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

Aimee K. Gardner, PhD, Katelyn J. Cavanaugh, PhD, Ross E. Willis, PhD, and Brian J. Dunkin

Abstract

**Purpose**
Use of the United States Medical Licensing Examination (USMLE) for residency selection has been criticized for its inability to predict clinical performance and potential bias against underrepresented minorities (URMs). This study explored the impact of altering traditional USMLE cutoffs and adopting more evidence-based applicant screening tools on inclusion of URMs in the surgical residency selection process.

**Method**
Multimethod job analyses were conducted at 7 U.S. general surgical residency programs during the 2018–2019 application cycle to gather validity evidence for developing selection assessments. Unique situational judgment tests (SJTs) and scoring algorithms were created to assess applicant competencies and fit. Programs lowered their traditional USMLE Step 1 cutoffs and invited candidates to take their unique SJT. URM status (woman, racial/ethnic minority) of candidates who would have been considered for interview using traditional USMLE Step 1 cutoffs was compared with the candidate pool considered based on SJT performance.

**Results**
A total of 2,742 general surgery applicants were invited to take an online SJT by at least 1 of the 7 programs. Approximately 35% of applicants who were invited to take the SJT would not have met traditional USMLE Step 1 cutoffs. Comparison of USMLE-driven versus SJT-driven assessment results demonstrated statistically different percentages of URMs recommended, and including the SJT allowed an average of 8% more URMs offered an interview invitation ($P < .01$).

**Conclusions**
Reliance on USMLE Step 1 as a primary screening tool precludes URMs from being considered for residency positions at higher rate than non-URMs. Developing screening tools to measure a wider array of candidate competencies can help create a more equitable surgical workforce.
found that adverse impact exists, that the organization will investigate suitable alternative selection methods that have as little adverse impact as possible. Similarly, the Standards for Educational and Psychological Testing require an ongoing program of validation to collect evidence about such tests’ validity, reliability, and fairness, among other things. It is likely for these and other reasons that the USMLE test developers themselves have warned the medical education community about use of USMLE scores for residency selection and have even acknowledged that there is a paucity of evidence linking Step 1 performance to residency success.

In this study, we sought to examine if reliance on the USMLE Step 1 for resident selection decisions served as a barrier for entrance into postgraduate surgery training programs for underrepresented minorities (URMs). Additionally, we investigated the extent to which alternative assessments could affect the composition of the pool of applicants considered for further screening.

Method

We conducted multimethod job analyses in July 2018 across 7 general surgery residency programs in Florida, Georgia, Ohio, and Texas to gather validity evidence to inform development of the new selection assessments to be used during the 2018–2019 application cycle. Each program identified 15–20 subject matter experts (SMEs) central to their residency, including the chair, PD, associate PD(s), critical members of the Clinical Competency Committee, high-performing incumbent trainees, and other key stakeholders associated with the program.

Two industrial organizational psychologists (IOPs) conducted an on-site job analysis for each program. The IOPs met with each SME and conducted a 1-hour semistructured interview using the critical incident technique to obtain input on the program’s culture, values, and demands. Each SME also completed a quantitative survey indicating the extent to which a number of competencies (professionalism, team orientation, resilience, self-directed learning, etc.) were required for success in the program and the extent to which they felt each competency might change in the future.

The list of competencies was derived from national accreditation bodies (Accreditation Council of Graduate Medical Education, CanMEDS) and the surgical education literature. IOPs provided SMEs a definition key with descriptions of each competency listed, and were also able to write in competencies not listed. Programs also provided historical documentation of performance remediation instances and reasons for attrition.

Based on the job analysis data, the IOPs determined the most required and desired competencies for trainees upon entry at each program. Situational judgment tests (SJTs), hypothetical but realistic scenarios in which respondents must indicate the effectiveness of a number of potential responses, were created to assess each competency. We chose SJTs as a selection tool because of their ability to assess multiple competencies simultaneously, high predictive validity for predicting future on-the-job performance, demonstrated ability to produce less adverse impact (negative effects for individuals from underrepresented groups) than traditional written examinations and resistance to applicant faking. The IOP team created unique SJTs, and scoring algorithms were created for each program based on SME input and scoring to maximize relevance and effectiveness of the assessment (see Supplemental Digital Appendix 1, available at http://links.lww.com/ACADMED/A776, for example).

Each program lowered its traditional USMLE Step 1 cutoff (which previously ranged from 220 to 240) to 210 (stage 1) and invited all otherwise eligible candidates to take their unique SJT as the next hurdle in their application process (stage 2). URM status (women, racial/ethnic minority) of candidates who would have been considered for an interview using traditional USMLE Step 1 cutoffs (i.e., the minimum score programs would have used if they did not have the SJT) was compared with the candidate pool considered based on SJT performance.

