Examining Demographics, Prior Academic Performance, and United States Medical Licensing Examination Scores

Jonathan D. Rubright, PhD, Michael Jodoin, PhD, and Michael A. Barone, MD, MPH

Abstract

Purpose
To examine whether demographic differences exist in United States Medical Licensing Examination (USMLE) scores and the extent to which any differences are explained by students' prior academic achievement.

Method
The authors completed hierarchical linear modeling of data for U.S. and Canadian allopathic and osteopathic medical graduates testing on USMLE Step 1 during or after 2010, and completing USMLE Step 3 by 2015. Main outcome measures were computer-based USMLE examinations: Step 1, Step 2 Clinical Knowledge, and Step 3. Test-taker characteristics included sex, self-identified race, U.S. citizenship status, English as a second language, and age at first Step 1 attempt. Covariates included composite Medical College Admission Test (MCAT) scores, undergraduate grade point average (GPA), and previous USMLE scores.

Results
A total of 45,154 examinees from 172 medical schools met the inclusion criteria. The sample was 67% white and 48% female; 3.7% non-U.S. citizens; and 7.4% with English as a second language. Hierarchical linear models examined demographic variables with and without covariates including MCAT scores and GPA. All Step examinations showed significant differences by gender after adding covariates, varying by Step. Racial differences were observed for each Step, attenuated by the addition of covariates.

Conclusions
Demographic differences in USMLE performance were tempered by previous examination performance and undergraduate performance. Additional research is required to identify factors that contribute to demographic differences, can aid educators' identification of students who would benefit from assistance preparing for USMLE, and can assist residency program directors in assessing performance measures while meeting diversity goals.

The United States Medical Licensing Examination (USMLE) has a mission to protect the health of the public. Passing these examinations is required for U.S. states and territories to consider granting an unrestricted medical license to a physician. USMLE comprises three Steps (four exams). Step 1 is a multiple-choice examination assessing an examinee's knowledge of foundational science concepts applicable to medicine. Step 2 Clinical Knowledge (CK) assesses the ability to apply scientific concepts to clinical medicine. Step 2 Clinical Skills (CS) uses standardized patients to test the examinee's ability to gather information from patients, perform physical examinations, and communicate findings to patients and colleagues. Step 3 uses multiple-choice questions and computerized patient cases to assess an examinee's ability to practice in an unsupervised setting.

This examination series may represent a barrier to practice for certain aspiring physicians. A rich body of research exists for the USMLE, including research on demographic differences in USMLE scores. A number of subgroups have been examined, including analyses grouped by sex and self-identified race. These previous studies have examined total reported scores with a focus on secondary use, such as postgraduate residency screening and selection.

Examining differences by sex on the precursor to the current USMLE Step 1, the National Board of Medical Examiners (NBME) Part I examination, Case and colleagues[1] found that men performed better than women on average by about 0.3 standard deviations (SDs). This difference was at least partly explained by covariates such as Medical College Admission Test (MCAT) scores, undergraduate grade point average (GPA), and college selectivity. This finding has been replicated.[2] A later study analyzing Step 1 scores showed a similar pattern of men performing better than women, even after controlling for covariates.[3] Analyses on NBME Part II, the precursor to Step 2 CK, showed women performing as well as or better than men.[1] This effect was again seen using the current Step 2 format, showing women moderately outperforming men on Step 2 CS and CK.[4–6]

Comparably less research has been performed on racial differences in USMLE scores. Our literature search identified only one study, using data from the older Part I format. That analysis showed racial differences wherein white students performed highest among self-identified racial groups, followed by Asian/Pacific Islanders, Hispanics, then blacks. Controlling for the MCAT, undergraduate GPA, and college selectivity reduced, but did not eliminate, differences.[2]

USMLE Step 3 scores have received less attention than Steps 1 or 2. Successfully passing Step 3 was most associated with being a native English-speaking U.S. citizen from a U.S. school. Although sex appeared statistically significant, with men outperforming women, the practical significance was small.[7] Together, previous work suggests that men outperform

Please see the end of this article for information about the authors.

Correspondence should be addressed to Jonathan D. Rubright, National Board of Medical Examiners, 3750 Market St., Philadelphia, PA 19104; telephone: (215) 590-9885; e-mail: jrubright@nbme.org.

