2023 APDS SPRING MEETING

Maximizing Standardization While Ensuring Equity: Exploring the Role of Applicant Experiences, Attributes, and Metrics on Performance of a Surgery-Specific Situational Judgment Test

Jennifer H. Chen, MD, [*] *Paula Costa, PhD,* [†,‡] *and Aimee K. Gardner, PhD* [*,†]

[*]Department of Surgery, Baylor College of Medicine, Houston, Texas; [†]SurgWise Consulting, Houston, Texas; and [‡]ICF International, Fairfax, Virginia

BACKGROUND: Situational judgment tests (SJT) are hypothetical but realistic scenario-based assessments that allow residency programs to measure judgment and decision-making among future trainees. A surgery-specific SJT was created to identify highly valued competencies among residency applicants. We aim to demonstrate a stepwise process for validation of this assessment for applicant screening through exploration of two often-overlooked sources of validity evidence − relations with other variables and consequences.

METHODS: This was a prospective multi-institutional study involving 7 general surgery residency programs. All applicants completed the SurgSJT, a 32-item test aimed to measure 10 core competencies: adaptability, attention to detail, communication, dependability, feedback receptivity, integrity, professionalism, resilience, self-directed learning, and team orientation. Performance on the SJT was compared to application data, including race, ethnicity, gender, medical school, and USMLE scores. Medical school rankings were determined based on the 2022 U.S. News & World Report rankings.

RESULTS: In total, 1491 applicants across seven residency programs were invited to complete the SJT. Of these, 1454 (97.5%) candidates completed the assessment. Applicants were predominantly White (57.5%), Asian (21.6%), Hispanic (9.7%), Black (7.3%), and 52% female. A total of 208 medical schools were represented, majority were allopathic (87.1%) and located in United States (98.7%). Less than a quarter of applicants (22.8%; N=337) were from a top 25 school based on U.S. News & World Report rankings for primary care, surgery, or research. Average USMLE Step 1 score was 235 (SD 37) and Step 2 score was 250 (SD 29). Sex, race, ethnicity, and medical school ranking did not significantly impact performance on the SJT. There was no relationship between SJT score and USMLE scores and medical school rankings.

CONCLUSIONS: We demonstrate the process of validity testing and importance of two specific sources of evidence−consequences and relations with other variables, in implementing future educational assessments. (J Surg Ed 000:1−8. © 2023 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: Situational Judgment Tests, Validity, Medical School Ranking, Core Competencies, Assessment, Nontechnical Skills

COMPETENCY: Professionalism, Interpersonal and Communication Skills, Systems-Based Practice

INTRODUCTION

Adoption of pass/fail scoring for USMLE Step 1 has placed increasing emphasis on alternative screening and selection tools in the residency application process.[1-6] Trainee selection in recent years has also shifted away from numerical scoring through incorporation of more noncognitive competencies such as interpersonal skills, professionalism, and feedback receptivity.[7-9] This trend is reflected in the 2022 National Resident Matching Program (NRMP) Program Directory Survey,[10] where an

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Correspondence: Inquiries to Jennifer Chen, MD, Baylor College of Medicine, One Baylor Plaza MS390, Houston, TX 77030; e-mail: Jennifer.chen2@bcm.edu

  1931-7204/$30.00 https://doi.org/10.1016/j.jsurg.2023.05.024

increasing number of programs are utilizing a holistic approach in applicant review, and the ACGME Surgery Milestones,[11] where 12 out of the 18 surgery milestones focus on nonmedical and nonprocedural skill sets. However, despite these changes, current tools in residency selection have not evolved to adequately assess nontechnical skills in residency applicants. Traditional metrics such as USMLE Step 2 CK scores, MSPE letters, preclinical and clerkship grades, and letters of recommendations often leave room for implicit bias and can further perpetuate inequities.[12-20] New emerging assessments such as personality trait analysis,[21,22] emotional intelligence scoring,[22-26] use of grit scales,[2,3] and various institution-specific screening tools[27-29] are being explored to determine their utility in the residency screening process.

