Global Surgical Education - Journal of the Association for Surgical Education (2024) 3:51
https://doi.org/10.1007/s44186-024-00250-9

COMMENTARY

Proposed reforms to the American Board of Surgery In-Training Examination: aligning test design and implementation with accepted standards and intended use

Tejas S. Sathe[1] · Jaeyun Jane Wang[1] · Ava Yap[1] · Nina W. Zhao[2,3] · Patricia S. O'Sullivan[1] · Adnan Alseidi[1]

Received: 6 October 2023 / Revised: 15 January 2024 / Accepted: 11 April 2024

Keywords ABSITE · Validity · Formative assessment · Summative assessment

The American Board of Surgery In-Training Examination (ABSITE®) is a multiple-choice exam administered yearly to surgical residents. The ABS provides the following information on its website:

"The ABSITE is furnished to program directors as a formative evaluation instrument to assess residents' progress. The results are released only to program directors and should not be shared outside of the department GME division." [1]

However, the use of ABSITE scores is often inconsistent with the stated guidelines. While formative assessment is a low-stakes evaluation of an individual's current learning with an emphasis on actionable feedback, summative assessment is a high-stakes evaluation in which the primary objective is to hold an individual accountable for a body of knowledge [2]. Currently, ABSITE scores are used summatively in many fellowship applications and are routinely shared outside departments. This misalignment between intended and actual use challenges the validity of ABSITE scores.

Validity refers to the body of evidence that supports the use of a test score to accurately interpret a construct, such as "surgical knowledge" or "preparedness for fellowship" [3]. Specifically, there should be evidence that (1) test content maps to the construct in question (content), (2) test questions evoke the intended thought processes (response process), (3) test questions collectively measure the intended construct (internal structure), (4) test scores correlate in expected ways to related measures of the construct (relationship to other variables), and (5) use of test scores leads to intended outcomes (consequences) [3]. This framework was initially proposed by Messick and later operationalized into practical guidelines in *The Standards for Educational and Psychological Testing* [3, 4]. Using *The Standards*, we discuss the limitations of the ABSITE in its current form, including (1) invalid score interpretation, (2) inconsistent use of scores, (3) disparate learning opportunities for test-takers, and (4) variability in test administration. We then suggest future reforms.

First, **Standard 1.4** states, "If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use" [4]. While studies differ on the importance of the ABSITE in fellowship ranking decisions, all of them indicate that ABSITE scores are a considered component [5, 6]. In doing so, fellowships implicitly view them as a measure of candidate quality. Although some studies have correlated low ABSITE scores with failure of the ABS Qualifying Exam (QE), the data are mixed [7, 8]. Furthermore, studies demonstrate poor correlation between ABSITE scores and clinical evaluations, suggesting that the exam is a poor predictor of clinical performance [9]. Thus, while the ABS does provide a blueprint on how questions map to domains of surgical knowledge (content evidence), there is insufficient evidence that ABSITE scores actually correlate with clinical competence (relationship to other variables) [10].

Second, **Standard 5.24** states, "When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly" [4]. Currently, the ways in which

✉ Tejas S. Sathe
tejas.sathe@ucsf.edu

1 Department of Surgery, University of California San Francisco, San Francisco, CA, USA

2 Department of Otolaryngology - Head and Neck Surgery, University Hospitals Cleveland Medical Center, Cleveland, OH, USA

3 School of Medicine, Case Western Reserve University, Cleveland, OH, USA

Published online: 02 May 2024

fellowship programs use ABSITE scores are heterogeneous and opaque. The number of ABSITE scores that fellowships request from applicants can vary, even across fellowships in the same specialty. Moreover, when they exist, score cutoffs are not transparently published. One study demonstrated that although a minority of fellowship programs enforced ABSITE score cutoffs, those that did had cutoffs ranging from the 10th percentile to the 90th percentile [5, 11]. Because no rigorous methodology for determining ABSITE score cutoffs has been described, fellowships may be inadvertently excluding qualified candidates. As a result, the summative use of the ABSITE lacks consequence validity evidence [12].