Acad Med. 2019;94:364–370.
First published online July 17, 2018
doi: 10.1097/ACM.0000000000002366

women on Step 1, yet the trend is reversed for Step 2 and negligible for Step 3. Some racial differences have also been seen, albeit from a study using older data on a test format no longer used.

These studies have told a story of average demographic differences across the USMLE series. Yet the story goes back 24 years, spans outdated test formats, examines demographic characteristics individually, and uses a variety of methodological approaches. To provide data on possible subgroup performance differences, this study examines many demographic characteristics of interest simultaneously within one modeling framework, under the current Step testing format, for all computer-based USMLE Step exams. Current information on subgroup performance differences may inform how accreditation organizations, medical schools, and postgraduate training programs use USMLE data above and beyond the primary intended use of assessing passing scores for medical licensure.

Method

Design, sample, and data collection

We used a cross-sectional analysis of historical, deidentified data. Ethical approval with "exempt" status was granted by the American Institutes for Research, Washington, DC. Examinees' first-time scores for Step 1, Step 2 CK, and Step 3 were included if the examinee took Step 1 during or after 2010, completed Step 3 by 2015, and reported demographic information. As our research was intended to address secondary use of scores, we sampled examinees who had progressed through the examination series and taken each of the computerized Steps. To focus on results from U.S. and Canadian allopathic and osteopathic medical schools, we did not include international medical graduates in this analysis.

Measurements

Dependent variables were scores on computer-based USMLE Step examinations: Step 1, Step 2 CK, and Step 3. Test-taker characteristics were self-reported on the application to sit for the first USMLE examination, and included sex (male as reference category), race (self-identified: Asian/Pacific Islander; black not of Hispanic origin; and Hispanic, with white as reference category), U.S. citizenship status (U.S. citizen as reference category), English as a second language (ESL) (native English speaker as reference category), and age at first Step 1 attempt (grand mean centered). Composite MCAT scores (from first take, grand mean centered) and undergraduate GPA (grand mean centered) were obtained from the Association of American Medical Colleges (AAMC). The MCAT composite included the verbal reasoning, biological sciences, and physical sciences sections and excluded the writing sample, as the former sections have been shown to be related to USMLE scores and one another while the latter section has not.[3] We did not include racial categories with too few examinees (American Indian/Alaskan Native, n = 175), nor from the categories "do not wish to respond," "multiple," or "other." Examinees were included if they agreed to allow their deidentified data to be used for research purposes.

Data analysis

Hierarchical linear modeling (HLM)[8] has been used previously in this line of research, with most score variance within, not between, schools[9] or cases.[5] Still, HLM is more appropriate in datasets with a nested structure. Medical students were nested within medical schools for this analysis performed using SAS statistical software, version 9.3 (SAS Institute Inc., Cary, North Carolina) with maximum likelihood estimation. Multicollinearity among predictors is not a concern here because variables likely to be correlated are used as control variables and not variables of interest. Additionally, centering of variables is used to aid in the interpretation of the resulting coefficients, and has the secondary benefit of reducing the relationships among the variables under study.

First, we produced descriptive statistics for all included variables. Principally interested in how examinee characteristics predicted USMLE performance and not in how these relationships varied by school, we estimated random intercept models allowing schools to have different intercepts but not slopes. This decision was driven by our interest in overall demographic effects and also by small sample sizes from school-level clusters. These models constrain the relationships between demographic characteristics and USMLE performance to remain the same across schools, although school intercepts may vary.

Because the research questions were to understand demographic differences among scores and whether covariates attenuated these differences, model building was guided by the research questions. We ran the following models with Step 1, Step 2 CK, and then Step 3 as the dependent variable:

- An unconditional model to calculate the intraclass correlation (ICC), which is the ratio of between-to-total variance. This value tells us the proportion of variance attributable to clustering at the medical school level.
- A random intercept model using the demographic characteristics U.S. citizenship, self-identified racial category, ESL status, sex, and age at first Step 1 attempt. Here, this will be referred to as the demographics model.
- A random intercept model including the variables above, along with GPA and MCAT score as covariates, to assess whether demographic relationships associated with USMLE performance are attenuated. Here, this will be referred to as the covariates model. With Step 2 CK scores as the dependent variable, Step 1 was entered in the covariates model grand mean centered. With Step 3 scores as the dependent variable, both Step 1 and Step 2 CK were added grand mean centered.

