

Unintended Consequence: Diversity as a Casualty of Eliminating United States Medical Licensing Examination Step 1 Scores



Felipe M. Campos, MS^a, Lars J. Grimm, MD, MHS^b, Charles M. Maxfield, MD^c, and the Radiology Residency Education Research Alliance

Abstract

Purpose: The purpose of this study was to use a discrete-choice experiment to model the trade-offs evaluators make between academic attributes and demographics when the United States Medical Licensing Examination (USMLE) Step 1 switches to pass/fail.

Methods: A discrete-choice experiment was administered to faculty members from a geographically diverse mix of 14 academic and community radiology departments in the United States from August through November 2020. Reviewers reviewed 10 applicant pairs with numeric Step 1 scores (part 1) and 10 applicant pairs with a pass Step 1 result (part 2). Applicant attributes included medical school rank, gender, race/ethnicity, USMLE Step 1 score, USMLE Step 2 score, class rank, clerkship honors, and publications. Conditional logistic regression modeled the influence of attribute levels.

Results: Two hundred twelve evaluators completed the study (response rate 59%). The most influential attribute was Step 1 score in part 1 and medical school rank in part 2. The relative importance of race/ethnicity and gender decreased by 25% and 29%, respectively, when Step 1 switches to pass/fail. Evaluators weigh race/ethnicity the strongest when applicants have the same Step 1 score (preference weights of 0.85 for African American, 1.42 for Hispanic, and 0 for White and Asian applicants). Race/ethnicity is relatively more important when Step 1 scores are higher (preference weights of 1.58 for African American, 0.90 for Hispanic, and 0 for White and Asian applicants).

Conclusions: The loss of numeric Step 1 scores reduced the residency evaluator preference for diversity. Reviewers prioritize underrepresented-in-medicine applicants when Step 1 scores are higher and comparable with White and Asian applicants.

Key Words: USMLE Step 1, residency, recruitment, diversity, discrete-choice experiment

J Am Coll Radiol 2023;20:1177-1187. Copyright © 2023 American College of Radiology

INTRODUCTION

Black and Hispanic physicians remain underrepresented in medicine (URiM) [1], especially in radiology, which ranks among medicine's least racially and ethnically diverse specialties [2]. Radiology residency program directors and selection committees, as gatekeepers of the profession, strive for more inclusive resident selection to improve diversity in our field [3-6]. These efforts center on a holistic application review, which balances academic metrics with life experiences

and personal attributes to gauge an applicant's likelihood of contributing to learning, practice, and teaching [6].

The growing challenge with holistic application reviews is the progressive loss of traditional measures of academic achievement to balance the nonacademic attributes. Increasingly, US medical schools are reporting grades as pass/fail, and fewer schools are providing comparative performance rankings [7]. Additionally, Step 1 of the United States Medical Licensing Examination (USMLE), a

^aSchool of Pharmacology, University of Washington, Seattle, Washington.

^bDepartment of Radiology, Duke University, Durham, North Carolina.

^cVice Chair of Education, Department of Radiology, Duke University, Durham, North Carolina.

Corresponding author and reprints: Lars Grimm, MD, MHS, Department of Radiology, Duke University, Box 3808, Durham, NC 27710; e-mail: lars.grimm@duke.edu.



Follow this author via Twitter: Lars J. Grimm @Dr_Lars_Grimm

The authors state that they have no conflict of interest related to the material discussed in this article. The authors are non-partner/non-partnership track/employees.

What are the tradeoffs when the United States Medical Licensing Exam (USMLE) Step 1 switches to pass/fail?



STEP 1

Many believe that reporting USMLE Step 1 as pass/fail will advantage those underrepresented in medicine (URiM), although these scores have previously been good predictors of residency success.



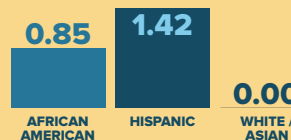
DISCRETE CHOICE EXPERIMENT

Reviewers reviewed ten Step 1 applicant pairs. Each had:

- An applicant with numerical scores (Part 1)
- An applicant with pass/fail determination (Part 2)

Pass/fail caused noticeable reduction in preference given to **race/ethnicity (25%)** and **gender (29%)**

Ethnicity mattered most for pairs with like Step 1 scores



The loss of numerical Step 1 scores reduced the residency evaluator preference for diversity. Reviewers prioritize URiM applicants when Step 1 scores are higher and comparable to white/Asian applicants.

JACR VISUAL ABSTRACT

historically reliable predictor of radiology resident success when reported as a three-digit numeric score [8,9], is now reported only as a pass/fail result [10]. Radiology residency program directors and selection committees must now assess and compare medical student applicants without their most trusted academic metrics [11].

The impact of pass/fail Step 1 scoring on URiM applicants has been a source of considerable conjecture. Most believe that URiM applicants will be advantaged by the loss of a metric on which they as a group have performed less well than other demographic groups [12]. Holistic application reviews seek to strike a balance between academic and nonacademic factors in building a residency class, but an understanding of how applicant reviewers will assess URiM applicants in the absence of Step 1 scores is currently unavailable.

Therefore, the purpose of this study was to use a discrete-choice experiment (DCE) to model the trade-offs evaluators make between academic attributes and demographics, specifically race/ethnicity and gender, in selecting radiology residency applicants. A DCE is a quantitative method for studying the relative importance of different attributes in decision-making processes. By eliciting individual preferences between hypothetical resident applicant profiles containing multiple varying attributes, the relative influence of each application attribute and the willingness of participants to trade one attribute for another can be calculated [13]. We used a DCE dataset that previously showed the prioritization

of URiM applicants in a simulated residency selection exercise to perform more sophisticated modeling that allows us to quantify the specific trade-offs residency reviewers make between diversity and academic metrics [14]. Furthermore, we collected real-world applicant information to validate the DCE methodology as a means of studying residency selection.