Third, **Standard 12.8** states, "When test results contribute substantially to decisions about student promotion or graduation, evidence should be provided that students have had an opportunity to learn the content and skills measured by the test" [4]. Thus, without adequate study conditions, ABSITE scores cannot be used summatively. Given institutional differences in protected education time and clinical schedules, residents may not have adequate opportunities to prepare for the ABSITE. In a study analyzing the relationship between burnout and ABSITE scores, 48% of the respondents viewed "no time secondary to clinical duties" as a barrier to studying. Moreover, burnout due to exhaustion was associated with low ABSITE scores in a multivariable regression analysis [13]. Proponents of maintaining the ABSITE in its current form believe that it incentivizes residents to study, with a recent opinion article noting "an important by-product of a rigorous, scored exam is the incentive for residents to read even though they may be too tired or distracted to do so" [14]. While this may be true, it also reflects the reality that residents are frequently studying under suboptimal conditions and are not given a reasonable opportunity to learn the required content.

Fourth, **Standard 3.0** states, "All steps in the testing process, including…administration…should be designed…to minimize…variance" [4]. Currently, residents are assigned a time slot within a five day testing window to take the ABSITE. However, some residents may take the test while on-call or following overnight call when fatigued and sleep deprived. While previous studies failed to show a correlation between prior night call and exam scores, subjecting a resident to a five hour test post-call seems unnecessary [15–17]. If ABSITE scores continue to be misused in a summative manner, then test takers must be given a fair and equitable chance to perform their best. The prior two examples highlight situations in which scores may inadvertently capture structural differences related to learning conditions or fatigue instead of how well residents can understand and answer the questions being asked. As a result, the summative use of ABSITE scores is again undermined by a lack of consequence and response process evidence [3, 12].

Considering these issues, we can improve the design and implementation of the ABSITE. Ideally, we propose that fellowships stop requesting ABSITE scores. While residents technically report ABSITE scores voluntarily, in practice, this is a forced choice since residents perceive score reporting as the expectation and the norm. Alternatively, current practice can be improved by standardizing score reporting and optimizing the testing environment. Specifically, fellowships should be consistent and transparent about how many scores they request, what score cutoffs they use, and how these cutoffs are determined. In addition, the testing window can be widened to accommodate residents' clinical schedules. Notably, this was done in 2021 due to the COVID-19 pandemic with minimal issues [18]. Alternatively, multiple testing windows can be offered, as is done with the MCAT and the USMLE. Finally, residency programs should prioritize using the ABSITE in a formative manner. For example, scores can be used to guide educational quality improvement or formulate personalized learning plans [19]. As an added benefit, transitioning to a low-stakes examination reduces the incentive to cheat and thus the need for stringent test security.

Adoption of the proposed reforms invites the question of how fellowships will evaluate applicants without ABSITE scores, as they are currently one of the few quantitative metrics available. As one alternative, the newly developed entrustable professional activities (EPAs) may offer an evidence-based approach to determine resident preparedness [20]. However, since EPAs are also intended to be formative, summative use would need to be thoroughly examined. Ultimately, we recommend that fellowship programs shift toward a more holistic review process for prospective fellows, paralleling initiatives already taking place at the medical school level [21–23].

In conclusion, the current use of ABSITE scores neither complies with ABS statements nor follows well-accepted testing guidelines. With exception to content evidence, the ABSITE lacks validity evidence in all other domains. It is essential that we recognize these limitations and use the test in an evidence-based manner. With proper use, we believe that the ABSITE can be a powerful learning tool that can help maximize the educational potential of every future surgeon.

Declarations

Conflict of interest No authors have any conflict of interest that is relevant to this paper. No Generative Artificial Intelligence was used in the development of this manuscript.