For the dichotomous variables in all models, we generated an effect size measure along with each coefficient. Because coefficients are interpretable in terms of USMLE score points, and all Step examinations are scaled to a base reference group with an SD of 20 points, the effect size used was the coefficient divided by 20 and is interpretable as differences in SD units. Cohen suggested that an effect size in SD units could be considered small if ≥ 0.2 yet < 0.5, medium if ≥ 0.5 yet < 0.8, and large if ≥ 0.8.[10] We provided effect sizes because, given the sample size we used, statistical significance is likely.

Results

A total of 45,154 examinees from 172 schools fit study criteria (average

262.52 examinees per school, SD 190.27, range 1–820). Table 1 shows descriptive statistics for the sample. Tables 2, 3, and 4 sequentially show the modeling results with USMLE Steps 1, 2, and 3 as the dependent variable. The ICC for predicting Step 1 scores is 0.12. Therefore, 88% of the variance in scores was due to student differences. Examining Step 1 results in Table 2, the intercept for the demographics model is the predicted performance when all demographic variables represent the reference category—that is, for a native English-speaking white male U.S. citizen at average age. The coefficients are interpreted as the difference in predicted Step 1 scores compared with the reference group with all others constant. Thus, a female ESL test taker, or any nonwhite test taker, would be predicted to have a lower Step 1 score. Similarly, scores are predicted to be lower for each year of age above average. Being a non-U.S. citizen would increase the predicted score.

Adding GPA and MCAT score to arrive at the covariates model (penultimate column of Table 2) improved predictions of Step 1 scores, as shown by the lower error variance at both levels along with improved fit indices (−2 log likelihood, Akaike information criterion and Bayesian information criterion). Because the added covariates were grand mean centered, the intercept is now interpreted as the predicted Step 1 performance of a test taker with the demographic characteristics described above who is also of average GPA and MCAT score. For every 1-point increase in GPA above the average value, predicted Step 1 performance increased by 11.91 points. Predicted scores also increased if an individual had above-average composite MCAT performance. After including these variables, the variables representing U.S. citizenship and ESL status were no longer significant. That is, these demographic differences were explained by differences in GPA and MCAT scores. The coefficients for black or Hispanic test takers were attenuated, although the Asian coefficient remained similar.

The ICC for Step 2 CK is similar to that of Step 1: 0.10. Table 3 displays results with Step 2 CK scores as the dependent variable; all demographic variables under study were statistically significant. The intercept retained the same interpretation as that of the Step 1 demographics model, albeit for the prediction of Step 2 CK scores. All demographic variables alter the prediction of Step 2 CK performance in the same direction as the Step 1 model, except for sex. Similar to previous studies of Step 2 performance, we found that women were predicted to have higher performance than men (by 0.34 points). Adding covariates again improved the model as shown by the decrease in error variance and fit indices. The demographic variable coefficients again changed under this model, with the impact of sex increased and U.S. citizenship status no longer a significant model predictor. Individuals with above-average GPA, composite MCAT, and Step 1 scores were predicted to have higher performance, while those with above-average age were predicted to be lower. And, the addition of the GPA and MCAT covariates again attenuated differences for Asian, black, Hispanic, and ESL examinees.

The ICC for Step 3 is 0.12. Lastly, Table 4 reports the parameters for the prediction of USMLE Step 3 performance. The direction and magnitude of the demographic variables were similar to those from Tables 2 and 3, except for sex, which is nonsignificant. Adding covariates to the model again aided in the prediction of Step 3 scores, with higher levels of Step 1, Step 2 CK, GPA, and composite MCAT increasing the prediction of Step 3 performance and higher age decreasing the predicted score. With added covariates, U.S. citizenship was no longer significant; racial and ESL indicators are attenuated when covariates were included.

