Diversity, Equity, Inclusion, and Justice

Racial and Ethnic Disparities in Situational Judgment Testing Among Applicants to an Anesthesiology Residency Program

Victoria Rosales, MD
Christopher Conley, MD
Mark C. Norris, MD

ABSTRACT

Background The Computer-Based Assessment for Sampling Personal Characteristics (CASPer) is a situational judgment test (SJT) that assesses noncognitive skills like professionalism, communication, and empathy. There are no reports of the effects of race/ethnicity and sex on CASPer scores among residency applicants.

Objective We examined the effects of race/ethnicity, sex, and United States vs international medical school attendance on CASPer performance.

Methods Our anesthesiology residency program required all applicants for the 2021-2022 Match cycle to complete an online video and text-based SJT (CASPer). We compared these results, reported as z-scores, with self-identified race/ethnicity, sex, United States vs international medical school attendance, and United States Medical Licensing Examination (USMLE) Step 1 scores.

Results Of the 1245 applicants who completed CASPer, 783 identified as male. The racial/ethnic distribution was 512 White, 412 Asian, 106 Black, 126 Hispanic, and 89 Other/No Answer. CASPer z-scores did not differ by sex. White candidates scored higher than Black (0.18 vs -0.57, $P<.001$) and Hispanic (0.18 vs -0.52, $P<.001$) candidates. Applicants attending US medical schools scored higher than those attending international medical schools (z-scores: 0.15 vs -0.68, $P<.001$). There was no correlation between CASPer z-scores and USMLE Step 1 scores.

Conclusions Our results suggest that CASPer scores favor White applicants over Black and Hispanic ones and applicants attending US medical schools over those attending international medical schools.

Introduction

A diverse health care workforce can improve access to care for underserved communities and expand the scope of research agendas.[1] Unfortunately, the current US physician workforce is less diverse than the population. The Association of American Medical Colleges (AAMC) defines underrepresented in medicine (UIM) as: "those racial and ethnic populations that are underrepresented in the medical profession relative to their numbers in the general population."[2] Physicians who identify as Black or Hispanic make up less than 15% of the US residency workforce.[1] In 2018, Black and Hispanic physicians represented only 5% and 5.8% of practicing doctors, although these groups constitute one-third of the US population.[1] Thus, Black and Hispanic physicians are significantly underrepresented in the medical workforce and are considered UIM. Many applicants to US residency programs are international medical school graduates (IMGs) and represent 23% of US physicians. Although 18% of IMGs are American or Canadian citizens, others bring diverse languages and cultures and represent a kind of diversity not often measured.[3] There is widespread consensus on the need to increase the UIM physician population in the United States. The Accreditation Council for Graduate Medical Education (ACGME) explicitly views diversifying the physician workforce as essential to eradicating health care inequalities.[1]

US residency programs review an ever-increasing number of applications, and many programs initially screen candidates using United States Medical Licensing Examination (USMLE) Step scores. This disadvantages UIM candidates, who score worse as a group on these examinations than their White counterparts.[4] These differences in performance can be related to disparities in resources or biases in test construction.[5] To address this, the ACGME and other groups advocate for holistic review of applications, considering each

DOI: http://dx.doi.org/10.4300/JGME-D-23-00360.1

candidate's unique background and potential.[6] Situational judgement tests (SJTs) may be a component of holistic review.[7]

SJTs aim to measure noncognitive skills such as empathy and judgment through responses to hypothetical scenarios. They can be used as a screening tool in personnel selection. While considerable data exists correlating SJT scores with job performance, there is less available information examining differences in SJT performance among different racial and ethnic groups.[7] The Computer-Based Assessment for Sampling Personal Characteristics (CASPer) is a commercially available online video-based SJT that was developed to assist in evaluating candidates for medical school admission. CASPer is currently used to screen applicants to 22 US allopathic and osteopathic medical schools.[8] More recently, CASPer is being marketed as a tool to evaluate residency applicants.[9] There is little published information concerning racial and ethnic differences in CASPer scores. However, 2 studies of medical school applicants have reported significant racial and ethnic differences in CASPer performance.[8,10] To our knowledge, there are no peer-reviewed publications examining CASPer scores in residency applicants.