METHODS

Subjects

Faculty participants were solicited from 12 academic and 2 community radiology departments in the United States. Geographically, four departments were in the South, three in the Midwest, two in the Northeast, and two in the West. All participating institutions were members of the Radiology Residency Education Research Alliance, a research collaborative of US radiology residency programs interested in resident education research [15]. All sites were instructed to invite all faculty members involved in residency selection to participate. Participants were not individually identified.

Experimental Design

Hypothetical residency applicants were presented as alternative choices differing in important application attributes, which were selected on the basis of published data and randomized for presentation [16]. Eight attributes were selected,

and each attribute was assigned up to four fixed levels. The levels were chosen to be realistic (ie, attribute values span the typical distribution of radiology residency applications) and discriminating (ie, adequate separation between attribute levels). Attributes and levels are shown in Table 1. Fifteen medical school were grouped into three categories for subsequent analysis on the basis of the 2020 *US News & World Report* rankings: top 10 ranked, midlevel ranked, and unranked [17]. A single race or ethnicity attribute was used with four levels (Asian, Black, Hispanic, White).

Twenty pairs of applicant profiles were presented as hypothetical candidates for discrete choices: the first 10 choice pairs included numerical Step 1 scores (part 1), and the next 10 included only a pass result (part 2). The combinations of attribute levels were determined according to a balanced overlap design generated by Sawtooth (Sawtooth Software). Each participant was assigned to 1 of 300 blocks of choice pairs to randomize the presentation of applicants. Each participant evaluated up to total 20 choice pairs. For each choice pair, participants were asked, “Which applicant would you choose to invite for an interview?” A sample choice set is presented in Figure 1. The questionnaire was piloted among nonparticipating faculty members at one institution, feedback from which was iteratively included to improve formatting and question clarity.

Procedure

The experiment was conducted from August through November 2020. Fourteen institutional site investigators sent e-mails to a total of 360 departmental faculty members who have participated in their residency programs’ selection of residents through screening, interviewing, or ranking applicants (mean, 25.7 faculty members per institution; range, 11-43). The e-mail stated that the purpose of the study was to evaluate application factors important in resident selection and contained a common link to the web-based DCE survey hosted by Sawtooth Software. Optional demographic questions inquired as to the participants’ self-identified gender, race/ethnicity, academic rank, and experience in residency selection. Reviewers were assured that their choices would be confidential and their anonymity maintained.

Statistical Analysis

A conditional logistic regression was chosen to model the choice of applicant A or applicant B. This model was used to relate the dichotomous choice between two residency candidates conditional on the attribute levels shown for each candidate. The modeling approach was shown to be consistent with random utility theory, which posits that a respondent will chose among alternatives to maximize their utility. The results of the conditional logit model provide log-odds preference weights and corresponding standard errors.

Table 1. Attributes and levels in the discrete-choice experiment

| Attribute | Levels |
|-----------------------------|--|
| Medical school rank | Top 10, midlevel ranked, unranked |
| Gender | Male, female |
| Race/ethnicity | White, Black, Asian, Hispanic |
| USMLE Step 1 score (part 1) | 202, 228, 246, 269 |
| USMLE Step 1 score (part 2) | Pass |
| USMLE Step 2 score | 213, 229, 248, 267 |
| Class rank | First, second, third, fourth quartiles |
| Core clerkship honors | 1, 3, 5, 6 |
| Number of publications | 0, 1, 3, 6 |

Note: Medical school rank refers to the *US News & World Report* rankings from 2020. USMLE = United States Medical Licensing Examination.

Formally,

$$P(A \text{ chosen and } B \text{ not chosen} | \text{one chosen applicant}) = \frac{\exp(\mathbf{x}_A \mathbf{b})}{\exp(\mathbf{x}_A \mathbf{b}) + \exp(\mathbf{x}_B \mathbf{b})},$$

where \mathbf{x}_A and \mathbf{x}_B correspond to the vector of values of the attributes of applicants A and B, respectively, and \mathbf{b} is the vector of parameters associated with the attributes.

In the initial model, all attributes were dummy-coded to evaluate whether each of the levels within an attribute were significantly associated with participants’ choices and to evaluate the functional form for the continuous attributes: USMLE scores, class rank, clerkship honors, and publications. Akaike information criteria and log-likelihood values were used to assess model fit across various specifications, including comparisons of dummy-coded attributes versus continuous functional forms and main-effects models and models with interactions. Odds ratios were computed to represent the relative influence of attribute levels on participants’ choices. URiM groups (Black and Hispanic) were compared with represented groups (White and Asian) for the analysis of race and ethnicity,

Relative Importance

To facilitate comparisons of attributes on residency selections between part 1 (USMLE Step 1 numeric score) and part 2 (USMLE Step 1 pass), we computed the relative importance weights for each attribute in each part. To accomplish this, we computed the difference in log-odds

Which applicant would you choose to invite for an interview?

(1 of 10)

| | Applicant A | Applicant B |
|---------------------|---------------------------|---------------------|
| Medical School | Stanford | University of Miami |
| Gender | Male | Female |
| Race/ Ethnicity | Black or African American | White |
| USMLE Step 1 Score | 246 | 269 |
| USMLE Step 2 Score | 229 | 267 |
| Class Quartile Rank | 2nd | 3rd |
| Clerkship Honors | 6 of 6 | 1 of 6 |
| Publications | 6 | 1 |
| | Select | Select |

Back

Next

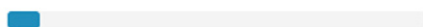
0%  100%

Fig. 1. Example of discrete choice from part 1 with numeric Step 1 scores. In part 2, the United States Medical Licensing Examination (USMLE) Step 1 score was changed to “pass.” No applicants had a “fail” score, as they would not be allowed to apply to residency.

parameter estimates from the conditional logit model between the least preferred level and the most preferred level within an attribute to determine the attribute importance conditional on the levels included in the experiment. Using this metric, we calculated two measures of relative importance across attributes. For relative importance 1, we chose the attribute with the greatest absolute difference as a reference attribute and set its value to 100 and rescaled the value of the other attributes as a proportion of importance relative to the reference attribute for parts 1 and 2. For relative importance 2, we rescaled the relative importance of all attributes such that total weights summed to 100 for parts 1 and 2.