Discussion

This study extends and updates previous analyses by using the modern USMLE Step format, examining the impact of all self-reported examinee characteristics simultaneously across all computerized Steps, and examining the impact of important premedical school covariates. Our findings show that, on average, demographic differences exist in USMLE scores. In the nonadjusted models, sex effects were present, although they varied depending on the Step under consideration. Men outperformed women on Step 1, women outperformed men on Step 2, and there was no difference on Step 3. ESL test takers and self-identified nonwhite groups consistently performed lower on all three Steps; although their practical significance varies, the size of the coefficients remained similar across Steps. Citizenship and ESL status showed statistical, yet not practical, significance. Age consistently showed a negative relationship with Step scores, with examinees above average age predicted to have lower scores.

Another consistent finding emerged: Adding covariates on a test taker's

Table 1

Descriptive Statistics for 45,154 Examinees From 172 Medical Schools,[a] From a Study of Demographic Differences in USMLE Scores, 2010–2015

Variable	Value
Step 1 score (first attempt), mean ± SD (range)	228.13 ± 20.60 (131–280)
Step 2 score (first attempt), mean ± SD (range)	240.60 ± 18.20 (159–288)
Step 3 score (first attempt), mean ± SD (range)	223.75 ± 15.67 (146–273)
Total GPA, mean ± SD (range)	3.67 ± 0.26 (1.89–4)
Total MCAT score, mean ± SD (range)	29.57 ± 4.84 (8–44)
Age at first Step 1 attempt, mean ± SD (range)	25.35 ± 2.59 (13–61)
Step 2 CS pass, no. (%)	44,070 (97.60)
Non-U.S. citizen, no. (%)	1,656 (3.67)
Asian/Pacific Islander, no. (%)	9,365 (20.74)
Black (not of Hispanic origin), no. (%)	2,780 (6.16)
Hispanic, no. (%)	2,918 (6.46)
White (not of Hispanic origin), no. (%)	30,091 (66.64)
ESL, no. (%)	3,348 (7.41)
Female, no. (%)	21,725 (48.11)

Abbreviations: USMLE indicates United States Medical Licensing Examination; SD, standard deviation; GPA, grade point average; MCAT, Medical College Admission Test; CS, clinical skills; ESL, English as a second language.
[a]Average 262.52 examinees per school, standard deviation 190.27, range 1–820.

Table 2

Results for Predicting First-Time USMLE Step 1 Performance Using a Demographics-Only Model and Fully Adjusted Model, From a Study of Demographic Differences in USMLE Scores, 2010–2015[a]

	Demographics model			Covariates model		
Characteristic	Coefficient	95% CI	Effect size[b]	Coefficient	95% CI	Effect size[b]
Intercept	233.17	232.06 to 234.28[c]	—	230.86	230.20 to 231.51[c]	—
Non-U.S. citizen	1.78	0.80 to 2.76[c]	0.09	−0.42	−1.34 to 0.50	−0.02
Asian	−4.45	−4.91 to −3.98[c]	−0.22	−3.96	−4.40 to −3.52[c]	−0.20
Black	−16.52	−17.32 to −15.72[c]	−0.83	−5.10	−5.90 to −4.29[c]	−0.26
Hispanic	−12.10	−12.90 to −11.29[c]	−0.61	−4.79	−5.57 to −4.01[c]	−0.24
ESL	−1.43	−2.16 to −0.71[c]	−0.07	−0.14	−0.82 to 0.55	−0.01
Female	−5.92	−6.27 to −5.57[c]	−0.30	−4.07	−4.40 to −3.73[c]	−0.20
Age at Step 1 attempt	−1.23	−1.29 to −1.16[c]	—	−0.58	−0.65 to −0.51[c]	—
Total GPA	—	—	—	11.91	11.16 to 12.66[c]	—
Total MCAT	—	—	—	1.49	1.44 to 1.53[c]	—
			Estimate (SE)			Estimate (SE)
Error variance						
Level 1			350.53 (2.34)[c]			312.29 (2.08)[c]
Level 2 intercept			43.90 (5.49)[c]			12.73 (1.83)[c]
Model fit						
−2 log likelihood			393,232.5			387,861.2
AIC			393,252.5			387,885.2
BIC			393,283.9			387,923.0

Abbreviations: USMLE indicates United States Medical Licensing Examination; CI, confidence interval; ESL, English as a second language; GPA, grade point average; MCAT, Medical College Admission Test; SE, standard error; AIC, Akaike information criterion; BIC, Bayesian information criterion.