In 2021 Boston Medical Center (BMC)/Boston University Chobanian & Avedisian School of Medicine joined 22 other residency programs in 4 specialties in a project sponsored by Altus Assessments (now Acuity Insights) piloting the use of CASPer as a screening tool to evaluate residency applicants.[9] In this study we report the effects of race and sex, as well as US vs international medical school graduation on CASPer scores. Because Black and Hispanic candidates made up the largest component of our UIM residency program applicants, we focused our analysis on these groups. Examining racial and ethnic differences in CASPer scores will allow us to determine whether this screening tool will effectively increase the diversity of applicants selected for interview.

KEY POINTS

What Is Known
As holistic review for residency applicants expands, it is critical to ensure new biases are not introduced into the selection process.

What Is New
This study of applicants to one anesthesiology residency program found that CASPer, a situational judgment test, favored White applicants in this particular use of the test.

Bottom Line
Programs looking to add new selection tools to their application process should be advised of this limitation of CASPer.

Methods

BMC is a northeastern urban safety net hospital with a primarily clinically focused anesthesiology residency program. We participate in the National Resident Matching Program and were one of 22 residency programs to require applicants to the 2022 Match to complete CASPer. The BMC Anesthesiology residency has both categorical and advanced positions, with 8 and 4 in each group respectively.

The SJT used in our study, CASPer, is a proprietary test created to assess noncognitive skills as a part of medical professions selections.[11] CASPer is comprised of 12 sections presenting ethical dilemmas through video and text prompts. Test-takers respond in writing or video recording. The video scenarios encompass a variety of nonclinical themes such as collaboration, communication, professionalism, and confidentiality.[11] These topics align with key nonmedical attributes identified by both the Royal College of Physicians and Surgeons of Canada and the ACGME.[11] These video scenarios offer a more immersive context than written options and demand fewer resources than live standardized assessments.[11] To evaluate candidates' performance, raters undergo training on the scoring process with background information and theoretical knowledge for each scenario.[11] These raters gauge the candidates' communication abilities, the strength of their arguments, their suitability for a medical career, and overall performance within the given situation.[11] For further reference and materials, see Dore et al.[11]

Candidates applied to our program through the Electronic Residency Application Service (ERAS). We obtained permission from the AAMC to use demographic and other information for applicants for the 2021-2022 Match cycle. We extracted the fields: AAMC ID, Applicant Name, USMLE Step 1 Score, Medical School of Graduation, Self-Identified Race/Ethnicity, and Sex from the CSV file "All Applicants" downloaded from the AAMC ERAS Program Director's Workstation (https://pdws.aamc.org/eras-pdws-web/home/dashboardPanel/). Additionally, we obtained permission from Altus Suite to use data regarding applicants who completed the CASPer assessment and applied to our residency program. We extracted the fields: First Name, Last Name, AAMC (ID), and CASPer z-score from a second file containing the Altus Suite assessments of all applicants downloaded from Altus Insights. These files were then combined by matching AAMC ID and Last Name. The name fields were deleted after matching to deidentify the candidates.

Altus Assessments reports the results of the CASPer assessment as a standard score, or z-score, which represents the number of standard deviations that a raw

score is above or below the mean score. The z-score is calculated using the formula:

$$Z = \frac{x - \mu}{\sigma}$$

Z = standard score
x = observed score
μ = sample mean
σ = standard deviation of the sample

Our data were not normally distributed, therefore nonparametric tests were selected for analysis. Mann-Whitney U and Kruskal-Wallis tests were used to look performance differences based on sex, self-identified race/ethnicity, and US vs international medical school of graduation. Dunn's test is a post hoc test that was used in conjunction with the Kruskal-Wallis test to examine subgroup differences among self-identified race/ethnicity.

The present study was deemed to meet the criteria for exemption from institutional review board review.

