Using the American Board of Surgery In-Training Examination to Predict Board Certification: A Cautionary Study

Andrew T. Jones, PhD, Thomas W. Biester, MS, Jo Buyske, MD, Frank R. Lewis, MD, and Mark A. Malangoni, MD

American Board of Surgery, Philadelphia, Pennsylvania

OBJECTIVE: Although designed as a low-stakes formative examination, the American Board of Surgery In-Training Examination (ABSITE) is often used in high-stakes decisions such as promotion, remediation, and retention owing to its perceived ability to predict the outcome of board certification. Because of the discrepancy between intent and use, the ability of ABSITE scores to predict passing the American Board of Surgery certification examinations was analyzed.

METHODS: All first-time American Board of Surgery qualifying examination (QE) examinees between 2006 and 2012 were reviewed. Examinees' postgraduate year (PGY) 1 and PGY5 ABSITE standard scores were linked to QE scores and pass/fail outcomes ($n = 6912$ and 6846, respectively) as well as first-time certifying examination (CE) pass/fail results ($n = 1329$). Linear and logistic regression analyses were performed to evaluate the utility of ABSITE scores to predict board certification scores and pass/fail outcomes.

RESULTS: PGY1 ABSITE scores accounted for 22% of the variance in QE scores ($p < 0.001$). PGY5 scores were a slightly better predictor, accounting for 30% of QE score variance ($p < 0.001$). Analyses showed that selecting a PGY5 ABSITE score that maximized overall decision accuracy for predicting QE pass/fail outcomes (86% accuracy) resulted in 98% sensitivity, 13% specificity, a positive predictive value of 87%, and a negative predictive value of 57%. ABSITE scores were not predictive of success on the CE.

CONCLUSIONS: ABSITE scores are a useful predictor of QE scores and outcomes but do not predict passing the CE. Although scoring well on the ABSITE is highly predictive of QE success, using low ABSITE scores to predict QE failure results in frequent decision errors. Program directors and other evaluators should use additional sources of information when making high-stakes decisions about resident performance. (J Surg 71:e144-e148. © 2014 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: certification, in-training exam, qualifying exam, prediction

COMPETENCY: Medical Knowledge

INTRODUCTION

The American Board of Surgery In-Training Examination (ABSITE) is a multiple-choice question examination developed by the American Board of Surgery (ABS) to provide feedback to surgery program directors about the level of progress their residents have attained in applied science and management of clinical problems related to surgery.[1] The ABS does not use ABSITE scores as a requirement in the certification process, and this is intended to be a low-stakes examination that will provide formative feedback. For reasons discussed later, however, the ABSITE results are often used as a tool for high-stakes decisions.

Although the ABSITE does not have a direct effect on board certification, it has been shown to be predictive of ABS Qualifying Examination (QE) performance.[2-4] An objective predictor of performance on the QE can be quite useful to programs directors as residency programs are evaluated on QE performance in the accreditation process. It is a requirement that programs must have at least 65% of graduates pass the qualifying and certifying examinations (CEs) on the first attempt (based on a 5-year average).[5] Thus, program directors may be reluctant to allow residents who may not become certified to advance or remain in a program as their poor performance may have a negative effect on the programs' accreditation status.

Surveys about the use of the ABSITE scores show that surgery residency programs make a variety of "high-stakes" decisions based on poor ABSITE performance. A survey of

Correspondence: Inquiries to Andrew T. Jones, PhD, American Board of Surgery, 1617 JFK Blvd, Suite 860, Philadelphia, PA 19103; fax: (215) 563-5718; e-mail: ajones@absurgery.org