[a]Intraclass correlation coefficient = 0.12.

[b]Reported for dichotomous variables only.

[c]$P < .001$.

previous examination and undergraduate performance increases the accuracy of prediction and, with the exception of sex, substantially reduces the predicted effects of demographic characteristics. In some cases, the effects of citizenship and ESL status were erased entirely. In others, the effects were attenuated. For example, self-identified blacks were predicted to score 16 points lower on all Step examinations compared with whites in the demographics-only model, representing more than three-fourths of an SD. When additional premedical school covariates were included, these differences were reduced to 4 or 5 points, around one-quarter of an SD. More than 10 points of a black test taker's predicted performance were explained by covariates.

There are limitations to this study. First, although our analysis aimed at understanding individual characteristics and their association with USMLE performance, 10% to 12% of score performance remains to be explained by medical school characteristics. Medical schools have different ways of supporting students through their curricula, and different policies concerning whether students need to take USMLE Steps for promotion or graduation (see, for example, https://www.aamc.org/initiatives/cir/406442/10b.html). Measuring and understanding how schools contribute to examination performance across demographic groups could be useful in understanding examinee performance and may further attenuate the demographic effects seen here. Second, additional aspects of training, included self-selected specialties, also have been shown to affect USMLE performance[11] yet are not considered here. Third, undergraduate institutions vary in their grading standards, which affects the comparability of GPAs for individuals across institutions. Fourth, this analysis only examines the computer-based USMLE Step exams; comparable analyses for Step 2 CS are planned.

Implications of these findings are relevant to two increasingly important concerns in medicine and medical education: the use of a score, on an examination intended for medical licensure, as a high-stakes screen or selection criterion for residency selection; and the recruitment and retention of a diverse physician workforce.

It is widely accepted that residency program directors, with the daunting task of screening numerous applications, use USMLE scores to screen applicants for interviews.[12,13] Furthermore, this practice has been associated in the past with potential bias against certain racial and ethnic minorities.[14] If applicants do not meet this screen, they are no longer considered despite their potentially having qualities or experiences that translate to becoming effective physicians. More recently, there has been a consistent message from leaders in the academic community as well as from the NBME to reduce or eliminate the use of USMLE scores, particularly

Table 3

Results for Predicting First-Time USMLE Step 2 CK Performance Using a Demographics-Only Model and Fully Adjusted Model, From a Study of Demographic Differences in USMLE Scores, 2010–2015[a]

	Demographics model			Covariates model		
Characteristic	Coefficient	95% CI	Effect size[b]	Coefficient	95% CI	Effect size[b]
Intercept	243.33	242.48 to 244.18[c]	—	239.60	239.14 to 240.07[c]	—
Non-U.S. citizen	1.05	0.18 to 1.92[d]	0.05	−0.41	−1.03 to 0.22	−0.02
Asian	−6.77	−7.18 to −6.35[c]	−0.34	−4.02	−4.32 to −3.72[c]	−0.20
Black	−15.97	−16.68 to −15.26[c]	−0.80	−4.04	−4.59 to −3.49[c]	−0.20
Hispanic	−10.55	−11.27 to −9.84[c]	−0.53	−1.94	−2.47 to −1.42[c]	−0.10
ESL	−2.19	−2.84 to −1.54[c]	−0.11	−1.11	−1.58 to −0.65[c]	−0.06
Female	0.34	0.03 to 0.66[d]	0.02	4.20	3.97 to 4.43[c]	0.21
Age at Step 1 attempt	−1.26	−1.33 to −1.20[c]	—	−0.40	−0.45 to −0.35[c]	—
Total GPA	—	—	—	2.53	2.02 to 3.05[c]	—
Total MCAT	—	—	—	0.26	0.23 to 0.29[c]	—
Step 1 (centered)	—	—	—	0.60	0.59 to 0.61[c]	—
			Estimate (SE)			Estimate (SE)
Error variance						
Level 1			279.14 (1.86)[c]			142.99 (0.95)[c]
Level 2 intercept			24.54 (3.14)[c]			6.64 (0.87)[c]
Model fit						
−2 log likelihood			382,899.1			352,605.6
AIC			382,919.1			352,631.6
BIC			382,950.6			352,672.5

Abbreviations: USMLE indicates United States Medical Licensing Examination; CK, Clinical Knowledge; CI, confidence interval; ESL, English as a second language; GPA, grade point average; MCAT, Medical College Admission Test; SE, standard error; AIC, Akaike information criterion; BIC, Bayesian information criterion.