Results

We had 1494 applicants to our residency program; 1245 also had a CASPer score. Four hundred and sixty-two of 1245 (37%) identified as female, 783 (63%) as male. The racial/ethnic distribution was: 512 (41%) White, 412 (33%) Asian, 106 (9%) Black, 126 (10%) Hispanic, and 89 (7%) Other/No Answer (TABLE). We did not include the "Other/No Answer" group in any subsequent analysis. Nine hundred and ninety-seven (80%) were graduates of US medical schools. CASPer z-scores did not differ by sex (FIGURE 1). There were significant racial/ethnic differences in CASPer z-scores. White candidates scored higher than Black (0.18 [IQR 1.13] vs -0.57 [IQR 1.29], $P<.001$) and Hispanic (0.18 [IQR 1.13] vs -0.52 [IQR 1.22], $P<.001$) candidates (FIGURE 2). However, there was no difference between White and Asian candidates (FIGURE 2). Applicants attending US medical schools scored higher than IMGs (0.15 [IQR 1.15] vs -0.68 [IQR 1.52], $P<.001$) (FIGURE 3). There was no correlation between CASPer z-scores and USMLE Step 1 scores.

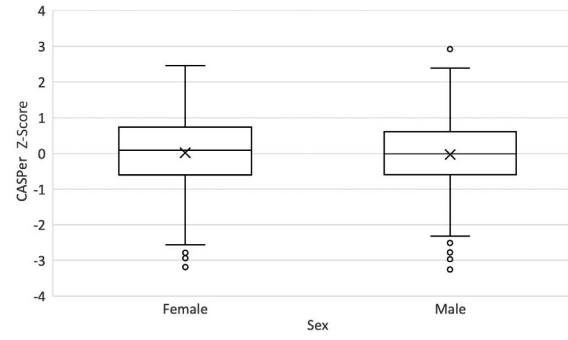

FIGURE 1
Differences in CASPer Z-Score by Sex

Discussion

This study suggests that including CASPer scores in a holistic review of residency applicants is unlikely to overcome the existing biases against UIM candidates. White applicants performed significantly better on CASPer than their Black and Hispanic counterparts. US medical school applicants also performed significantly better than IMGs.

Our results with CASPer scores are consistent with studies of SJTs designed for use outside of medical education, which also produce performance disparities between majority and minority groups. For instance, a study of Belgian job seekers showed significant differences in SJT performance between

TABLE
Participant Characteristics, N=1245

Characteristic	n (%)	Z-Score Mean	Z-Score Standard Deviation
Sex			
Female	462 (37)	0.02	1.03
Male	783 (63)	-0.04	0.95
Race			
White	512 (41)	0.18	0.89
Black	106 (9)	-0.57	1.00
Hispanic	126 (10)	-0.52	1.02
Asian	412 (33)	0.03	0.96
Other	89 (7)		
Medical school			
US medical school	997 (80)	0.15	0.88
Non-US medical school	248 (20)	-0.68	0.15

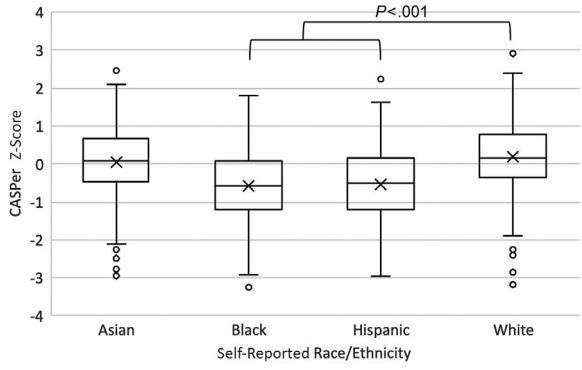

FIGURE 2
Differences in CASPer Z-Score by Self-Identified Race/Ethnicity

native-born and immigrant candidates.[12] In another study, candidates from majority racial groups and higher socioeconomic backgrounds outperformed racial and ethnic minoritized groups and those from lower socioeconomic backgrounds on an SJT designed for screening prospective teachers in the United Kingdom.[13]

Studies of medical school applicants also report biases in SJT results. One study reported results of a computer simulation examining the impact of including CASPer scores when selecting applicants for medical school interview.[10] Increasing the weight of CASPer scores led to more female, Black, and Hispanic/Latino candidates being selected.[10] However, the inclusion of CASPer decreased the number of self-declared disadvantaged applicants.[10] Additionally applicants from racial and ethnic minoritized groups and lower socioeconomic backgrounds performed worse on CASPer than their White or economically advantaged peers.[10] Notably, the differences in CASPer scores were of lesser magnitude than differences in Medical College Admission Test scores.[10] A more recent study of applicants to a southeastern US medical school also reported significant differences in CASPer scores for gender and race.[8] Additionally, they found only a weak correlation between CASPer and traditional interview scores.[8]