Pairwise Comparison by Step 1 Score Difference

We exploited variations in the Step 1 scores in part 1 (attribute levels 202, 228, 246, and 269) to explore the

heterogeneity of preferences on race/ethnicity when participants encounter pairs of applicants with different combinations of Step 1 scores. For this part of the analysis, we used only part 1 of the experiment, and we defined four groups of applicant pairs. The criterion for defining the groups was how far or close the pair of applicants are in terms of the Step 1 score (Table 2). Then, we estimated a conditional logistic regression in each of the subgroups including all the attributes and plotted the log odds of school of medicine, Step 2 score, race/ethnicity, and gender.

However, two applicants who have close but high Step 1 scores (eg, 246 versus 269) may not be viewed the same as two applications with close by low Step 1 scores (eg, 202 versus 228). We therefore defined an additional subgroup among pairs one category apart in Step 1 score as shown in Table 3: low academic quality, mid academic quality, and high academic quality. Then, we estimated the conditional logit in these three subgroups defined previously to

Table 2. Definitions for the subgroups of paired Step 1 scores used in part 1 of the experiment

| Group | Definition |
|--------------------------|--|
| Same Step 1 score | Same Step 1 scores (n = 85) |
| Small Step 1 difference | Step 1 scores one category apart (n = 968) |
| Medium Step 1 difference | Step 1 scores two categories apart (n = 651) |
| Big Step 1 difference | Step 1 scores three categories apart (n = 312) |

Note: n refers to the number of applicant pairs in each group.

analyze if the importance of attributes changes when applicants are in each subgroup.

Validation

Because DCE is an experimental modality meant to simulate real-world conditions, we performed a validation experiment. Information from actual radiology residency applicants to Duke University in 2021 (n = 912) were collected, as well as whether the applicants were invited for an interview. A linear probability model and logistic regression model were created using medical school rank, gender, race/ethnicity, and Step 1 and Step 2 as dependent variables. Step 1 and Step 2 scores were converted into four-level quartiles to mirror the DCE levels on the basis of the actual distribution of Step 1 and Step 2 scores. The dependent variable was whether the applicant was invited for an interview (1 = yes, 0 = no).

RESULTS

Respondent Demographics

Of the 360 individuals invited to participate, 243 opened the survey and 212 completed at least one choice question (response rate 59% [212 of 360]). Six participants answered only one query; 175 participants answered all 20 queries (mean, 17.9; median, 20). We used data from the 212 participants who responded to at least one choice question in part 1 and the 177 participants who responded to at least one choice question in part 2. Of participants who self-reported gender, the plurality was female (49% [104 of 212]), and of those self-reporting race, the plurality was White (65% [138 of 212]). Academic rank was as follows: 33% (69 of 212) assistant professor, 21% (45 of 212) associate professor, and 19% (40 of 212) professor. As for experience reviewing applications, 36% (76 of 212) reported 1 to 5 years' experience, 20% (43 of 212) reported greater

than 10 years, and 18% (38 of 212) reported 6 to 10 years. The demographics of participants are shown in Table 4.

Relative Importance

The relative importance of attributes in parts 1 and 2 via the relative importance 1 metric is shown in Figure 2. Relative to the reference class (Step 1 score in part 1 and medical school rank in part 2), there is either no change or an increase in importance for most of the academic attributes (medical school rank, Step 2, publications, clerkship honors, and class rank) from part 1 to part 2. However, there was a discrete decrease in the importance of race/ethnicity and gender. In Figure 3, the relative importance of the most important attribute, race/ethnicity, and gender is shown. Although the absolute decreases in the relative importance of the race/ethnicity (from 49 to 37) and gender (from 14 to 10) attributes from part 1 to part 2 are small, the relative decreases are larger: a 25% decrease for race/ethnicity and a 29% decrease for gender. Similarly, in the relative importance 2 metric, the aggregate importance of the academic attributes (medical school rank, clerkship honors, class rank, Step 2 score, and publications) increased in part 2 (from 67 to 73, a 9% increase), and the importance of demographic (race/ethnicity and gender) attributes decreased (from 33 to 27, an 18% decrease), as shown in Figure 3.

Pairwise Comparison by Step 1 Score Difference

The conditional logistic regression for the pairwise comparisons for school of medicine, Step 2 score, race/ethnicity, and gender are plotted in Figure 4 and in Supplementary Tables 1 to 4. The more evaluators can separate applicants using the Step 1 score metric (ie, big Step 1 difference versus small Step 1 difference), the smaller the preference weights (log-odds ratio) are for school of medicine and Step 2 score. Similarly, evaluators weigh race/ethnicity strongest when the two applicants have the same or very close Step 1 scores. In contrast, the impact of gender is relatively flat regardless of the Step 1 score difference.