  1931-7204/$30.00 http://dx.doi.org/10.1016/j.jsurg.2014.04.004

TABLE 1. Demographic Characteristics of Overall Sample and Analysis Samples

	Overall Sample ($n = 7372$) *n* (%)	First-Year ABSITE/ QE ($n = 6912$) *n* (%)	Fifth-Year ABSITE/ QE ($n = 6846$) *n* (%)	First-Year ABSITE/ CE ($n = 1344$) *n* (%)	Fifth-Year ABSITE/ CE ($n = 1329$) *n* (%)
Sex*					
Male	5279 (71.6)	4940 (71.5)	4924 (71.9)	903 (67.2)	900 (67.7)
Female	2093 (28.4)	1972 (28.5)	1922 (28.1)	441 (32.8)	429 (32.3)
Medical school†					
USMG	5887 (79.9)	5571 (80.6)	5457 (79.8)	1083 (80.6)	1052 (79.2)
IMG	1447 (20.0)	1334 (19.3)	1382 (20.2)	259 (19.3)	275 (20.7)
Fellowship*					
Yes	5170 (70.1)	4826 (69.8)	4775 (69.7)	1048 (78.0)	1035 (77.9)
No	2202 (29.9)	2086 (30.2)	2071 (30.3)	296 (22.0)	294 (22.1)

IMG, International Medical Graduate; USMG, United States Medical Graduate.
*Significant chi-square differences were observed between the overall sample and the CE samples for sex and fellowship status.
†Missing medical school data for 8 examinees.

more than 200 program directors found that 7% of programs had residents repeat a training year and 6% dismissed residents with low ABSITE scores. Nearly half of the programs had some form of remediation plan in place for low ABSITE scores.[6] A more recent survey (albeit with a lower response rate of 54%) found that 41% of programs had deferred resident promotion based on low ABSITE scores but that none had suspended a resident for ABSITE performance.[7] Other "high-stakes" decisions, such as prohibiting low-scoring residents from taking the ABS QE, are explicit policies in some programs.

Because of the discrepancy between intent (low-stakes formative feedback) and use (high-stakes decisions), the ability of ABSITE scores to predict passing ABS certification examinations was analyzed. Although previous studies have demonstrated a moderate relationship between the ABSITE and surgery QE,[4] no study has addressed the actual decision accuracy of using the ABSITE to predict certification outcomes on a large-scale basis. This study uses national ABSITE, surgery QE, and CE data across multiple years to address the accuracy of using the ABSITE to predict board certification success in surgery.

METHODS

Sample

The 2 criteria in the study for evaluating ABSITE utility were scores and pass/fail outcomes from the QE and CE, the 2 examinations used in the ABS surgery certification process. The QE is a 300-question multiple-choice written examination designed to evaluate knowledge in general surgical principles and applied science.[8] Meanwhile, the CE is an oral examination with the intent to evaluate a surgeon's ability to organize diagnostic evaluations of common surgical problems and determine appropriate therapy.[9]

The scores for first-time QE examinees from 2006 to 2012 were aggregated from ABS archival data. QE results were then linked to each individual's ABSITE scores (QE scores served as the baseline data in the set; ABSITE and CE scores were then merged with QE data). ABSITE scores included both first- and fifth-year results. QE scores were also linked with first-time CE scores and outcomes from 2011 to 2012. There were 7372 examinees who took the QE for the first time from 2006 to 2012. We were able to link fifth-year ABSITE scores for 6846 examinees and first-year ABSITE scores for 6912. There were 1429 first-time CE records that could be linked to the QE scores in the data set. For first-time CE pass/fail outcomes in 2011 to 2012, we were able to link first- and fifth-year ABSITE scores with 1344 and 1329 candidates, respectively. CE pass/fail outcomes from the complete data set were not significantly different from scores in the merged file for first- or fifth-year ABSITE scores. However, there were sex and fellowship participation rate differences between the total sample and the CE samples. Table 1 presents demographic analyses for the full data and the data used for analyses.

ABSITE scores were calculated as a standard score, wherein the mean of the entire resident group was set to 500 with a standard deviation of 100. Despite the popular use of percentile rankings,[7] percentiles are not equal interval scores, and small changes in actual performance near the mean can result in large shifts in percentile scores. Conversely, greater changes near the extremes of the score distribution would typically result in smaller percentile

TABLE 2. Linear Regression of Model Predicting QE Scores From ABSITE Scores

	b	*SEb*	β	R^2	p
First-year ABSITE scores					
Constant	64.39	0.39			<0.001
ABSITE	0.04	0.00	0.47	0.22	<0.001
Fifth-year ABSITE scores					
Constant	58.65	0.42			<0.001
ABSITE	0.04	0.00	0.55	0.30	<0.001

SE, standard error.