[a]Intraclass correlation coefficient = 0.10.

[b]Reported for dichotomous variables only.

[c]$P < .001$.

[d]$P < .05$.

Step 1, as a barrier to residency selection.[15,16] These calls acknowledge the mission of the USMLE program, and point to evidence where USMLE scores can be predictive of performance on subsequent assessments, such as specialty in-training and certification examinations.[17] Relationships have also been demonstrated between scores on subcomponents of the USMLE and residency program director performance ratings, as well as for scores on certain USMLE Steps and disciplinary action in practice.[18–20] While research is ongoing regarding the predictive value of licensing examinations on clinical practice measures,[21] the debate remains over the evidence, or lack thereof, for using USMLE scores as a threshold for residency candidate consideration.[22] Some investigators have reported that, despite consistently lower scores on the USMLE obtained by underrepresented minority residents, no difference existed in observed structured clinical examinations at the start of residency.[23]

In 2015, black medical students comprised less than 6% of medical school graduates in the United States, and Latinos less than 5%.[24] Over the past 10 years, the AAMC's Holistic Review initiative has provided guidance and resources for medical admissions programs to "widen the lens" when viewing prospective candidates, emphasizing the applicants' experiences and personal attributes, in addition to their academic metrics.[25] An admissions process that focuses on mission-based initiatives is likely to produce diverse students, viewpoints, experiences, and ultimately a workforce reflecting the same. The concept of holistic review has carried into graduate medical education, particularly given the need for program directors to assess professionalism and communication competencies during the brief selection season, as well as the priority that graduate medical education programs are placing on recruiting and retaining diverse cohorts of trainees.[26,27] Given our findings, residency program directors may be able to more effectively engage in holistic review of applicants, and may also be motivated to provide additional resources to trainees in need of support for success on licensure and certification examinations. Some health professions education programs have demonstrated the effectiveness that targeted resources or mentoring may have on standardized test scores.[28] Furthermore, it would be important to consider how traditional program evaluation metrics—such as certifying board pass rates—might hinder efforts to advance diversity in medicine across specialties.[29]

Subgroup examinee performance on standardized tests need not be equal for a test to meet the standard of fairness.[30] In the case of our study, as in one previous study,[2] prior academic performance explains much of the demographic differences in scores. Although mean performance

Table 4

Results for Predicting First-Time USMLE Step 3 Performance Using a Demographics-Only Model and Fully Adjusted Model, From a Study of Demographic Differences in USMLE Scores, 2010–2015[a]

	Demographics model			Covariates model		
Characteristic	Coefficient	95% CI	Effect size[b]	Coefficient	95% CI	Effect size[b]
Intercept	226.76	225.99 to 227.53[c]	—	223.79	223.49 to 224.09[c]	—
Non-U.S. citizen	1.37	0.63 to 2.10[c]	0.07	0.23	−0.30 to 0.76	0.01
Asian	−6.79	−7.14 to −6.44[c]	−0.34	−3.22	−3.48 to −2.97[c]	−0.16
Black	−15.94	−16.54 to −15.34[c]	−0.80	−3.73	−4.20 to −3.27[c]	−0.19
Hispanic	−9.18	−9.79 to −8.58[c]	−0.46	−1.04	−1.49 to −0.59[c]	−0.05
ESL	−2.64	−3.18 to −2.09[c]	−0.13	−1.06	−1.45 to −0.66[c]	−0.05
Female	0.05	−0.21 to 0.31	0.00	1.19	1.00 to 1.39[c]	0.06
Age at Step 1 attempt	−0.95	−1.00 to −0.90[c]	—	−0.08	−0.12 to −0.04[c]	—
Total GPA	—	—	—	2.48	2.04 to 2.92[c]	—
Total MCAT	—	—	—	0.49	0.47 to 0.52[c]	—
Step 1 (centered)	—	—	—	0.11	0.10 to 0.11[c]	—
Step 2 CK (centered)	—	—	—	0.45	0.44 to 0.46[c]	—
			Estimate (SE)			Estimate (SE)
Error variance						
Level 1			199.02 (1.33)[c]			103.71 (0.69)[c]
Level 2 intercept			20.86 (2.56)[c]			2.25 (0.32)[c]
Model fit						
−2 log likelihood			367,646.7			338,004.7
AIC			367,666.7			338,032.7
BIC			367,698.2			338,076.8