One modality that has been shown to be effective in assessing an array of nontechnical competencies is the situational judgment test (SJT).[22,30] SJTs consist of hypothetical but realistic scenario-based assessments that measure judgment and decision-making during times of uncertainty. By placing residents in scenarios mimicking real-life situations in residency, SJTs can provide incremental validity data beyond that of knowledge-based assessments and ultimately help predict future performance in training.[31-33] The basic principles of SJTs rely on the theory of behavioral consistency, where past behavior serves as the best predictor of future behavior.[34] Hypothetical scenarios require applicants to extrapolate from prior experiences in similar settings and use previous actions to guide future decisions. Patterson et al. performed a systematic review of use of SJTs in measurement of nonacademic attributes and found that SJTs demonstrated reliability, predictive validity, and incremental validity in measuring nontechnical dimensions such as empathy, integrity, and resilience in the work setting.[35] While SJTs have been utilized extensively in industrial settings, their use in residency selection have largely been limited to countries abroad (UK, Belgium, Canada, Israel, Singapore, Australia).[31-33,35,36] Given its successes in predicting performance in other occupational settings, we aimed to examine the suitability of a surgery-specific SJT in residency applicant screening and selection.

The overall goal of this study was to collect validity evidence to evaluate appropriateness of use of a surgery-specific SJT in surgical trainee selection. As with any new assessment, multiple sources of validity evidence are necessary to confirm its reliability and assumptions prior to interpretation and application of its results. Our validation process is based off the framework proposed by Messick, which uses five sources of validity evidence: content, internal structure, relationships with other variables, response process, and consequences.[37] Relations with other variables and consequence validity evidence represent two commonly underreported sources of evidence. Relations with other variable examines the association of the assessment with other variables and known metrics that measure the same construct. Consequences examine the future impact of the assessment, whether it be beneficial or harmful, intentional or unintentional. In particular, consequence validity evidence is often viewed as the most crucial to obtain because if an assessment has unintended impacts, one should argue that it should not be employed at all. Validity evidence related to content and internal structure were previously collected during development of the SJT. We aim to provide two additional sources of validity evidence, relationship with other metrics of applicant evaluation and consequences of scores on applicant performance.

MATERIALS AND METHODS

Study Design

We conducted a prospective cohort study of general surgery applicants to seven academic medical centers throughout the United States. All applicants were asked to complete a surgery-specific SJT as a part of their application. The SJT measures 10 core competencies: adaptability, attention to detail, communication, dependability, feedback receptivity, integrity, professionalism, resilience, self-directed learning, and team orientation. Applicants were asked to respond to SJT items on a scale of 1 (not effective at all) to 6 (highly effective). Additional applicant information including race, ethnicity, sex, USMLE Step 1 and Step 2 scores, and medical school of graduation were obtained from application packets. Applicants' medical schools of graduation were tallied according to the 2022 U.S. News and World Report rankings based on primary care, research, and surgery.[38-40] Schools were sorted based on their assigned ranking numbers and categorized to indicate whether they were in the top 5, 10, or 25 of each category.

Situational Judgment Tests (SurgSJT)

The SurgSJT is a surgery-specific SJT developed from five years of data from customized selection assessments used across general surgery residency programs.[41] It was collated from over 1000 SJT items and measures 10 core competencies deemed to be those most valued in future residents from a national sample of faculty and trainees. Each item underwent iterative review by subject matter experts (SME), demonstrated high level of interrater agreements, exhibited broad generalizability across programs, and underwent concordance analysis against ACGME milestones. In total, 504 surgeons across 12 subspecialties were involved in its development and each

SJT item was reviewed by an average of 38 surgeons. The final assessment was tested in a pool of over 5000 residency and fellowship applicants, which represented over 92% of medical schools in the United States. The SurgSJT used in this study consisted of the final 32 items derived from this process.

