Association Between American Board of Surgery In-Training Examination Scores and Resident Performance

Juliet J. Ray, MD; Joshua A. Sznol, BS; Laura F. Teisch, BS; Jonathan P. Meizoso, MD; Casey J. Allen, MD; Nicholas Namias, MD, MBA; Louis R. Pizano, MD, MBA; Danny Sleeman, MD; Seth A. Spector, MD; Carl I. Schulman, MD, PhD, MSPH

IMPORTANCE The American Board of Surgery In-Training Examination (ABSITE) is designed to measure progress, applied medical knowledge, and clinical management; results may determine promotion and fellowship candidacy for general surgery residents. Evaluations are mandated by the Accreditation Council for Graduate Medical Education but are administered at the discretion of individual institutions and are not standardized. It is unclear whether the ABSITE and evaluations form a reasonable assessment of resident performance.

OBJECTIVE To determine whether favorable evaluations are associated with ABSITE performance.

DESIGN, SETTING, AND PARTICIPANTS Cross-sectional analysis of preliminary and categorical residents in postgraduate years (PGYs) 1 through 5 training in a single university-based general surgery program from July 1, 2011, through June 30, 2014, who took the ABSITE.

EXPOSURES Evaluation overall performance and subset evaluation performance in the following categories: patient care, technical skills, problem-based learning, interpersonal and communication skills, professionalism, systems-based practice, and medical knowledge.

MAIN OUTCOMES AND MEASURES Passing the ABSITE (≥30th percentile) and ranking in the top 30% of scores at our institution.

RESULTS The study population comprised residents in PGY 1 (n = 44), PGY 2 (n = 31), PGY 3 (n = 26), PGY 4 (n = 25), and PGY 5 (n = 24) during the 4-year study period (N = 150). Evaluations had less variation than the ABSITE percentile (SD = 5.06 vs 28.82, respectively). Neither annual nor subset evaluation scores were significantly associated with passing the ABSITE (n = 102; for annual evaluation, odds ratio = 0.949; 95% CI, 0.884-1.019; P = .15) or receiving a top 30% score (n = 45; for annual evaluation, odds ratio = 1.036; 95% CI, 0.964-1.113; P = .33). There was no difference in mean evaluation score between those who passed vs failed the ABSITE (mean [SD] evaluation score, 91.77 [5.10] vs 93.04 [4.80], respectively; P = .14) or between those who received a top 30% score vs those who did not (mean [SD] evaluation score, 92.78 [4.83] vs 91.92 [5.11], respectively; P = .33). There was no correlation between annual evaluation score and ABSITE percentile (r² = 0.014; P = .15), percentage correct unadjusted for PGY level (r² = 0.019; P = .09), or percentage correct adjusted for PGY level (r² = 0.429; P = .91).

CONCLUSIONS AND RELEVANCE Favorable evaluations do not correlate with ABSITE scores, nor do they predict passing. Evaluations do not show much discriminatory ability. It is unclear whether individual resident evaluations and ABSITE scores fully assess competency in residents or allow comparisons to be made across programs. Creation of a uniform evaluation system that encompasses the necessary subjective feedback from faculty with the objective measure of the ABSITE is warranted.
The American Board of Surgery In-Training Examination (ABSITE) is an annual multiple-choice examination used to assess medical and applied knowledge of general surgery residents. It was originally designed as a tool for program directors to assess residents' progress and is not a requirement for certification. Studies show that ABSITE performance is predictive of passing the American Board of Surgery Qualifying Examination; therefore, program directors turn to this tool as an objective comparable measure. Extensive variability exists between how scores are used by each program and whether they affect promotion. A uniform standard is not determined by the Accreditation Council for Graduate Medical Education (ACGME). Regardless of the design of the ABSITE, this examination is now often used in manners beyond its original intent owing to a lack of other standardized evaluation techniques.

Semiannual review of residents is mandated and outlined by the General Surgery Milestone Project, a joint initiative of the ACGME and the American Board of Surgery. Evaluations are administered by the individual institutions depending on resident rotations. At our university-affiliated hospital, a comprehensive rotation-specific evaluation system has been used to assess essential qualities not readily testable on standardized examinations. The 7 categories evaluated are adapted from the General Surgery Milestone Project, including the following 6 core competencies in addition to technical skill: medical knowledge, patient care, interpersonal and communication skills, professionalism, practice-based learning, medical knowledge, interpersonal and communication skills, professionalism, and systems-based practice. End-of-rotation evaluations in these competencies are a common approach at many institutions, but the structure is variable. Furthermore, training of faculty on the use of the scoring system as well as their education in knowledge or skills assessment of residents is not uniform.

Studies on the relationship between evaluations and board performance are limited and controversial. In the medical student population, one study showed high interrater reliability between resident and attending evaluations of medical students but poor correlation with standardized examination scores on the surgical clerkship. Conversely, another study showed a strong positive correlation between ward evaluation and National Board of Medical Examiners examination performance. To our knowledge, only 2 studies have evaluated this phenomenon in surgical residents. In regard to resident medical knowledge, one showed that there was poor correlation with ABSITE performance and that faculty evaluations cannot predict residents who will perform poorly. It should be noted that this study looked only at the relationship between medical knowledge assessed on evaluation and ABSITE score. The other showed that when using a standardized assessment of ACGME core competencies, faculty ratings were internally consistent and correlated with ABSITE and United States Medical Licensure Examination scores.