Abbreviations: USMLE indicates United States Medical Licensing Examination; CI, confidence interval; ESL, English as a second language; GPA, grade point average; MCAT, Medical College Admission Test; CK, Clinical Knowledge; SE, standard error; AIC, Akaike information criterion; BIC, Bayesian information criterion.

[a]Intraclass correlation coefficient = 0.12.

[b]Reported for dichotomous variables only.

[c]$P < .001$.

between racial categories, especially for blacks and Hispanics, appears initially large, "the observed racial and ethnic differences reflect the lower mean MCAT scores and GPAs of underrepresented minority students."[2(p678)] And, MCAT scores themselves have not shown evidence of bias against underrepresented minority test takers.[31] As the remaining performance differences are unexplained, additional work is required to identify factors contributing to the remaining demographic differences and identify factors that can aid medical educators in identifying candidate examinees who may need additional help with USMLE preparation.

Acknowledgments: The authors thank Monica Cuddy and Kimberly Swygert for their valuable comments on early drafts of this manuscript.

Funding/Support: None reported.

Other disclosures: Drs. Rubright, Jodoin, and Barone are employed by the National Board of Medical Examiners.

Ethical approval: Institutional review board approval with "exempt" status granted by American Institutes for Research, Washington, DC.

J.D. Rubright is senior psychometrician, National Board of Medical Examiners, Philadelphia, Pennsylvania.

M. Jodoin is vice president of psychometrics and data analysis, National Board of Medical Examiners, Philadelphia, Pennsylvania.

M.A. Barone is vice president of licensure, National Board of Medical Examiners, Philadelphia, Pennsylvania.

References

1 Case SM, Becker DF, Swanson DB. Performances of men and women on NBME Part I and Part II: The more things change. Acad Med. 1993;68(10 suppl):S25–S27.
2 Dawson B, Iwamoto CK, Ross LP, Nungester RJ, Swanson DB, Volle RL. Performance on the National Board of Medical Examiners. Part I examination by men and women of different race and ethnicity. JAMA. 1994;272:674–679.
3 Cuddy MM, Swanson DB, Clauser BE. A multilevel analysis of examinee gender and USMLE Step 1 performance. Acad Med. 2008;83(10 suppl):S58–S62.
4 Cuddy MM, Swygert KA, Swanson DB, Jobe AC. A multilevel analysis of examinee gender, standardized patient gender, and United States medical licensing examination Step 2 clinical skills communication and interpersonal skills scores. Acad Med. 2011;86(10 suppl):S17–S20.
5 Swygert KA, Cuddy MM, van Zanten M, Haist SA, Jobe AC. Gender differences in examinee performance on the Step 2 Clinical Skills data gathering (DG) and patient note (PN) components. Adv Health Sci Educ Theory Pract. 2012;17:557–571.
6 Cuddy MM, Swanson DB, Clauser BE. A multilevel analysis of the relationships between examinee gender and United States Medical Licensing Exam (USMLE) Step 2 CK content area performance. Acad Med. 2007;82(10 suppl):S89–S93.
7 De Champlain A, Sample L, Dillon GF, Boulet JR. Modeling longitudinal performances on the United States Medical Licensing Examination and the impact of sociodemographic covariates: An application of survival data analysis. Acad Med. 2006;81(10 suppl):S108–S111.
8 Raudenbush SW, Bryk AS. Hierarchical Linear Models: Applications and Data