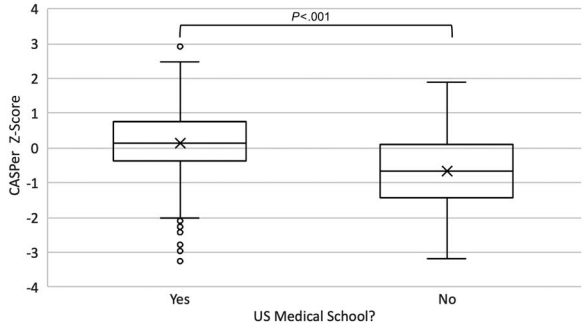

FIGURE 3
Differences in CASPer Z-Score by attendance at US Medical School

There are isolated reports of the use of SJTs in residency and fellowship selection, but none use CASPer. One study reported the effects of adding an SJT and lowering the USMLE Step 1 cutoff on the diversity of candidates interviewed by surgical residencies.[14] The combination resulted in an 8% increase in the number of UIM candidates who were invited to interview ($P<.01$).[14] However, lowering the USMLE Step 1 cutoff alone also decreased the bias against applicants of color.[14] In another study of candidates for surgical fellowships, the percentage of UIM applicants who were interviewed and subsequently ranked increased from 70% to 92% when an SJT was used.[15] Importantly both studies included female candidates in their definition of "UIM." In addition, neither describe how they used the SJT scores in their screening and selection process.[14,15] Thus the impact of SJTs on UIM inclusion in residency selection is unclear from the literature.

As is often the case with other standardized tests, disparities in SJT performance may stem from hidden biases in their development, administration, and interpretation. Lievens et al have reviewed the impact of SJT construction on majority-minority differences. Tests that require timed, written responses may disadvantage non-English speakers.[12] Questions that are culturally specific or difficult to understand may discriminate against those from different cultural backgrounds or non-native English speakers.[12] SJT response types range from multiple choice, ranked choice, and either written, or video-based free responses. This can influence the majority-minority differences in SJT scores. Multiple-choice questions are associated with greater majority-minority disparities than a free text format. Minority-majority differences are further reduced in the audiovisual format compared to written responses.[12] These differences may relate to the greater cognitive load required by multiple-choice or written responses. Culturally specific questions, time limits, and multiple-choice questions may all favor majority over minority applicants.[12]

Bias can also affect the grading of SJTs. CASPer uses both written and video recorded responses. Written responses afford greater anonymity to candidates, minimizing the impact of factors such as sex, ethnicity, and other potential sources of bias. However, time constraints, language fluency, and cultural background may unintentionally disadvantage non-native English speakers and those from minority backgrounds.[12] The grading of recorded or video responses may also be subject to bias. For instance,

Muslim applicants and those with a foreign accent receive lower ratings if visual or auditory information is considered.[12] In CASPer validation studies, the authors provide minimal information on measures to reduce scoring bias. Thus, hidden or unacknowledged prejudices may contribute to the racial/ethnic disparities observed in CASPer and similar SJTs.

Our study has some limitations. Our sample includes 49% of 2022 US anesthesiology applicants but only 3% of total 2021-2022 Match cycle applicants to US residency programs.[16] We used US vs international medical school as a proxy for native language. This choice is likely flawed as many non-native students attend US medical schools and many US students enroll in international medical schools.

Conclusions

Although SJTs, like CASPer, assess different skills than traditional multiple-choice examinations, our experience of applicants to a single anesthesiology residency program suggests SJT results still favor White applicants over Black or Hispanic applicants. Applicants attending US medical schools outperformed IMG applicants.