The conditional logistic regression results for the subgroup of paired applicants with only small Step 1 difference broken down into low quality, mid quality, and high quality are

Table 3. Definitions for the subgroups of applicants with small Step 1 scores

| Group | Definition |
|-----------------------|----------------------------|
| Low academic quality | Pairs 202 vs 228 (n = 326) |
| Mid academic quality | Pairs 228 vs 246 (n = 308) |
| High academic quality | Pairs 246 vs 269 (n = 334) |

Table 4. Participant demographics

| Demographic | n (%) |
|---|-----------|
| Total | 212 (100) |
| Gender | |
| Male | 75 (35) |
| Female | 104 (49) |
| Prefer not to say or skipped | 32 (15) |
| Race | |
| White | 138 (65) |
| Asian | 29 (14) |
| Black or African American | 2 (1) |
| American Indian or Alaska Native | 1 (0) |
| Native Hawaiian or other Pacific Islander | 1 (0) |
| Other | 6 (3) |
| Prefer not to say or skipped | 35 (17) |
| Ethnicity | |
| Not Hispanic or Latino | 170 (80) |
| Hispanic or Latino | 5 (2) |
| Prefer not to say or skipped | 37 (17) |
| Academic rank | |
| Clinical or medical instructor | 12 (6) |
| Assistant professor | 69 (33) |
| Associate professor | 45 (21) |
| Professor | 40 (19) |
| Prefer not to say or skipped | 46 (22) |
| Experience reviewing applications | |
| None | 26 (12) |
| 1-5 y | 76 (36) |
| 6-10 y | 38 (18) |
| >10 y | 43 (20) |
| Prefer not to say or skipped | 29 (14) |

plotted in Figure 5 and in Supplementary Tables 5 to 7. Step 1 score is most important when Step 1 scores overall are lower. In contrast, as Step 1 scores increases, the relative importance of medical school rank, Step 2 score, and race/ethnicity increases as these factors replace the importance of Step 1. However, the relative importance of gender was relatively flat.

Finally, the interaction of medical school rank and race/ethnicity in the conditional logit model for the four subgroups was plotted, as shown in Figure 6 and Supplementary Tables 8 to 10. Only when the two applicants are close in Step 1 score and in the high-quality group is the effect of race/ethnicity important. Otherwise, the effect of race/ethnicity is driven by the interaction with medical school rank.

Validation

The findings from the DCE align with real-world interview invitations, as shown in Table 5. Relative to unranked universities, applicants from top 10 ranked schools were 30% (95% confidence interval [CI], 21%-39%; $P < .001$) more likely to be invited for an interview. Female (relative to male) applicants were 38% (95% CI, 0.08%-76%; $P = .045$) more likely to be invited for an interview, whereas Black or Hispanic (relative to White or Asian) applicants were 4% (95% CI, 0.06%-8.2%; $P = .047$) more likely to be invited for an interview. Finally, a USMLE Step 1 score in the highest quartile (relative to the lowest quartile) makes the applicant 13% (95% CI, 7.2%-18%; $P < .001$) more likely to receive an interview invitation. There were no significant differences according to USMLE Step 2 score.

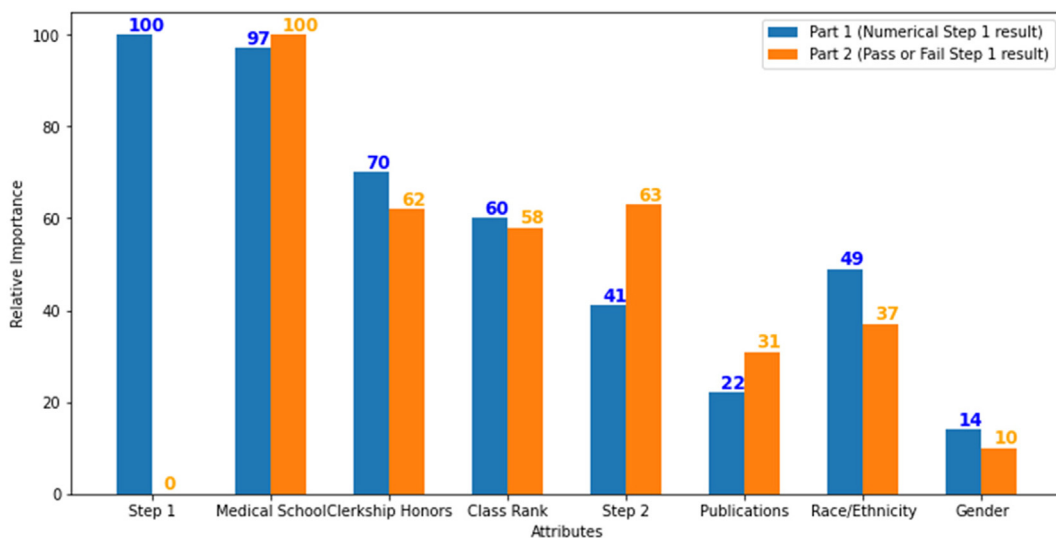


Fig. 2. Relative importance 1 metric for all the attributes in parts 1 and 2. The attribute with the greatest absolute difference was chosen as the reference attribute and set to 100, with all other attributes proportionately rescaled. The reference class for part 1 is Step 1 score, and the reference class for part 2 is medical school.