We analyzed basic descriptive statistics and frequencies via SPSS statistical software, version 25 (SPSS Inc., Armonk, New York). Interclass correlation coefficients were used to identify level of agreement across SMEs within each program. Via independent-samples t tests, we compared the number of URMs recommended for an interview based on the use of traditional USMLE cutoffs with the use of lowered USMLE cutoff plus SJT assessments. We also compared the proportion of URMs in each group via chi-square tests.

This study was deemed quality improvement through the University of Texas Institutional Review Board, and thus no IRB approval was required.

Results

Seven general surgery residency programs across the United States participated in this study. An average of 14 (± 2.52) SMEs participated in job analysis data collection at each site, for a total of 98 SMEs interviewed. As the competencies deemed most critical across programs differed significantly, IOPs developed unique SJT items to measure competencies for each program, in accordance with methods of other studies. Examples of critical competencies that were common across more than one program are integrity, communication, teamwork, dependability, and professionalism. To determine which items SMEs agreed upon (indicating shared values within a program) and to develop the scoring algorithm for each program, SMEs reviewed a large batch of items (50–55 items) and provided feedback and input indicating how they would prefer a junior resident in their program to respond to each of the scenarios. The Kendall coefficient of concordance was computed for each ranking item, and only those items above 0.70, indicating adequate interrater agreement, were retained. The final SJT assessment for each respective program included 20 items, which allowed for 100 unique data points (5 data points per item) to be collected and scored from each applicant.

A total of 2,742 categorical applicants (1,625 unique applicants; approximately 68% of all U.S. applicants to general surgery programs in 2018) who exceeded the new lowered USMLE Step 1 cutoff score of 210 were invited to take an online SJT assessment by at least 1 of the 7 general surgery programs. Slightly over one-half of these invited applicants were either male (54%; 878) or white.
(53%; 861), and 72% (1,170) represented at least one underrepresented group (woman, nonwhite).

Traditional USMLE thresholds would have resulted in a pool composed of 55.7% (1,527) male and 56.1% (1,538) white applicants. Lowering the USMLE Step 1 threshold resulted in approximately 35% (698) more applicants invited to take the SJT who would not have met traditional USMLE cutoffs by the programs and been excluded from further consideration. These individuals were more likely to be nonwhite (52.7% versus 43.9%, \( P < .001 \)) and female (47.5% versus 44.3%, \( P < .05 \)) compared with individuals permitted by traditional (higher) USMLE Step 1 cutoffs. Overall, more URM applicants were included in the initial applicant pool (74.1% versus 66.0%, \( P < .01 \)) as a result of the first stage of intervention (using a cutoff of 210 on USMLE Step 1 compared with each program’s typical USMLE cutoff).

For example, Figure 1 illustrates that program #1 would have considered 436 URM applicants for an on-site interview using traditional USMLE cutoffs, versus considering 587 URM applicants by using lower cutoffs. Programs were seeking to fill an average of 6.14 (standard deviation [SD] = 2.19) positions and invited an average of 391.71 (SD = 251.28) applicants who exceeded the new lower USMLE Step 1 cutoff to complete the SJT. Programs invited an average of 100 (SD = 46.94; range, 22–151) more underrepresented applicants to complete the SJT than the number who would have been considered with typical USMLE Step 1 cutoffs. Ninety-seven percent of invited applicants (2,662/2,744) completed the SJT within the respective program deadlines (3–14 days). Program completion rates ranged from 95% to 98%. Average time to complete the assessment across all programs was 35 minutes (SD = 21.60). Only 0.001% (n = 3) of applicants started but did not complete the assessment, indicating almost no test abandonment.

Figure 2 shows the impact of the second stage, SJT performance, on URM representation versus reliance on traditional USMLE cutoff scores. The new 2-step process (lower USMLE Step 1 cutoff plus SJT) increased the percentage of URM applicants offered an interview invitation by 8% on average across programs compared with the use of only traditional USMLE Step 1 cutoffs (\( P < .01 \)). All but one program invited more URM applicants for an on-site interview, with increases ranging from 1% to 17%. Figure 3 displays these differences for each program along with the overall percentage change in URM candidates across all stages of the process.