Data Analyses

Basic descriptive statistics (means, ranges, frequencies, standard deviations) were used to examine demographic data and USMLE scores. Independent t-tests were used to examine differences between two groups, such as medical school ranking and sex. Multiple analyses of variance with post hoc Tukey tests were used to explore potential differences between multiple groups, such as race and ethnicity and competency area. Relationships between competencies, USMLE scores, and medical school ranking were examined using correlation coefficients. All statistical tests were two-sided and $p < 0.05$ was considered statistically significant. All data were analyzed using SPSS version 28.0.

RESULTS

Baseline Characteristics

In total, 1491 applicants were invited to complete the SJT. Of these, 1454 candidates responded (97.5%) and 1437 (98.9%) had at least some demographic data available. Applicants consisted of 52% women and were predominantly White (57.5%) and Asian (21.6%). The remaining applicant pool consisted of Hispanic (9.7%), Black (7.3%), American Indian (0.8%), Hawaiian and Pacific Islander (0.1%), and Other (2.9%). A total of 208 medical schools were represented, with majority of them allopathic (87.1%) and located in the United States (98.7%). Average USMLE Step 1 score was 235 (SD 37) and Step 2 score was 250 (SD 29). Less than a quarter (22.8%; N=337) of applicants were from a top 25 school based on U.S. News & World Report rankings for primary care, surgery, or research. Specifically, 16.2% (N=240) were from a Top 25 program in primary care, approximately 10% (N=146) were from a Top 25 program for research, and slightly more (10.8%; N=160) were from a Top 25 program for surgery. Of these, 4.7% (N=69) were from a medical school with a Top 25 ranking in at least two categories, and 4.7% (N=70) were from a medical school with a Top 25 ranking in all three categories. Schools that were ranked in all three categories include Baylor College of Medicine, Harvard University, University of California – Los Angeles, University of California–San Francisco, University of Michigan, University of Pennsylvania, University of Pittsburgh, University of Texas Southwestern Medical Center, and University of Washington.

SJT and Applicant Demographic Data

Performance on the SJT did not differ by sex, race, or ethnicity of the applicant. More specifically, the 10 core competencies measured by the SJT demonstrated no significant differences across sex, race, or ethnicity.

SJT and Medical School Ranking

Overall, SJT scores did not differ by medical school rankings categorized by primary care, research, or surgery. This result was consistent across the top 5, 10, and 25 medical schools in each category. When comparing across the 10 core competencies, applicants from a top 10 ranked primary care medical school performed worse on SJT items measuring resiliency ($p = 0.028$) and those from a top 25 ranked primary care school performed better on items measuring self-directed learning ($p = 0.019$). Additionally, candidates from top 5 and top 25 ranked medical schools in research performed worse on SJT items measuring dependability ($p = 0.008$ and 0.048, respectively). A similar trend was seen in applicants from top 5 and top 10 ranked schools in surgery ($p = 0.026$ (N=17) and $p < 0.001$ (N=46), respectively), though those from top 25 ranked schools demonstrated no difference in performance on SJT items measuring dependability.

SJT and USMLE Scores

There was no significant difference between overall SJT scores and USMLE Step 1 and Step 2 scores (Step 1 CI: -0.1 to 0.002 and Step 2 CI: -0.08 to 0.03). Additionally, there was no significant difference between USLME scores and core competencies assessed with the SJT items.

SJT and Type of Medical School

We found no significant difference in SJT scores between applicants from allopathic versus osteopathic schools. Additionally, USMLE scores and core competencies measured by SJT did not differ based on the type of medical school attended.

DISCUSSION

Our study explores the stepwise process of collecting validity evidence to support use of a new screening and selection tool among surgery trainees. Specifically, we focused on two underexplored sources of validity evidence—relations with other variables and consequences—to confirm that the tool is not redundant with other

TABLE 1. SurgSJT Performance by Experiences, Attributes, and Metrics

	SurgSJT		
	Low (Bottom 25th Percentile)	**Average (25-75th Percentile)**	**High (Top 25th Percentile)**
USMLE1	234	236	237
USMLE2	237	237	241
Women	52%	54%	50%
Underrepresented Race	23%	21%	19%
Top 25, Primary Care	14%	18%	16%
Top 25, Research	7%	12%	8%
Top 25, Surgery	7%	13%	11%

Note: USMLE mean values are provided. All others are frequencies.

applicant metrics and has no unintended consequences for women or minorities. Our findings reveal that a surgery-specific SJT can allow for a standardized approach to screen candidates while also ensuring equitable consideration of applicants regardless of sex, race, and ethnicity Table 1.