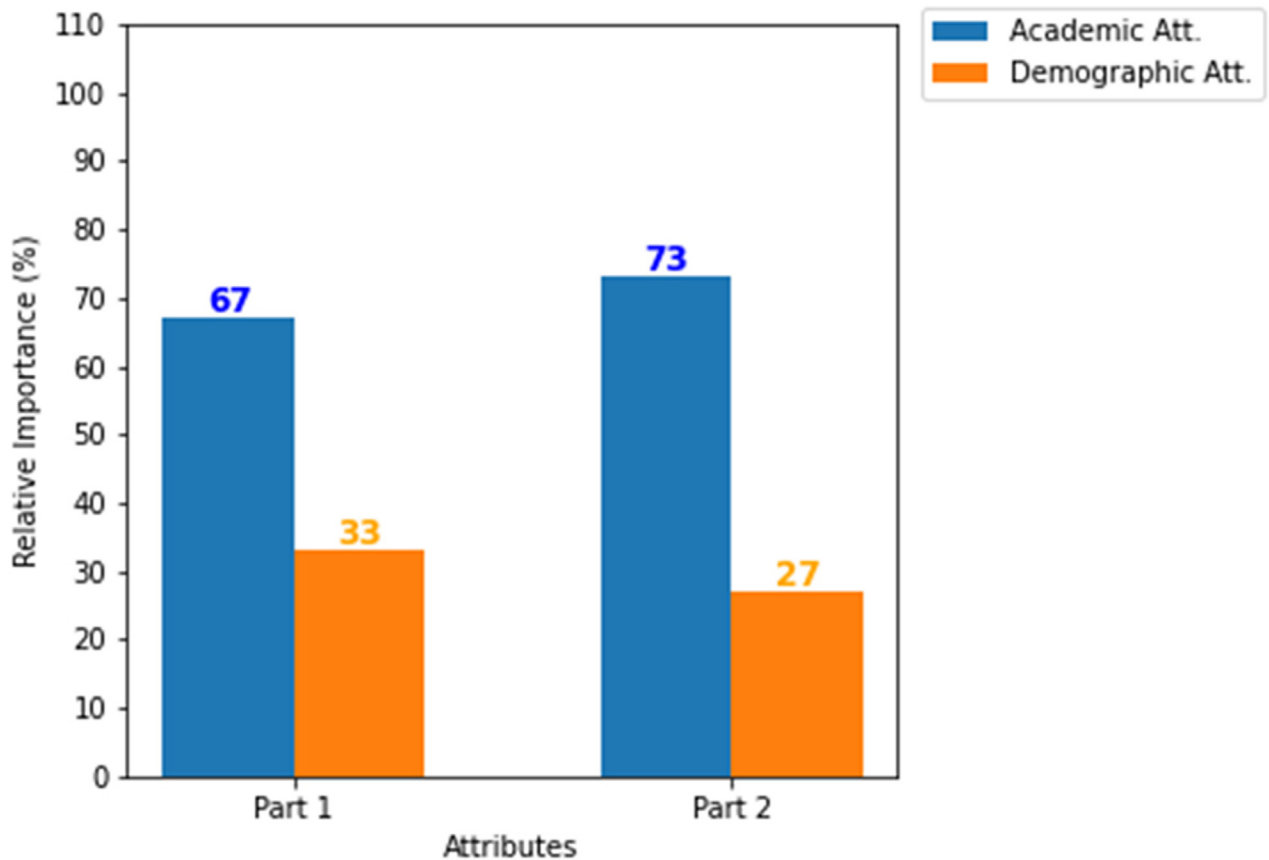


Fig. 3. Relative importance 2 metric: academic versus demographic attributes. All attributes were rescaled so that the sum of parts 1 and 2 is 100. Att. = attribute.

DISCUSSION

The aim of our multi-institutional DCE was to understand the relative emphasis residency selection committees place on academic and demographic variables and how that balance might change with the loss of objective academic

metrics. A loss of academic metrics might be expected to increase focus on diversity by default [18], but our data suggest a complex relationship in which focus could be diverted from efforts toward diversity in holistic reviews of applicants. In this study, residency programs prioritize

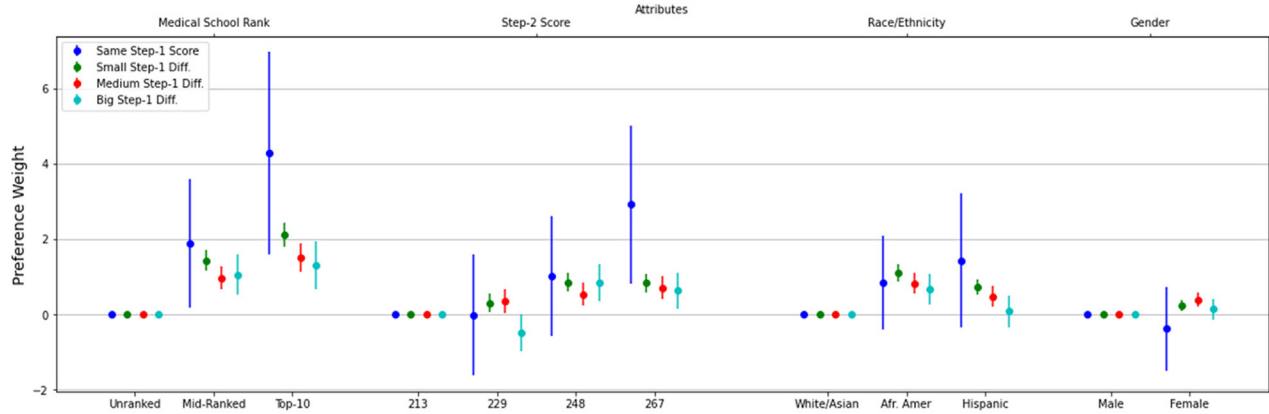


Fig. 4. Pairwise comparison of the coefficients for the school of medicine, Step 2 score, race/ethnicity, and gender by Step 1 score difference. The dots represent the value of the log odds derived from the conditional logit model, and the bar corresponds to the 95% confidence interval. The color represents the subsample used to estimate the conditional logit. Diff. = difference.