Discussion

Reliance on the USMLE for initial screening of applicants may bring with it a host of unintended consequences, including minimizing the relationship between the assessments and resident performance criteria, decreasing the number of URMs considered for later stages of the selection process, and opening programs to potential litigation. These and other reasons are likely why the developers of the exam, along with leaders in medical

---

**Figure 1** Impact of lower USMLE cutoff on diversity of candidates considered for next stage, from a study of USMLE Step 1 score cutoffs and situational judgment tests as applicant screening tools in resident selection at 7 surgical residency programs, 2018–2019. Abbreviations: URM, underrepresented minority; USMLE, United States Medical Licensing Examination; SJT, situational judgment test.
education, have admonished against its use for residency selection.6,7,15 However, if the movement toward making the USMLE a pass/fail test continues to gain momentum and succeeds, residency programs will need to develop or adopt additional tools designed for the purposes of selection. In fact, it is critical that programs work toward this aim now so the residency community is not left at a loss if or when new reporting standards emerge. Otherwise, we as a community might be at a risk of taking 1 step forward and 2 steps back.

Our study describes the efforts of 7 separate general surgery residency programs across the United States to implement a more evidence-based selection process. By adopting techniques from industry (initial job analysis led by IOPs and development and validation of assessment tools designed for the purposes of selection), we have demonstrated how tools with less potential for adverse impact can be successfully incorporated into the residency selection process while also incorporating many sources of validity evidence that align with contemporary validity frameworks.31 Through an in-depth job analysis with key SMEs associated with each training program, we built a foundation of specific and relevant content from which to develop the assessment. Additionally, we had each of these SMEs review the scenarios, potential responses, and assessment instructions to ensure appropriate content. This process also allowed for establishing evidence on internal structure by ensuring scenarios related to intended constructs and high interrater agreement among SMEs. Finally, we describe the preliminary consequences evidence as the beneficial impact that use of the assessment, and the decisions that arose from its results, had on consideration of underrepresented groups for on-site interviews. Given work showing that few academics and practitioners include evidence related to test consequences,32 these outcomes should be considered a strength of this study. Future work could incorporate additional sources of evidence by specifically questioning test takers about their performance strategies, responses to particular items, and understanding of the items through feedback surveys or interviews. Additionally, these efforts will become even more powerful once we are able to establish evidence based on relations to other variables.

As there is little standardized and objective information in each student's application packet, we were unable to compare performance on items from the online assessment to other established competency metrics. However, we plan future work to follow these individuals into residency and examine the relationships between competencies measured in the online assessment and relevant performance outcomes intended to measure those competencies.

Importantly, our data demonstrate that deemphasizing reliance on the USMLE for residency selection and instead relying on tools developed to be valid for resident selection can create more opportunities for URM candidates to be considered. The reasons for these differences are

---

**Figure 2** Increase in URM percentages in interview recommendation based on SJT, from a study of USMLE Step 1 score cutoffs and situational judgment tests as applicant screening tools in resident selection at 7 surgical residency programs, 2018–2019. Abbreviations: URM, underrepresented minority; USMLE, United States Medical Licensing Examination; SJT, situational judgment test.
likely multifaceted. For example, while the USMLE Step 1 primarily evaluates cognitive abilities, SJTs capture both cognitive and noncognitive traits and abilities and thus may reduce the influence of various environmental influences that have traditionally disadvantaged individuals from underrepresented or historically disadvantaged groups. Other factors, such as socioeconomic status and access to test preparation materials, may have less of an influence on SJT performance as well. On average, programs experienced an 8% increase in the percentage of URM candidates

**Table 1**

**Overview of Changes in Number of URM Candidates Across Screening Methods, From a Study of USMLE Step 1 Score Cutoffs and Situational Judgment Tests as Applicant Screening Tools in Resident Selection at 7 Surgical Residency Programs, 2018–2019**

<table>
<thead>
<tr>
<th>Program</th>
<th>Total no. applicants invited to take SJT</th>
<th>No. applicants invited to complete SJT who met traditional USMLE threshold</th>
<th>URM in traditional USMLE cutoff, no. (overall %)</th>
<th>Additional URM after lowering USMLE 1 threshold to 210, no. (overall %)</th>
<th>URM in SJT recommendation, no. (overall %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>812</td>
<td>632</td>
<td>436 (69)</td>
<td>151 (76)</td>
<td>61 (75)</td>
</tr>
<tr>
<td>2</td>
<td>138</td>
<td>91</td>
<td>54 (59)</td>
<td>22 (67)</td>
<td>30 (74)</td>
</tr>
<tr>
<td>3</td>
<td>511</td>
<td>367</td>
<td>250 (68)</td>
<td>122 (78)</td>
<td>69 (69)</td>
</tr>
<tr>
<td>4</td>
<td>173</td>
<td>76</td>
<td>48 (63)</td>
<td>73 (77)</td>
<td>33 (81)</td>
</tr>
<tr>
<td>5</td>
<td>290</td>
<td>114</td>
<td>84 (74)</td>
<td>65 (76)</td>
<td>42 (88)</td>
</tr>
<tr>
<td>6</td>
<td>329</td>
<td>144</td>
<td>85 (59)</td>
<td>132 (74)</td>
<td>37 (62)</td>
</tr>
<tr>
<td>7</td>
<td>579</td>
<td>276</td>
<td>190 (69)</td>
<td>133 (71)</td>
<td>67 (68)</td>
</tr>
</tbody>
</table>

