tool by raters contributes to these differences. Ultimately, the scoring differences found in our study support the use of rater training, which has been shown to improve the accuracy of workplace-based entrustment ratings of medical learners.<sup>41</sup>

Additionally, ongoing efforts to promote an equitable and diverse workforce are necessary, noting the role of concordance on some of these disparities. Ultimately, achieving fairness and accuracy in NCAT-EM assessments is crucial to promoting gender and racial equity among EM applicants, especially with national implementation of the NCAT-EM tool. In addition, clerkships in other specialties should note that despite the positive impacts of using a standardized and national assessment instrument with strong reliability, testing by itself is not the solution for overcoming observed differences not explained by student performance.

## LIMITATIONS

Although we identified compelling findings, this study had several limitations. Our study analytic approach involved multiple regressions, and we did not adjust for multiple comparisons due to the exploratory nature of the study, which increased the risk of type 1 error in our conclusions. However, we feel the presence of findings in our sensitivity analysis suggests that this finding is a true trend. In addition, the study

was limited to a single institution. Although it included six different sites within the institution (each with a unique culture and patient population), obtaining a multicenter study across distinct geographical regions is an important next step to fully evaluate the effect of student and rater demographics on NCAT-EM performance nationally. Furthermore, it is not yet clear whether these findings among second- and third-year clerkship students can be generalized to more senior medical students on sub-internships or electives, and additional work evaluating the impact of URM status on advanced students is needed. Our study did not include individuals who identified as gender diverse, which would also be important to include in future research. Finally, while we importantly noted some differences in NCAT-EM use by site and by level (residents NCAT-EM scores were higher as compared to faculty), we were unable to assess PGY level in residency training nor the impact of faculty development on the tool. This is an important area of future work.

## CONCLUSION

While we found no association between student gender with NCAT-EM scores, we did find an association between student under-represented in medicine status in two of six NCAT-EM performance domains, an effect that was mediated by concordance in URM status with the rater. Future multi-institution research is needed to verify grading disparities based on student and rater characteristics on the national level, which would further support the use of multifaceted interventions to mitigate disparities in ratings, including diversity efforts in recruitment practices, equitable access to medical school resources, gender- or URM-specific student support and rater training to ultimately promote equity among emergency physicians.

**Address for Correspondence:** Janae K. Heath, MD, MSCE, Hospital of the University of Pennsylvania, Department of Medicine, 3600 Spruce St, 822 West Gates Building, Philadelphia, PA 19104. Email: Janae.heath@penmedicine.upenn.edu

**Conflicts of Interest:** By the *WestJEM* article submission agreement, all authors are required to disclose all affiliations, funding sources and financial or management relationships that could be perceived as potential sources of bias. No author has professional or financial relationships with any companies that are relevant to this study. There are no conflicts of interest or sources of funding to declare.

**Copyright:** © 2025 Amin et al. This is an open access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) License. See: <http://creativecommons.org/licenses/by/4.0/>

## REFERENCES

1. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med.* 2006;119(2):166.e7-16.
2. Lawson L, Jung J, Franzen D, et al. Clinical assessment of medical students in emergency medicine clerkships: a survey of current practice. *J Emerg Med.* 2016;51(6):705-11.
3. Negaard M, Assimacopoulos E, Harland K, et al. Emergency medicine residency selection criteria: an update and comparison. *AEM Educ Train.* 2018;2(2):146-153.
4. Katzung KG, Ankel F, Clark M, et al. What do program directors look for in an applicant? *J Emerg Med.* 2019;56(5):e95-101.
5. Jung J, Franzen D, Lawson L, et al. The National Clinical Assessment Tool for Medical Students in the emergency department (NCAT-EM). *West J Emerg Med.* 2018;19(1):66-74.
6. Hiller K, Jung J, Lawson L, et al. Multi-institutional implementation of the national clinical assessment tool in emergency medicine: data from the first year of use. *AEM Educ Train.* 2021;5(2):e10496.
7. Gauer JL, Mustapha T, Violato C. Race and gender bias in clerkship grading. *Teach Learn Med.* 2024;36(3):304-11.
8. Nguemeni Tiako MJ, Ray V, South EC. Medical schools as racialized organizations: how race-neutral structures sustain racial inequality in medical education—a narrative review. *J Gen Intern Med.* 2022;37(9):2259-66.
9. Winfield A, Amin DP. To the Editor: Where is equity in the SLOE? *J Grad Med Educ.* 2022;14(3):357.
10. Calles I. To the Editor: For equity in assessment: a comment on bias in the emergency medicine Standardized Letter of Evaluation. *J Grad Med Educ.* 2023;15(1):129.
11. Kukulski P, Ahn J. Validity evidence for the emergency medicine Standardized Letter of Evaluation. *J Grad Med Educ.* 2021;13(4):490-9.
12. Colson ER, Pérez M, Blaylock L, 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 Dec;95(12S Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments):S131-5.
13. Cheng D, Promes S, Clem K, et al. Chairperson and faculty gender in academic emergency medicine departments. *Acad Emerg Med.* 2006;13(8):904-6.
14. Madsen TE, Linden JA, Rounds K, et al. Current status of gender and racial/ethnic disparities among academic emergency medicine physicians. *Acad Emerg Med.* 2017;24(10):1182-92.
15. Wiler JL, Wendel SK, Rounds K, et al. Salary disparities based on gender in academic emergency medicine leadership. *Acad Emerg Med.* 2022;29(3):286-93.
16. Jena AB, Khullar D, Ho O, et al. Sex differences in academic rank in US medical schools in 2014. *JAMA.* 2015;314(11):1149-58.
17. Bullock JL, Lockspeiser T, Del Pino-Jones A, et al. They don't see a lot of people my color: a mixed methods study of racial/ethnic stereotype threat among medical students on core clerkships. *Acad Med.* 2020;95(11S Association of American Medical Colleges Learn Serve Lead: Proceedings of the 59th Annual Research in Medical Education Presentations):S58-S66.
18. Ackerman-Barger K, Boatright D, Gonzalez-Colaso R, et al. Seeking