Our study aims to add to the limited body of literature regarding the relationship between evaluations and ABSITE scores. It is unclear whether the ABSITE and evaluations form a reasonable assessment of residents. We hypothesize that there is no relationship between rotation-specific evaluation and ABSITE scores as measures of resident performance.

Methods

A retrospective study of deidentified evaluation and ABSITE scores was conducted at our institution from July 1, 2011, through June 30, 2014. Evaluations in postgraduate years (PGYs) 1 through 5 were reviewed for each rotation completed during that academic year, ranging from 6 to 11 evaluations per resident depending on PGY. A mean annual score and a mean score for each of the 7 evaluation subsets were calculated for each resident: patient care, technical skill, practice-based learning, medical knowledge, interpersonal and communication skills, professionalism, and systems-based practice. Evaluations are completed by faculty through the New Innovations Residency Management Suite online system. Residents are required to review these nonanonymous evaluations semiannually with the program director. Data for each resident, separated by PGY at the time of testing, were acquired from the evaluation and ABSITE reports for the academic year corresponding to the examination. This study was approved by the Institutional Review Board of the University of Miami. Informed consent was waived owing to the nature of the study, which involved a retrospective analysis of previously collected and stored deidentified data.

All data were analyzed in SAS version 9.3 statistical software (SAS Institute, Inc). The type I error rate was set to 5% and $P < .05$ was considered statistically significant. For continuous variables, normally distributed data are reported as mean (standard deviation). Continuous variables were compared with $t$ test for parametric data. Evaluations were compared with the corresponding year's ABSITE score. The ABSITE percentile and percentage correct scores were considered in the analyses. Binary outcomes included passing the ABSITE with a score above or equal to the national 30th percentile as well as ranking in the top 30% of all scores at our institution during the combined years of the study. Multivariable regression was performed to predict passing the ABSITE or achieving a top 30% score. Data were also analyzed in terms of percentage correct score adjusted for PGY. All components of the ABSITE and all evaluation subsets were regressed to determine whether an association existed.

Results

The population comprised residents in PGY 1 ($n = 44$), PGY 2 ($n = 31$), PGY 3 ($n = 26$), PGY 4 ($n = 25$), and PGY 5 ($n = 24$) during the 4-year study period ($N = 150$). One hundred fifty ABSITE scores and 1131 evaluations were included for analysis. The distribution of resident scores for annual evaluation and ABSITE scores is as follows: mean ($SD$), 92.24 [5.06]; median, 92.65; interquartile range, 7.82; ABSITE percentile (mean $SD$), 49.18 [28.82]; median, 49.50; interquartile range, 53.00; and ABSITE percentage correct (mean $SD$), 72.51 [8.07]; median, 42.00; interquartile range, 53.00; and ABSITE percentage correct.
range, 11.00) are shown in Figure 1. Evaluations had less variation compared with the ABSITE percentile (SD = 5.06 vs 28.82, respectively). Overall, there was no correlation between annual evaluation score and ABSITE percentile ($r^2 = 0.014; P = .15$), percentage correct unadjusted for PGY level ($r^2 = 0.019; P = .09$), or percentage correct adjusted for PGY level ($r^2 = 0.429; P = .91$).

On binary logistic regression, the annual evaluation score was not significantly associated with passing the ABSITE (odds ratio = 0.949; 95% CI, 0.884-1.019; $P = .15$) or receiving a top 30% score (odds ratio = 1.036; 95% CI, 0.964-1.113; $P = .33$). There was no significant relationship between annual or any of the subset evaluation scores and ABSITE scores adjusted for PGY on multivariable linear regression. Figure 2A and B show the lack of correlation between annual or medical knowledge evaluations and ABSITE percentile by PGY, and Figure 2C and D show this for ABSITE percentage correct. No evaluation subset scores were predictive of passing the ABSITE on regression models.

There was no difference in mean evaluation score between those who passed vs failed the ABSITE (mean [SD] evaluation score, 91.77 [5.10] vs 93.04 [4.80], respectively; $P = .14$) (Figure 3) or between those who received a top 30% score vs those who did not (mean [SD] evaluation score, 92.78 [4.83] vs 91.92 [5.11], respectively; $P = .33$) (Figure 4).

**Discussion**

In the times of William Stewart Halsted, MD, surgical residents were continuously and rigorously evaluated through an apprenticeship model that was perhaps more reflective of true medical knowledge and skill than the prototype we have today. The modern-day residency model has shifted—residents frequently rotate between services and work with a multitude of attending surgeons. Perhaps now more than ever, the need for a comprehensive evaluation is even more critical, especially to identify struggling residents early in their training experience. A reform like this may help create a standard by which to compare residents during the fellowship application and promotion process but would require faculty education and “buy in” to ensure that the evaluation method is consistent within and across programs.