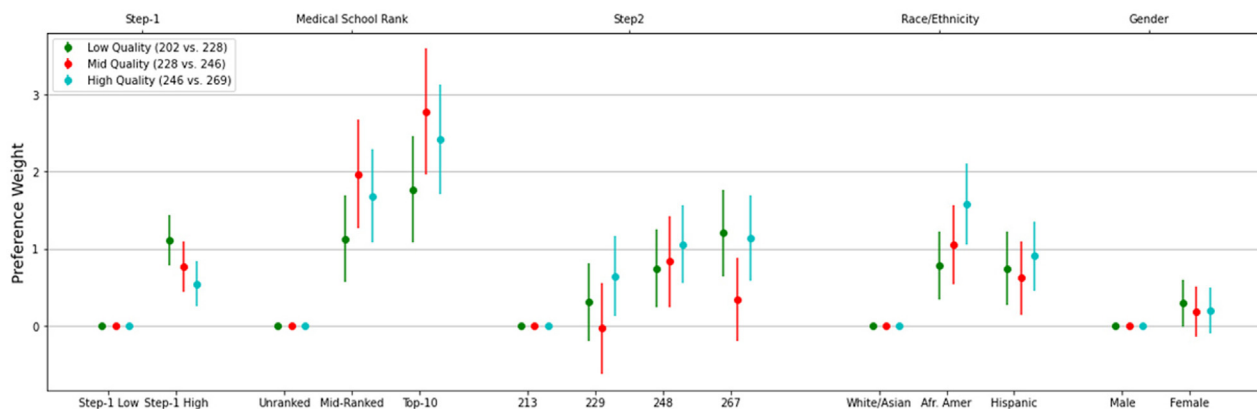


Fig. 5. Pairwise comparison of the coefficients for the Step 1 score, medical school rank, Step 2 score, race/ethnicity, and gender for the subgroup of paired applicants with only small Step 1 difference. The dots represent the value of the log odds derived from the conditional logit model and the bar corresponds to the 95% confidence interval. The color represents the subsample used to estimate the conditional logit.

diversity only above a threshold of academic quality, typically the USMLE Step 1 score, which is an efficient signal of academic quality and the first-order preference of participants. When the numerical Step 1 score was removed, evaluators sought alternative signals of academic quality through second-order academic metrics (USMLE Step 2 scores, medical school rank and prestige, clerkship honors, class rank, and publications). The search for alternative academic metrics had a transactional cost that deprioritized gender and race/ethnicity in resident selection. Specifically, the loss of the Step 1 score resulted in a 25% decrease in the relative importance of race/ethnicity, and a 29% decrease in the relative importance of gender, in residency selection decisions. Despite creating a culture of a holistic evaluation, the conversion to pass/fail decreases the numeric data

available to assess applicants holistically, which could paradoxically result in a less diverse pool of interviewees.

Moreover, our results suggest that the relative weight afforded to diversity in the selection process varies on the basis of absolute and relative differences in Step 1 scores between applicants. When comparing applicants with similar Step 1 scores, particularly if those scores are high, preference weights for race/ethnicity and gender are larger. When participants can easily discriminate the academic quality of applicants by Step 1 score differences, or if the Step 1 scores are low, preference rates for the second order academic metrics are larger. The belief that diversity is more influential in decisions when academic quality between two applicants is similar is reinforced by the analysis of the interaction of medical school rank and race/ethnicity, which

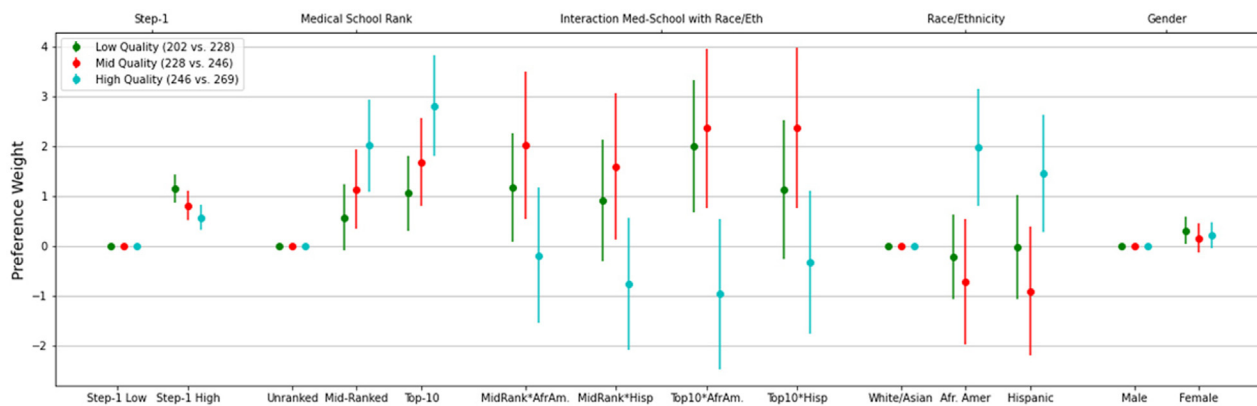


Fig. 6. Pairwise comparison of the coefficients for the Step 1 score, medical school rank, interaction between medical school rank and race/ethnicity, race/ethnicity, and gender for the subgroup of paired applicants with only small Step 1 difference. Each variable was dummy-coded with a reference class. The dots represent the value of the log odds derived from the conditional logit model, and the bar corresponds to the 95% confidence interval. The color represents the subsample used to estimate the conditional logit.

Table 5. Linear probability and logistic regression models validating selection attributes using interview invitations from a single radiology residency program

| Attribute | Linear Probability | | Logistic Regression | |
|---------------------|-------------------------|-------|-----------------------|-------|
| | Coefficient (95% CI) | P | Odds Ratio (95% CI) | P |
| Medical school rank | | | | |
| Unranked | Reference | | Reference | |
| Midlevel ranked | 0.07 (0.03 to 0.10) | <.001 | 5.80 (2.40 to 14.13) | <.001 |
| Top 10 | 0.30 (0.21 to 0.39) | <.001 | 27.43 (8.30 to 90.60) | <.001 |
| Gender | | | | |
| Male | Reference | | Reference | |
| Female | 0.38 (0.0008 to 0.76) | .045 | 2.21 (1.08 to 4.53) | .030 |
| Race/ethnicity | | | | |
| White or Asian | Reference | | Reference | |
| Black or Hispanic | 0.04 (0.0006 to 0.082) | .047 | 1.87 (1.01 to 3.43) | .043 |
| Step 1 (quartiles) | | | | |
| First | Reference | | Reference | |
| Second | 0.035 (−0.01 to 0.081) | .145 | 4.89 (1.01 to 23.78) | .049 |
| Third | 0.06 (0.01 to 0.10) | .023 | 7.81 (1.67 to 36.54) | .009 |
| Fourth | 0.13 (0.072 to 0.18) | <.001 | 15.62 (3.28 to 74.48) | .001 |
| Step 2 (quartiles) | | | | |
| First | Reference | | Reference | |
| Second | −0.003 (−0.05 to 0.043) | .90 | 1.53 (0.40 to 5.85) | .53 |
| Third | 0.014 (−0.04 to 0.067) | .60 | 2.14 (0.56 to 8.28) | .27 |
| Fourth | 0.038 (−0.017 to 0.095) | .17 | 2.79 (0.73 to 10.66) | .13 |