We derived our validation process from the framework first proposed by Messick in 1989, which views validation as a hypothesis of the interpretation and applications of the assessment of interest.[37,42,43] The validity hypothesis is supported by a body of evidence from five different sources: content, internal structure, relationships with other variables, response process, and consequences.[37] Consequence validity evidence is a crucial component of this process because it examines future impacts of assessments, whether it be intended or unintended and harmful or beneficial.[42,43] It is often viewed as the most important source of validity evidence because if an assessment does not meet its intended impact, then it should not be utilized. Conversely, if an assessment has unintended impact (e.g., lower scores for underrepresented groups), its use should be reconsidered. Despite this recognized importance, consequence evidence is infrequently reported, with only up to 5% to 20% of educational assessments in health professions examining it.[44,45] One of the goals of this study was to further contribute to the fields of surgical education and residency selection by examining this important source of validity evidence in a large national sample. We were able to examine the consequences of implementing a surgery-specific SJT on the demographic composition of test takers. Our study revealed that performance on the SJT was not related to sex, race, or ethnicity of applicants. Furthermore, individual performance on each of the 10 core competencies measured did not vary by demographic data. Overall consistency of performance on the SJT across these subgroups provides strong validity evidence in support of its use in surgical trainee screening and selection.

Our second source of validity evidence focuses on relationships with other variables, namely, the associations between performance on the SJT and other metrics used in applicant evaluation. Specifically, we focused on medical school rankings and USMLE Step 1 and 2 scores. Our finding that overall SJT scores did not differ based on medical school rankings for categories of primary care, research, and surgery is an important result that warrants further discussion. This topic is especially poignant due to the recent withdrawal of up to 13 medical schools from U.S. News and World Report rankings.[46,47] Institution ranking and reputation undeniably influences admission and selection process for medical school, residency, and fellowship positions, whether intentional or not.[48-50] Based on the 2021 NRMP Program Director Survey, 44.7% of program directors endorsed consideration of medical school reputation when deciding which applicants to interview and over 20% when making the rank list.[51] Additionally, medical school reputation was rated with a mean importance of 3.6 and 3.9 (1 = not at all important, 5 = very important) by program directors when making interview decisions and determining the rank list, respectively.[51] However, as several schools who recently withdrew from ranking participation have highlighted, the ranking process is a flawed system that lacks validity, accuracy, and objectivity.[52-55] Our study supports this initiative as we found no differences between overall SJT performance and school ranking. It is worth noting that our results did show some differences across medical school rankings on specific SJT items focused on resiliency, self-directed learning, and dependability. However, while this difference was statically significant, we did not consider this result meaningful in application due to substantially unequal sample sizes and presence of extreme outliers. For instance, one applicant from a top-rated school that performed poorly on 1 core competencies are likely driving the few differences observed.

Furthermore, we found no difference between performance on the SJT and performance on the USMLE. This finding is reassuring as several previous studies have shown that USMLE exams have poor predictive value and do not correlate with future performance in residency.[12,56-58] Additionally, the USMLE fails to adequately assess nontechnical core competencies such as teamwork, leadership, and communication.[5,8,9,12,57] Given that USMLE Step 1 scoring has already transitioned to pass/fail scoring, we hope SJTs and other rigorously developed assessments and approaches can be used to measure relevant competencies in a more objective and equitable manner.