- inclusion excellence: understanding racial microaggressions as experienced by underrepresented medical and nursing students. *Acad Med*. 2020;95(5):758-63.
19. Ross DA, Boatright D, Nunez-Smith M, et al. Differences in words used to describe racial and gender groups in Medical Student Performance Evaluations. *PLoS ONE*. 2017;12(8):e0181659.
  20. Axelson RD, Solow CM, Ferguson KJ, et al. Assessing implicit gender bias in Medical Student Performance Evaluations. *Eval Health Prof*. 2010;33(3):365-85.
  21. Hanson JL, Pérez M, Mason HRC, et al. Racial/ethnic disparities in clerkship grading: perspectives of students and teachers. *Acad Med*. 2022;97(11S):S35-S45.
  22. O'Sullivan L, Kagabo W, Prasad N, et al. Racial and ethnic bias in medical school clinical grading: a review. *J Surg Educ*. 2023;80(6):806-16.
  23. Lucey CR, Hauer KE, Boatright D, et al. Medical education's wicked problem: achieving equity in assessment for medical learners. *Acad Med*. 2020;95(12S Addressing Harmful Bias and Eliminating Discrimination in Health Professions Learning Environments):S98-S108.
  24. Kakara Anderson HL, Govaerts M, et al. Clarifying and expanding equity in assessment by considering three orientations: Fairness, inclusion and justice. *Med Educ*. 2025;59(5):494-502.
  25. Takeshita J, Wang S, Loren AW, et al. Association of racial/ethnic and gender concordance between patients and physicians with patient experience ratings. *JAMA Netw Open*. 2020;3(11):e2024583.
  26. McOwen KS, Bellini LM, Guerra CE, et al. Evaluation of clinical faculty: gender and minority implications. *Acad Med*. 2007;82(10):S94.
  27. Nguemeni Tiako MJ, Wages JE 3rd, Perry SP. Black medical students' sense of belonging and confidence in scholastic abilities at historically Black vs predominantly White medical schools: a prospective study. *J Gen Intern Med*. 2023 Jan;38(1):122-4.
  28. Ghanem N, Goldberg DG, Granger E, et al. A critical qualitative study to understand current Black women medical student perspectives on anti-racist reform in US medical education. *Med Educ Online*. 2024;29(1):2393436.
  29. Boatright D, Anderson N, Kim JG, et al. Racial and ethnic differences in internal medicine residency assessments. *JAMA Netw Open*. 2022;5(12):e2247649.
  30. Low D, Pollack SW, Liao ZC, et al. Racial/ethnic disparities in clinical grading in medical school. *Teach Learn Med*. 2019;31(5):487-96.
  31. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Adv Health Sci Educ Theory Pract*. 2002;7(3):235-41.
  32. Tavakol M, Dennick R. The foundations of measurement and assessment in medical education. *Med Teach*. 2017;39(10):1010-5.
  33. Wheeler M, de Bourmont S, Paul-Emile K, et al. Physician and trainee experiences with patient bias. *JAMA Intern Med*. 2019;179(12):1678-85.
  34. Colson ER, Pérez M, Chibueze S, et al. Understanding and addressing bias in grading: progress at Washington University School of Medicine. *Acad Med*. 2023;98(8S):S64.
  35. Osseo-Asare A, Balasuriya L, Huot SJ, et al. Minority resident physicians' views on the role of race/ethnicity in their training experiences in the workplace. *JAMA Netw Open*. 2018;1(5):e182723.
  36. Heath JK, Dine CJ, LaMarra D, et al. The impact of trainee and standardized patient race and gender on internal medicine resident communication assessment scores. *J Grad Med Educ*. 2021;13(5):643-9.
  37. Berg K, Blatt B, Lopreiato J, et al. Standardized patient assessment of medical student empathy: ethnicity and gender effects in a multi-institutional study. *Acad Med*. 2015;90(1):105-11.
  38. Pololi L, Conrad P, Knight S, et al. A Study of the relational aspects of the culture of academic medicine. *Acad Med*. 2009;84(1):106.
  39. Dayal A, O'Connor DM, Qadri U, et al. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. *JAMA Intern Med*. 2017;177(5):651-7.
  40. Santen SA, Yamazaki K, Holmboe ES, et al. Comparison of male and female resident milestone assessments during emergency medicine residency training: a national study. *Acad Med*. 2020;95(2):263.
  41. Kogan JR, Dine CJ, Conforti LN, et al. Can rater training improve the quality and accuracy of workplace-based assessment narrative comments and entrustment ratings? A randomized controlled trial. *Acad Med*. 2023;98(2):237-47.