Note: CI = confidence interval.

demonstrated that race/ethnicity was significant only when applicants are in the highest Step 1 score group. Otherwise, when applicants are in the medium or lowest Step 1 score group, race/ethnicity was significant only when applicants attended a similarly ranked medical school (ie, the academic quality of applicants is similar). In other words, when applicants do not have high enough Step 1 scores, selection committees need to see an additional academic attribute to make sure two applicants have similar academic quality, and only then is race/ethnicity considered.

There is no perfect methodology to understand the trade-offs residency selection committees make when evaluating applicants. However, the DCE methodology appears to be well suited to the task and may be the best methodology available in an experimental setting. The validation of our findings using an external set of real-world data demonstrated significant associations with medical school rank, gender, race/ethnicity, and Step 1 score. Only Step 2 score was nonsignificant, but at the time of the 2021 application cycle, the Step 2 score was not universally required and not relied upon as strongly as was the Step 1 score. A limitation of the DCE, however, is that it does not provide a complete overview of an applicant, since there are a limited number of variables that can be tested (a maximum

of eight in the Sawtooth platform). This makes the DCE less valuable when a more complete profile of an applicant is desired, such as after an interview, when rank lists are being prepared. However, it is better suited for the more superficial screening that occurs for interview invitations when very large numbers of applications must be reviewed and a more limited number of factors can be used for decision making by evaluators.

The differing influence of race/ethnicity compared with gender is a notable finding of our study. The underrepresentation of women in radiology has been well publicized and is of long standing [19-21]. However, our study demonstrates that race/ethnicity is given more relative importance than gender with or without Step 1 numeric scores. Furthermore, the impact of race/ethnicity changes much more with both relative and absolute differences in Step 1 score. For example, there were notable changes in the preference weights for Black and Hispanic applicants on the basis of differences in Step 1 score pairs but minimal changes on the basis of gender, as shown in Figure 3. This implies that evaluators are more actively incorporating race/ethnicity into their decision-making process and that gender is an afterthought. This is at odds with the large body of radiology work focused on gender-based diversity and may

reflect the more recent national climate, which has focused more directly on race/ethnicity [2].

These findings have potentially major implications for applicants, residency programs, and policymakers. Because a diverse workforce can help foster creativity, innovation, productivity, and cultural competence [22], we suspect that diversity is a goal for all radiology residency programs. The continued erosion of academic metrics may have the unintended and paradoxical consequences of diminishing efforts toward racial/ethnic diversity. Although we used the Step 1 score as a surrogate for academic quality, other medical school performance metrics, such as grades, classmate summative assessments [7], and inductions into the Alpha Omega Alpha honor society [23] are being progressively eliminated, and there are early calls to eliminate a numeric score on the USMLE Step 2 examination [24]. Policymakers should consider the unintended consequences when the transparency of academic performance is lost. Radiology residency program directors and selection committees should be mindful of the careful balance of academic and demographic factors in their holistic review of residency applicants. Institutions that provide diversity and inclusion training for their selection committees should include implications of a loss of academic metrics in their deliberation.

There are limitations to this study. Individual attribute levels must be selected for the DCE, but they do not represent the full breadth and scope of a typical pool of residency applicants. The DCE is designed to compare pairs of items in order to statistically extract the relative contributions of specific attributes, whereas real-world residency selection instead compares a pool of applicants simultaneously. Nonetheless, this methodology has been used to provide insights into important decision-making processes in residency selection and other areas of medicine, and we have now shown the results correlate with real-world resident selection decisions of a single residency program. The value of Step 1 score may vary by medical specialty. Gatekeepers for radiology, an information-based specialty in which the step has been shown to predict successful completion of program [8], diagnostic accuracy on independent call [25], and success on board certification examinations [26], may value the Step 1 results more than other specialties.

TAKE-HOME POINTS

- Diversity, especially in race/ethnicity, is primarily valued only when applicants meet an academic quality threshold.

- Race/ethnicity is given greater priority than gender overall and is more influenced by changes to available academic metrics.
- Discrete-choice methodology was validated using real-world data on resident interview invitations.
- The erosion of long-trusted academic metrics may deemphasize diversity, especially without replacement objective measures of academic performance.

ACKNOWLEDGMENTS

The Radiology Residency Education Research Alliance includes the following members: Sabina Amin, MD, David Bader, MD, Brooke Beckett, MD, Kevin Carter, MD, Teresa Chapman, MD, Bernard Chow, MD, Amanda Derylo, MD, Francis Flaherty, MD, Michael Fox, MD, Jennifer Gould, MD, Robert Groves, MD, Darel Heitkamp, MD, John Heymann, MD, Christopher Ho, MD, Marion Hughes, MD, Nathan Hull, MD, Abtin Jafroodifar, MD, Ann Jay, MD, Aaron Kamer, MD, Hillary Kelly, MD, Tabassum Kennedy, MD, Emily Knippa, MD, Nicholas Koontz, MD, Mary Marx, MD, James Milburn, MD, Megan Mills, MD, Marco Molina, MD, Desiree Morgan, MD, Rustain Morgan, MD, Toma Omofoye, MD, Ryan Peterson, MD, Donald Romanelli, MD, Johanna Schubert, MD, Andrew Schweitzer, MD, Jayne Seekins, MD, John Stanfill, MD, Kara Udager, MD, Geogy Vatakencherry, MD, Morlie Wang, MD, Mandy Weidenhaft, MD, Clint Williamson, MD, Andrij Wojtowycz, MD, and Jessica Zarzour, MD.