Analysis Methods. 2nd ed. Newbury Park, CA: Sage; 2002.
9 Cuddy MM, Swanson DB, Dillon GF, Holtman MC, Clauser BE. A multilevel analysis of the relationships between selected examinee characteristics and United States Medical Licensing Examination Step 2 Clinical Knowledge performance: Revisiting old findings and asking new questions. Acad Med. 2006;81(10 suppl):S103–S107.
10 Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
11 Sawhill AJ, Dillon GF, Ripkey DR, Hawkins RE, Swanson DB. The impact of postgraduate training and timing on USMLE Step 3 performance. Acad Med. 2003;78(10 suppl):S10–S12.
12 Green M, Jones P, Thomas JX Jr. Selection criteria for residency: Results of a national program directors survey. Acad Med. 2009;84:362–367.
13 National Resident Matching Program. Data Release and Research Committee. Results of the 2016 NRMP Program Director Survey. Washington, DC: National Resident Matching Program; 2016.
14 Edmond MB, Deschenes JL, Eckler M, Wenzel RP. Racial bias in using USMLE Step 1 scores to grant internal medicine residency interviews. Acad Med. 2001;76:1253–1256.
15 Prober CG, Kolars JC, First LR, Melnick DE. A plea to reassess the role of United States Medical Licensing Examination Step 1 scores in residency selection. Acad Med. 2016;91:12–15.
16 Katsufrakis PJ, Uhler TA, Jones LD. The residency application process: Pursuing improved outcomes through better understanding of the issues. Acad Med. 2016;91:1483–1487.
17 Dillon GF, Swanson DB, McClintock JC, Gravlee GP. The relationship between the American Board of Anesthesiology Part 1 certification examination and the United States Medical Licensing Examination. J Grad Med Educ. 2013;5:276–283.
18 Cuddy MM, Winward ML, Johnston MM, Lipner RS, Clauser BE. Evaluating validity evidence for USMLE Step 2 Clinical Skills data gathering and data interpretation scores: Does performance predict history-taking and physical examination ratings for first-year internal medicine residents? Acad Med. 2016;91:133–139.
19 Winward ML, Lipner RS, Johnston MM, Cuddy MM, Clauser BE. The relationship between communication scores from the USMLE Step 2 Clinical Skills examination and communication ratings for first-year internal medicine residents. Acad Med. 2013;88:693–698.
20 Cuddy MM, Young A, Gelman A, et al. Exploring the relationships between USMLE performance and disciplinary action in practice: A validity study of score inferences from a licensure examination. Acad Med. 2017;92:1780–1785.
21 Tamblyn R, Abrahamowicz M, Dauphinee WD, et al. Association between licensure examination scores and practice in primary care. JAMA. 2002;288:3019–3026.
22 McGaghie WC, Cohen ER, Wayne DB. Are United States Medical Licensing Exam Step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? Acad Med. 2011;86:48–52.
23 Lypson ML, Ross PT, Hamstra SJ, Haftel HM, Gruppen LD, Colletti LM. Evidence for increasing diversity in graduate medical education: The competence of underrepresented minority residents measured by an intern objective structured clinical examination. J Grad Med Educ. 2010;2:354–359.
24 Association of American Medical Colleges. Diversity in medical education: AAMC facts and figures 2016. http://www.aamcdiversityfactsandfigures2016.org. Accessed June 6, 2018.
25 Association of American Medical Colleges. Holistic admissions. https://www.aamc.org/initiatives/holisticreview/about. Accessed June 6, 2018.
26 King A, Mayer C, Starnes A, Barringer K, Beier L, Sule H. Using the Association of American Medical Colleges standardized video interview in a holistic residency application review. Cureus. 2017;9:e1913.
27 Van Voorhees AS, Enos CW. Diversity in dermatology residency programs. J Investig Dermatol Symp Proc. 2017;18:S46–S49.
28 Girotti JA, Park YS, Tekian A. Ensuring a fair and equitable selection of students to serve society's health care needs. Med Educ. 2015;49:84–92.
29 Berger JS, Cioletti A. Viewpoint from 2 graduate medical education deans: Application overload in the residency Match process. J Grad Med Educ. 2016;8:317–321.
30 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association; 2014.
31 Davis D, Dorsey JK, Franks RD, Sackett PR, Searcy CA, Zhao X. Do racial and ethnic group differences in performance on the MCAT exam reflect test bias? Acad Med. 2013;88:593–602.