References

1. American Board of Surgery. https://www.absurgery.org/default.jsp?newsabsite12.23. Accessed 27 June 2023
2. Norcini J, Anderson MB, Bollela V, et al. 2018 Consensus framework for good assessment. Med Teach. 2018;40(11):1102–9. https://doi.org/10.1080/0142159X.2018.1500016.
3. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–7. https://doi.org/10.1046/j.1365-2923.2003.01594.x.
4. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for Educational and Psychological Testing. American Educational Research Association; 2014. https://play.google.com/store/books/details?id=clI_mAEACAAJ
5. Miller AT, Swain GW, Widmar M, Divino CM. How important are American Board of Surgery In-Training examination scores when applying for fellowships? J Surg Educ. 2010;67(3):149–51. https://doi.org/10.1016/j.jsurg.2010.02.007.
6. Savoie KB, Kulaylat AN, Huntington JT, et al. The pediatric surgery match by the numbers: defining the successful application. J Pediatr Surg. 2020;55(6):1053–7. https://doi.org/10.1016/j.jpedsurg.2020.02.052.
7. de Virgilio C, Yaghoubian A, Kaji A, et al. Predicting performance on the American Board of Surgery qualifying and certifying examinations: a multi-institutional study. Arch Surg. 2010;145(9):852–6. https://doi.org/10.1001/archsurg.2010.177.
8. Jones AT, Biester TW, Buyske J, Lewis FR, Malangoni MA. Using the American Board of Surgery In-Training examination to predict board certification: a cautionary study. J Surg Educ. 2014;71(6):e144–8. https://doi.org/10.1016/j.jsurg.2014.04.004.
9. Ray JJ, Sznol JA, Teisch LF, et al. Association between American Board of Surgery In-Training examination scores and resident performance. JAMA Surg. 2016;151(1):26–31. https://doi.org/10.1001/jamasurg.2015.3088.
10. General surgery content outline for the abs in-training examination (absite ®. https://www.absurgery.org/xfer/GS-ITE.pdf. Accessed 15 Jan 2024
11. Gupta S, Jackson JE, Shindorf ML, Grier Arthur L, Chandler N, Danielson P, Downard C, Ehrlich P, Gaines B, Gray B, Javid P, Lallier M, Nwomeh B, Tagge E, Weiss R, Mak G, Garrison AP. Success in pediatric surgery: an updated survey of Program Directors 2020. J Pediatr Surg. 2022;57(10):438–44. https://doi.org/10.1016/j.jpedsurg.2021.10.055.
12. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166.e7-16. https://doi.org/10.1016/j.amjmed.2005.10.036.
13. Smeds MR, Thrush CR, McDaniel FK, Gill R, Kimbrough MK, Shames BD, Sussman JJ, Galante JM, Wittgen CM, Ansari P, Allen SR, Nussbaum MS, Hess DT, Knight DC, Frederick R, Bentley MD. Relationships between study habits, burnout, and general surgery resident performance on the American Board of Surgery In-Training Examination. J Surg Res. 2017;217:217–25. https://doi.org/10.1016/j.jss.2017.05.034.
14. Mullen JT, Cassidy DJ. Let's not throw the baby out with the bath water—keep the ABSITE a numerically scored exam. J Surg Educ. 2021;78(3):714–6. https://doi.org/10.1016/j.jsurg.2020.09.007.
15. Minion D, Plymale M, Donnelly M, Endean E. The effect of prior night call status on the American Board of Surgery In-Training Examination scores: eight years of data from a single institution. J Surg Educ. 2007;64(6):416–9. https://doi.org/10.1016/j.jsurg.2007.06.016.
16. Stone MD, Doyle J, Bosch RJ, Bothe A, Steele G. Effect of resident call status on ABSITE performance. Surgery. 2000;128(3):465–71. https://doi.org/10.1067/msy.2000.108048.
17. Sugar JG, Chu QD, Cole PA, Li BDL, Kim RH. Effect of January vacations and prior night call status on resident ABSITE Performance. J Surg Educ. 2013;70(6):720–4. https://doi.org/10.1016/j.jsurg.2013.06.013.
18. 2021 ABSITE Order Information. https://www.absurgery.org/default.jsp?news_absite12.23. Accessed July 31, 2023
19 Carter B, Sidrak J, Wagner B, Travis C, Nehler M, Christian N. Preliminary development of a program ABSITE dashboard (PAD) to guide curriculum innovation. J Surg Educ. 2024. https://doi.org/10.1016/j.jsurg.2023.10.014.
20. Brazelle M, Zmijewski P, McLeod C, Corey B, Porterfield JR Jr, Lindeman B. Concurrent validity evidence for entrustable professional activities in general surgery residents. J Am Coll Surg. 2022;234(5):938. https://doi.org/10.1097/XCS.0000000000000168.
21. Holistic Review. AAMC. https://www.aamc.org/services/member-capacity-building/holistic-review. Accessed July 31, 2023
22. Witzburg RA, Sondheimer HM. Holistic review–shaping the medical profession one applicant at a time. N Engl J Med. 2013;368(17):1565–7. https://doi.org/10.1056/NEJMp1300411.
23. Sklar DP. Diversity, fairness, and excellence: three pillars of holistic admissions. Acad Med. 2019;94(4):453–5. https://doi.org/10.1097/ACM.0000000000002588.