ADDITIONAL RESOURCES

Additional resources can be found online at: <https://doi.org/10.1016/j.jacr.2023.07.019>.

REFERENCES

1. Mora H, Obayemi A, Holcomb K, Hinson M. The national deficit of black and Hispanic physicians in the US and projected estimates of time to correction. *JAMA Netw Open* 2022;5:e2215485.
2. Wu X, Bajaj S, Khunte M, et al. Diversity in radiology: current status and trends over the past decade. *Radiology* 2022;305:640-7.
3. DeBenedictis CM, Heitkamp DE, England E, et al. A program director's guide to cultivating diversity and inclusion in radiology residency recruitment. *Acad Radiol* 2020;27:864-7.
4. Weaver JS, Revels JW, Wang SS. Approaching diversity and inclusion in the radiology department. *Abdom Radiol (N Y)* 2021;46:5471-4.
5. Gupta S, Choe AI, Hardy PA, et al. Multilevel approach to support diversity, equity and inclusion in radiology. *Acad Radiol* 2023;30:952-8.
6. Marbin J, Rosenbluth G, Brim R, Cruz E, Martinez A, McNamara M. Improving diversity in pediatric residency selection: using an equity framework to implement holistic review. *J Grad Med Educ* 2021;13:195-200.
7. Association of American Medical Colleges. Grading systems used in medical school programs. Available at: <https://www.aamc.org/data>

reports/curriculum-reports/interactive-data/grading-systems-used-medical-school-programs. Accessed February 14, 2022.

8. Maxfield CM, Grimm LJ. The value of numerical USMLE Step 1 scores in radiology resident selection. *Acad Radiol* 2020;27:1475-80.
9. Patel MD, Tomblinson CM, Benefield T, et al. The relationship between US Medical Licensing Examination step scores and ABR core examination outcome and performance: a multi-institutional study. *J Am Coll Radiol* 2020;17:1037-45.
10. USMLE program announces upcoming policy change. Available at: <https://www.newswise.com/articles/usmle-program-announces-upcoming-policy-changes>. Accessed January 19, 2021.
11. National Resident Matching Program. Data Release and Research Committee. Results of the 2018 NRMP Program Director Survey. Available at: <https://www.nrmp.org/wp-content/uploads/2018/07/NRMP-2018-Program-Director-Survey-for-WWW.pdf>. Accessed January 19, 2021.
12. McDade W, Vela MB, Sanchez JP. Anticipating the impact of the USMLE Step 1 pass/fail scoring decision on underrepresented-in-medicine students. *Acad Med* 2020;95:1318-21.
13. Ryan M, Bate A, Eastmond CJ, Ludbrook A. Use of discrete choice experiments to elicit preferences. *Qual Health Care* 2001;10(suppl 1):i55-60.
14. Maxfield CM, Montano-Campos JF, Chapman T, et al. Factors influential in the selection of radiology residents in the post-Step 1 world: a discrete choice experiment. *J Am Coll Radiol* 2021;18:1572-80.
15. Maxfield CM, Heitkamp D, Chapman T, Koontz NA, Kohr JR, Grimm LJ. The Radiology Resident Education Research Alliance: the evolution of a multi-institutional research cooperative. *J Am Coll Radiol* 2022;19:586-9.
16. National Resident Matching Program. Data Release and Research Committee. Results of the 2018 NRMP Program Director Survey. Available at: <https://www.nrmp.org/wp-content/uploads/2018/07/NRMP-2018-Program-Director-Survey-for-WWW.pdf>. Accessed January 19, 2021.
17. 2020 best medical schools: research. Available at: <https://www.usnews.com/best-graduate-schools/top-medical-schools/research-rankings>. Accessed August 8, 2020.
18. Nehemiah A, Roberts SE, Song Y, et al. Looking beyond the numbers: increasing diversity and inclusion through holistic review in general surgery recruitment. *J Surg Educ* 2021;78:763-9.
19. Cater SW, Yoon SC, Lowell DA, et al. Bridging the gap: identifying global trends in gender disparity among the radiology physician workforce. *Acad Radiol* 2018;25:1052-61.
20. Campbell JC, Yoon SC, Cater SW, Grimm LJ. Factors influencing the gender breakdown of academic radiology residency programs. *J Am Coll Radiol* 2017;14:958-62.
21. Grimm LJ, Lowell DA, Cater SW, Yoon SC. Differential motivations for pursuing diagnostic radiology by gender: implications for residency recruitment. *Acad Radiol* 2017;24:1312-7.
22. Rosenkranz KM, Arora TK, Termuhlen PM, et al. Diversity, equity and inclusion in medicine: why it matters and how do we achieve it? *J Surg Educ* 2021;78:1058-65.
23. Lynch G, Holloway T, Muller D, Palermo AG. Suspending student selections to Alpha Omega Alpha Honor Medical Society: how one school is navigating the intersection of equity and wellness. *Acad Med* 2020;95:700-3.
24. Wilson CM, Brown NJ, Detchou DKE. Letter to the editor. USMLE examination and implications of a recent change. *J Neurosurg* 2021;136:316-7.
25. Agarwal V, Bump GM, Heller MT, et al. Do residency selection factors predict radiology resident performance? *Acad Radiol* 2018;25:397-402.
26. Horn GL Jr, Herrmann S, Masood I, Andersen CR, Nguyen QD. Predictors for failing the American Board of Radiology core examination. *AJR Am J Roentgenol* 2019;213:485-9.