As with any study, there are limitations that warrant further discussion. First, while our sample size is relatively large, it is limited to only 7 general surgery training programs in the United States. According to the latest data available from the Electronic Residency Application Service (ERAS), there were 2718 U.S. MD grads in the 2021-2022 application cycle.[59] While our sample reflects the majority of these applicants (55%), it is not entirely inclusive. Further testing of the SJT in more residency programs and across geographical regions will provide further validity evidence and generalizability. Second, our applicant pool had substantial variability in sample sizes across demographic groups. For instance, we had 1 Hawaiian and Pacific islander applicant and 12 American Indian applicants, compared to 857 White applicants. Unfortunately, our demographic composition is reflective of the overall diversity of the general surgery applicant pool. Thus, we are limited in our ability to fully analyze the impact of large-scale educational assessments until the field of surgery becomes more diverse. Third, we were unable to evaluate performance on SJT against metrics in applicant packet with qualitative measures, such as letters of recommendation, personal statements, and MSPE letters. Based on the 2021 NRMP Program Director Survey results, these long-standing components of applicant profiles are equally as important when making interview and ranking decisions—with mean importance rankings ranging from 3.8-4.3 out of 5 (1 = not at all important and 5 = very important).[51] While several studies have identified implicit bias based on gender and race and ethnicity in these metrics,[13-17,19,20,55] if their use continues it may be fruitful to assess relationships between them and SJT scores. Lastly, we were unable to obtain evidence related to response process, which examines applicant thought processes or actions during the assessment. Given the context in which the SJT was administered, our study did not allow for applicant feedback. This is the only remaining source of validity evidence not explored with this surgery-specific SJT, and thus future research is warranted.

CONCLUSIONS

Principles and guidelines for test development require a stepwise validation process to ensure appropriateness of future interpretations, uses, and decisions derived from assessment scores.[60] Our study provides an example of how two underutilized sources of validity evidence—consequences & relations with other variables—can be used when exploring the utility and impact of a new screening tool for residency selection. We hope our detailing of validity testing will encourage others in the surgical education community to employ these measures when implementing future assessments for these purposes ().

REFERENCES

1. Asaad M, Drolet BC, Janis JE, Giatsidis G. Applicant familiarity becomes most important evaluation factor in USMLE step I conversion to pass/fail: a survey of plastic surgery program directors. *J Surg Educ*. 2021;78:1406–1412. https://doi.org/10.1016/j.jsurg.2021.01.007.
2. Kurian EB, Desai VS, Turner NS, et al. Is grit the new fit?—Assessing non-cognitive variables in orthopedic surgery trainees. *J Surg Educ*. 2019;76:924–930. https://doi.org/10.1016/j.jsurg.2019.01.010.
3. Camp CL, Wang D, Turner NS, Grawe BM, Kogan M, Kelly AM. Objective predictors of grit, self-control, and conscientiousness in orthopaedic surgery residency applicants. *J Am Acad Orthop Surg*. 2019;27:e227–e234 https://journals.lww.com/jaaos/Fulltext/2019/03010/Objective_Predictors_of_Grit,_Self_Control,_and.9.aspx.
4. Lund S, D'Angelo J, D'Angelo AL, Heller S, Stulak J, Rivera M. New heuristics to stratify applicants: predictors of general surgery residency applicant step 1 scores. *J Surg Educ*. 2022;79:349–354. https://doi.org/10.1016/j.jsurg.2021.10.007.
5. Naides AI, Ayyala HS, Lee ES. How do we choose? A review of residency application scoring systems. *J Surg Educ*. 2021;78:1461–1468. https://doi.org/10.1016/j.jsurg.2021.02.003.
6. Patel H, Yakkanti R, Bellam K, Agyeman K, Aiyer A. Innovation in resident selection: life without step 1. *J Med Educ Curric Dev*. 2022;9. https://doi.org/10.1177/23821205221084936.
7. Potts J R. General surgery residency: past, present and future. *Curr Probl Surg*. 2019;56:174–197. https://doi.org/10.1067/j.cpsurg.2019.01.005.

8. Potts JRIII. Shifting sands of surgical education. *J Am Coll Surg*. 2018;227:151–162 https://journals.lww.com/journalacs/Fulltext/2018/08000/Shifting_Sands_of_Surgical_Education.1.aspx.

9. Gardner AK, Cavanaugh KJ, Willis RE, et al. Great expectations? Future competency requirements among candidates entering surgery training. *J Surg Educ*. 2020;77:267–272. https://doi.org/10.1016/j.jsurg.2019.09.001.