References

1. Crites K, Johnson J, Scott N, Shanks A. Increasing diversity in residency training programs. *Cureus*. 2022;14(6):e25962. doi:10.7759/cureus.25962
2. Association of American Medical Colleges. Underrepresented in medicine. Accessed September 13, 2023. https://www.aamc.org/what-we-do/equity-diversity-inclusion/underrepresented-in-medicine
3. Nagarajan KK, Bali A, Malayala SV, Adhikari R. Prevalence of US-trained international medical graduates (IMG) physicians awaiting permanent residency: a quantitative analysis. *J Community Hosp Intern Med Perspect*. 2020;10(6):537-541. doi:10.1080/20009666.2020.1816274
4. Williams M, Kim EJ, Pappas K, et al. The impact of United States Medical Licensing Exam (USMLE) Step 1 cutoff scores on recruitment of underrepresented minorities in medicine: a retrospective cross-sectional study. *Health Sci Rep*. 2020;3(2):e2161. doi:10.1002/hsr2.161
5. Jones AC, Nichols AC, McNicholas CM, Stanford FC. Admissions is not enough: the racial achievement gap in medical education. *Acad Med*. 2021;96(2):176-181. doi:10.1097/acm.0000000000003837
6. Association of American Medical Colleges. Holistic review. Accessed April 11, 2023. https://www.aamc.org/services/member-capacity-building/holistic-review
7. Patterson F, Zibarras L, Ashworth V. Situational judgement tests in medical education and training: research, theory and practice: AMEE guide no. 100. *Med Teach*. 2016;38(1):3-17. doi:10.3109/0142159X.2015.1072619
8. Gustafson CE, Johnson CJ, Beck Dallaghan GL, et al. Evaluating situational judgment test use and diversity in admissions at a southern US medical school. *PLoS One*. 2023;18(2):e0280205. doi:10.1371/journal.pone.0280205
9. Acuity Insights. Casper. Accessed January 15, 2024. https://acuityinsights.com/product-overview/admissions-assessments/casper/
10. Juster FR, Baum RC, Zou C, ET AL. Addressing the diversity-validity dilemma using situational judgment tests. *Acad Med*. 2019;94(8):1197-1203. doi:10.1097/ACM.0000000000002769
11. Dore KL, Reiter HI, Eva KW, et al. Extending the interview to all medical school candidates—computer-based multiple sample evaluation of noncognitive skills (CMSENS). *Acad Med*. 2009;84(suppl 10):9-12. doi:10.1097/ACM.0b013e3181b3705a
12. Lievens F, Sackett PR, Dahlke JA, Oostrom JK, De Soete B. Constructed response formats and their effects on minority-majority differences and validity. *J Appl Psychol*. 2019;104(5):715-726. doi:10.1037/apl0000367
13. Bardach L, Rushby JV, Klassen RM. The selection gap in teacher education: adverse effects of ethnicity, gender, and socio-economic status on situational judgment test performance. *Br J Educ Psychol*. 2021;91(3):1015-1034. doi:10.1111/bjep.12405
14. Gardner AK, Cavanaugh KJ, Willis RE, Dunkin BJ. Can better selection tools help us achieve our diversity goals in postgraduate medical education? Comparing use of USMLE Step 1 scores and situational judgment tests at 7 surgical residencies. *Acad Med*. 2020;95(5):751-757. doi:10.1097/ACM.0000000000003092
15. Gardner AK, Dunkin BJ. Pursuing excellence: the power of selection science to provide meaningful data and enhance efficiency in selecting surgical trainees. *Ann Surg*. 2019;270(1):188-192. doi:10.1097/SLA.0000000000002806
16. National Resident Matching Program. Residency data reports. Accessed September 23, 2023. https://www.nrmp.org/match-data-analytics/residency-data-reports/

Victoria Rosales, MD, is an Anesthesiology Resident, Department of Anesthesiology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA; **Christopher Conley, MD,** is Clinical Associate Professor of Anesthesiology, Associate Residency Program Director, and Director of Pediatric Anesthesia, Department of Anesthesiology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA; and **Mark C. Norris, MD,** is Clinical Professor of Anesthesiology, Residency Program Director, and Director of

Obstetric Anesthesia, Department of Anesthesiology, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

This work was previously presented as a poster presentation at the Accreditation Council for Graduate Medical Education Annual Educational Conference, February 23-25, 2023, Nashville, Tennessee, USA.

The authors would like to thank Lan Zu, PhD, Boston Medical Center, for statistical analysis.

Corresponding author: Victoria Rosales, MD, Boston University Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA, victoria.rosales@bmc.org

Received May 19, 2023; revisions received September 25, 2023, and January 17, 2024; accepted January 23, 2024.