Abbreviations: URM, underrepresented minority; SJT, situational judgment test; USMLE, United States Medical Licensing Examination.
recommended for an interview by relying upon SJTs for interview decisions. This 8% difference is noteworthy for 2 reasons. First, while only a small percentage at first glance, this 8% represents over 300 medical students who would have been rejected for consideration outright. As such, this seemingly small percentage reflects a large practical value for programs. Additionally, we were able to increase the percentage of URMs recommended for interview, while also narrowing the applicant pool to only those candidates who objectively possessed desired program competencies. Specifically, if programs were to rely solely on their USMLE cutoffs, they still would have to review the applications of the 250 applicants who exceeded that cutoff to determine which applicants to invite for an interview, increasing both the workload and opportunity for bias in the process. With the inclusion of an SJT, however, programs had an average of about 55 applicants recommended for an on-site interview. Thus, we were able to simultaneously reduce the total number of interviewees while also increasing the percentage of URMs.

Even if PDs can agree that the USMLE is not an ideal screening tool, they may be at a loss for suitable alternatives. Not only might PDs lack awareness of other useful tools that may exist to facilitate applicant screening activities, but they may also be unable to successfully create them on their own. Our study demonstrates that partnership with experts outside of medicine is critical to efficiently and effectively navigate this process. Without this outside expertise, these clinically trained PDs would not have the expertise or bandwidth to conduct numerous one-on-one job analysis interviews, perform quantitative and qualitative analytics on interview data, develop customized assessments designed for resident selection, or gather appropriate validity evidence at each step along the way. Each of these components are critical to ensuring a fair and valid selection process. Fortunately, our study found that these medical and nonmedical partnerships can move the needle forward in creating tools that allow programs to capture desired and required competencies while also leveling the playing field for applicants from diverse backgrounds.

This study is not without its limitations. We do not yet have measures of the follow-up performance of these candidates, although work is underway to follow them into residency. However, other research showing the value of SJTs to predict later performance and remediation in surgery residency is promising. Furthermore, we did not have data to measure socioeconomic status among these candidates, nor do we have data on the relationship between these results and other important attributes, metrics, and experiences included in the application packet, such as clerkship performance, citizenship metrics, and other extracurricular activities. Other studies have shown that SJTs can similarly widen access to individuals from disadvantaged socioeconomic backgrounds. Future research is needed to explore how these mechanisms unfold in postgraduate training selection in the United States. Future work will also need to expand the sources of validity evidence (i.e., response process and relations to other variables) to create a more powerful assessment development process. Finally, we have provided a snapshot of how these mechanisms play out among a handful of general surgery residency programs. Further work will need to be undertaken to investigate generalizability among other specialties and among the wider array of general surgery programs.

Conclusions
Our study demonstrates that reliance on the USMLE Step 1 for selection decisions may serve as a barrier for entrance into postgraduate surgery training programs for URMs. Fortunately, alternative assessments, such as SJTs, can be created to capture competencies valued by programs while also providing more equitable opportunities for underrepresented candidates.

Funding/Support: None reported.

Other disclosures: A.K. Gardner, K.J. Cavanaugh, R.E. Willis, and B.J. Dunkin all provide advice on selection and assessment to training programs and organizations through SurgWise Consulting.

Ethical approval: This study was deemed quality improvement through the University of Texas Institutional Review Board, and thus no IRB approval was required.

Previous presentations: These data were presented at the International Association for Medical Education (AMEE) Conference, August 27, 2019, Austria, Vienna, and at the Association of American Medical Colleges Workforce Research conference, May 2, 2019, Alexandria, Virginia.

A.K. Gardner is assistant dean of evaluation and research, Baylor College of Medicine, and president and CEO, SurgWise Consulting, Houston, Texas.

K.J. Cavanaugh is a research analyst, MD Anderson Cancer Center, and senior associate, SurgWise Consulting, Houston, Texas.

R.E. Willis is director of surgical education, University of Texas Health Sciences Center at San Antonio, San Antonio, Texas, and senior associate, SurgWise Consulting, Houston, Texas.

B.J. Dunkin is executive vice president, SurgWise Consulting, Houston, Texas.

References
12. Brothers TE, Wetherholt S. Importance of the faculty interview during the


33 Tiffin PA, Dowell JS, McLachlan JC. Widening access to UK medical education for under-represented socioeconomic groups: Modelling the impact of the UKCAT in the 2009 cohort. BMJ. 2012;344:e1805.