10. National Resident Matching Program. *National Resident Matching Program, Data Release and Research Committee: Results of the 2022 NRMP Program Director Survey*.; 2022.

11. The Accreditation Council for Graduate Medical Education. Surgery Milestones. 2019.

12. 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 https://journals.lww.com/academicmedicine/Fulltext/2011/01000/Are_United_States_Medical_Licensing_Exam_Step_1.20.aspx.

13. Sobol DL, Berfield KS, Shalhub S, et al. Exploring the influence of gender on surgical clerkship grades and test scores: a single institution, multisite comparison. *J Surg Educ*. 2022;79:1132–1139. https://doi.org/10.1016/j.jsurg.2022.05.008.

14. Storino A, Polanco-Santana JC, Ruiz de Somocurcio J, Sampson R, Gangadharan SP, Kent TS. Impact of surgeon gender and seniority in use of agentic and communal language in letters of recommendation for surgery residency applicants. *J Surg Educ*. 2022;79:1140–1149. https://doi.org/10.1016/j.jsurg.2022.04.002.

15. Isaac C, Chertoff J, Lee B, Carnes M. Do students' and authors' genders affect evaluations? A linguistic analysis of medical student performance evaluations. *Acad Med*. 2011;86:59–66 https://journals.lww.com/academicmedicine/Fulltext/2011/01000/Do_Students__and_Authors__Genders_Affect.22.aspx.

16. Polanco-Santana JC, Storino A, Souza-Mota L, Gangadharan SP, Kent TS. Ethnic/racial bias in medical school performance evaluation of general surgery residency applicants. *J Surg Educ*. 2021;78:1524–1534. https://doi.org/10.1016/j.jsurg.2021.02.005.

17. Turrentine FE, Dreisbach CN, St Ivany AR, Hanks JB, Schroen AT. Influence of gender on surgical residency applicants' recommendation letters. *J Am Coll Surg*. 2019;228:356–356e3 https://journals.lww.com/journalacs/Fulltext/2019/04000/Influence_of_Gender_on_Surgical_Residency.6.aspx.

18. Papp KK, Polk HC, David Richardson J. The relationship between criteria used to select residents and performance during residency. *Am J Surg*. 1997;173:326–329. https://doi.org/10.1016/S0002-9610(96)00389-3.

19. Chen S, Beck Dallaghan GL, Shaheen A. Implicit gender bias in third-year surgery clerkship MSPE narratives. *J Surg Educ*. 2021;78:1136–1143. https://doi.org/10.1016/j.jsurg.2020.10.011.

20. Go C, Lang S, Byrne M, Brucha DL, Parviainen K, Sachdev U. Linguistic analysis of letters of recommendation for vascular surgery and obstetrics and gynecology applicants detects differences in attributable strengths based on gender. *J Surg Educ*. 2021;78:1535–1543. https://doi.org/10.1016/j.jsurg.2021.02.002.

21. Hughes BD, Perone JA, Cummins CB, et al. Personality testing may identify applicants who will become successful in general surgery residency. *J Surg Res*. 2019;233:240–248. https://doi.org/10.1016/j.jss.2018.08.003.

22. Gardner AK, Dunkin BJ. Evaluation of validity evidence for personality, emotional intelligence, and situational judgment tests to identify successful residents. *JAMA Surg*. 2018;153:409–416. https://doi.org/10.1001/jamasurg.2017.5013.

23. Leddy JJ, Moineau G, Puddester D, Wood TJ, Humphrey-Murto S. Does an emotional intelligence test correlate with traditional measures used to determine medical school admission? *Acad Med*. 2011;86:S39–S41 https://journals.lww.com/academicmedicine/Fulltext/2011/10001/Does_an_Emotional_Intelligence_Test_Correlate_With.10.aspx.

24. Carrothers RM, Gregory SW Jr, Gallagher TJ. Measuring emotional intelligence of medical school applicants. *Acad Med*. 2000;75:456–463 https://journals.lww.com/academicmedicine/Fulltext/2000/05000/Measuring_Emotional_Intelligence_of_Medical_-School.16.aspx.

25. Lin DT, Kannappan A, Lau JN. The assessment of emotional intelligence among candidates interviewing for general surgery residency. *J Surg Educ*. 2013;70:514–521. https://doi.org/10.1016/j.jsurg.2013.03.010.

26. McKinley SK, Petrusa ER, Fiedeldey-Van Dijk C, et al. A multi-institutional study of the emotional

intelligence of resident physicians. *Am J Surg*. 2015;209:26–33. https://doi.org/10.1016/j.amjsurg.2014.09.015.

27. Bowe SN, Weitzel EK, Hannah WN Jr, et al. Introducing a novel applicant ranking tool to predict future resident performance: a pilot study. *Mil Med*. 2017;182:e1514–e1520. https://doi.org/10.7205/MILMED-D-15-00436.

28. Villwock JA, Hamill CS, Sale KA, Sykes KJ. Beyond the USMLE: The STAR algorithm for initial residency applicant screening and interview selection. *J Surg Res*. 2019;235:447–452. https://doi.org/10.1016/j.jss.2018.07.057.

29. Lyons J, Bingmer K, Ammori J, Marks J. Utilization of a novel program-specific evaluation tool results in a decidedly different interview pool than traditional application review. *J Surg Educ*. 2019;76:e110–e117. https://doi.org/10.1016/j.jsurg.2019.10.007.

30. Gardner AK, Cavanaugh KJ, Willis RE, Dunkin BJ. Can better selection tools help us achieve our diversity goals in postgraduate medical education? Comparing use of USMLE step 1 scores and situational judgment tests at 7 surgical residencies. *Acad Med*. 2020;95:751–757 https://journals.lww.com/academicmedicine/Fulltext/2020/05000/Can_Better_-Selection_Tools_Help_Us_Achieve_Our.29.aspx.

31. Cullen MJ, Zhang C, Marcus-Blank B, et al. Improving our ability to predict resident applicant performance: validity evidence for a situational judgment test. *Teach Learn Med*. 2020;32:508–521. https://doi.org/10.1080/10401334.2020.1760104.

32. Lievens F, Patterson F. The validity and incremental validity of knowledge tests, low-fidelity simulations, and high-fidelity simulations for predicting job performance in advanced-level high-stakes selection. *J Appl Psychol*. 2011;96:927–940. https://doi.org/10.1037/a0023496.

33. Patterson F, Ashworth V, Zibarras L, Coan P, Kerrin M, O'Neill P. Evaluations of situational judgement tests to assess non-academic attributes in selection. *Med Educ*. 2012;46:850–868. https://doi.org/10.1111/j.1365-2923.2012.04336.x.

34. Motowidlo SJ, Dunnette MD, Carter GW. An alternative selection procedure: the low-fidelity simulation. *J Appl Psychol*. 1990;75:640–647. https://doi.org/10.1037/0021-9010.75.6.640.

35. Patterson F, Lievens F, Kerrin M, Munro N, Irish B. The predictive validity of selection for entry into postgraduate training in general practice: evidence from three longitudinal studies. *Br J Gen Pract*. 2013;63:734–741. https://doi.org/10.3399/bjgp13X674413.

36. UKCAT Consortium. UCAT University Clinical Aptitude Test. 2019.

37. Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. *Adv Simulation*. 2016;1:31. https://doi.org/10.1186/s41077-016-0033-y.

38. U.S. News and World Report. 2023 Best Medical Schools: Primary Care. 2022.

39. U.S. News and World Report. 2023 Best Medical Schools: Research. 2022.

40. U.S. News and World Report. Best Surgery Programs. 2022.

41. SurgWise. SurgSJT. 2019.

42. Cook DA, Lineberry M. Consequences validity evidence: evaluating the impact of educational assessments. *Acad Med*. 2016;91:785–795 https://journals.lww.com/academicmedicine/Fulltext/2016/06000/Consequences_Validity_Evidence__Evaluating_the.18.aspx.

43. Moss PA. The role of consequences in validity theory. *Educ Meas Issues Prac*. 1998;17:6–12. https://doi.org/10.1111/j.1745-3992.1998.tb00826.x.

44. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med*. 2005;20:1159–1164. https://doi.org/10.1111/j.1525-1497.2005.0258.x.

45. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Health Sci Educ*. 2014;19:233–250. https://doi.org/10.1007/s10459-013-9458-4.

46. Kayser A. 13 medical schools boycott US News rankings: who, why and what's next. *Becker's Hosp Rev*. 2023.

47. Daskivich TJ, Gewertz BL. Campaign reform for US news and world report rankings. *JAMA Surg*. 2023;158:114–115. https://doi.org/10.1001/jamasurg.2022.4511.

48. Kortz MW, McCray E, Strasser T, et al. The role of medical school prestige and location in neurosurgery residency placement: an analysis of data from 2016 to 2020. *Clin Neurol Neurosurg*. 2021;210:106980. https://doi.org/10.1016/j.clineuro.2021.106980.

49. Goshtasbi K, Tsutsumi K, Merna C, Kuan EC, Haidar YM, Tjoa T. Does medical school geography and ranking influence residency match in otolaryngology? *Ann Otol Rhinol Laryngol*. 2021;131:485–492. https://doi.org/10.1177/00034894211026482.

50. Holderread BM, Liu J, Craft HK, Weiner BK, Harris JD, Liberman SR. Analysis of current orthopedic surgery residents and their prior medical education: does medical school ranking matter in orthopedic surgery match? *J Surg Educ*. 2022;79:1063–1075. https://doi.org/10.1016/j.jsurg.2022.02.004.

51. National Resident Matching Program. *National Resident Matching Program, Data Release and Research Committee: Results of the 2021 NRMP Program Director Survey*.; 2021.

52. McGaghie WC. America's best medical schools: a renewed critique of the U.S. news & world report rankings. *Acad Med*. 2019;94:1264–1266 https://journals.lww.com/academicmedicine/Fulltext/2019/09000/America_s_Best_Medical_Schools__A_Renewed_Critique.8.aspx.

53. Phillips RL Jr, Bazemore AW, Westfall JM. Increasing transparency for medical school primary care rankings—moving from a beauty contest to a talent show. *JAMA Health Forum*. 2021;2:e213419. https://doi.org/10.1001/jamahealthforum.2021.3419.

54. Goldstein MJ, Lunn MR, Peng L. What makes a top research medical school? A call for a new model to evaluate academic physicians and medical school performance. *Acad Med*. 2015;90:603–608 https://journals.lww.com/academicmedicine/Fulltext/2015/05000/What_Makes_a_Top_Research_Medical_School__A_Call.20.aspx.

55. Bowen CJ, Kersbergen CJ, Tang O, Cox A, Beach MC. Medical school research ranking is associated with gender inequality in MSTP application rates. *BMC Med Educ*. 2018;18:187. https://doi.org/10.1186/s12909-018-1306-z.

56. McDade W, Vela MB, Sánchez JP. Anticipating the impact of the USMLE Step 1 pass/fail scoring decision on underrepresented-in-medicine students. *Acad Med*. 2020;95:1318–1321 https://journals.lww.com/academicmedicine/Fulltext/2020/09000/Anticipating_the_Impact_of_the_USMLE_-Step_1.29.aspx.

57. 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 https://journals.lww.com/academicmedicine/Fulltext/2016/01000/A_Plea_to_Reassess_the_Role_of_United_States.11.aspx.

58. Sutton E, Richardson JD, Ziegler C, Bond J, Burke-Poole M, McMasters KM. Is USMLE Step 1 score a valid predictor of success in surgical residency? *Am J Surg*. 2014;208:1029–1034. https://doi.org/10.1016/j.amjsurg.2014.06.032.

59. Association of American Medical Colleges. *ERAS Statistics Preliminary Data*. Washington, DC; 2023.

60. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Position Statement on High-Stakes Testing. 2000.