Our study found that favorable evaluation scores could not be used to predict ABSITE performance and that there was no difference in mean evaluation scores between those who passed and failed the ABSITE. We also showed that evaluations have low variability, which implies that they are not being used to their full potential to assess a resident’s competency and standing among their peers. These findings illustrate a concerning deficiency in the way we currently evaluate surgical trainees. The creation of a standardized, valid, and reliable system of evaluation is imperative to the future of surgical education.

The ACGME has required residency programs to base evaluations on the 6 core competencies for more than a decade, but the scales used are diverse. In addition to the structured variability inherent in these evaluations, we know that outside factors, such as personality, influence evalua-
Regression models of ABSITE percentile and annual (A) and medical knowledge (B) evaluation scores by PGY, and regression models of ABSITE percentage correct and annual (C) and medical knowledge (D) evaluation scores by PGY.
These results show promise that a standardized and reliable system can be implemented across institutions.

Prior to 2014, there was a junior version and a senior version of the ABSITE. The content differed on these examinations based on resident PGY. As of 2014, the ABSITE structure changed so that all residents now receive a single examination that is compared nationally against residents of the same PGY.1 The 2014 ABSITE provided overall percentage correct and percentile scores in addition to percentage correct scores in the subcategories of patient care and medical knowledge. The 2011 to 2013 examinations also reported percentage correct scores for individual organ system subsets. For the purposes of our study, only overall percentage correct and percentile scores were considered, to account for the change in structure of the test.

Our study should be considered in the context of certain limitations. First, the possibility of type II error exists as a sample size of 150 may not be large enough to detect an association. The change in structure of the ABSITE during our study period is also a limitation, as test results were evaluated together across this period. Furthermore, the status of each resident (ie, categorical vs preliminary) was not factored into the analysis. Finally, we are unable to control for the inconsistency in completing the evaluations by faculty members, which may contribute to decreased internal validity. A limitation on external validity should also be considered, as other programs’ systems for evaluating residents have their own sets of strengths and weaknesses.

Conclusions
Favorable evaluations do not correlate with ABSITE scores. Evaluations do not show discriminatory ability. It is unclear whether resident evaluations and ABSITE scores fully assess competency in residents or whether these tools allow comparisons to be made across programs. Creation of a uniform evaluation system that encompasses the necessary feedback from faculty with the objective measure of the ABSITE is warranted. This tool will be vital to create a fair and effective method to determine resident promotion and ensure timely intervention for residents with deficiencies. Furthermore, it would potentially allow evaluation of surgery programs across institutions and influence surgical training paradigms.

ARTICLE INFORMATION
Accepted for Publication: June 8, 2015.

Author Contributions: Drs Spector and Schulman had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Dr Ray is the first author and Dr Schulman is the senior author.
Study concept and design: Ray, Meizoso, Allen, Namias, Pizano, Spector, Schulman.
Acquisition, analysis, or interpretation of data: Ray, Sznol, Teisch, Meizoso, Allen, Namias, Selleman, Spector, Schulman.
Conflict of Interest Disclosures: None reported.
Previous Presentation: This paper was presented at the 39th Annual Meeting of the Association of VA Surgeons; May 3, 2015; Miami Beach, Florida.

Drafting of the manuscript: Ray, Sznol, Teisch, Meizoso, Schulman.
Critical revision of the manuscript for important intellectual content: Teisch, Meizoso, Allen, Namias, Pizano, Selleman, Spector, Schulman.
Administrative, technical, or material support: Namias, Spector.
Study supervision: Allen, Namias, Pizano, Selleman, Spector, Schulman.

Additional Contributions: Tanya Spencer, Leela Mundra, BA, and Manasa Narasimman, University of Miami, Miami, Florida, provided assistance in data collection; they received no compensation.

REFERENCES
The Continuing Quest for Meaningful Faculty Evaluations of Residents

John L. Tarpley, MD; Margaret J. Tarpley, MLS

In this issue of JAMA Surgery, Ray et al1 report their findings at their institution, the University of Miami, Miami, Florida, on the (non)relationship between performance on the American Board of Surgery In-Training Examination (ABSITE) and the mean scores of faculty evaluations. Several thoughts and conjectures stem from their study.

First, the authors cite their use of the measures from the General Surgery Milestone Project as their base for evaluating in the New Innovations Residency Management Suite online system for the study period 2011 to 2014; however, the General Surgery Milestone Project was under development in 2011 and only became effective for general surgery in July 2014. The 6-monthly resident performance review and conversation with the program director or associate program director have been part of the Accreditation Council for Graduate Medical Education Surgical Residency Review Committee program requirements for decades.

Second, their faculty evaluations were nonanonymous. Many of our own faculty, citing that they will be working with residents over years, are leery of providing less than laudatory comments (constructive criticism) even anonymously because they believe residents can identify the source of the less-than-A+ evaluation. They are not keen on nonanonymous evaluations. Of note, New Innovations now provides a way for faculty to make a “not for the resident to see” comment in a supplemental box at the bottom of the evaluation form. The residents in the study by Ray and colleagues were mostly perceived as “A” performers (mean evaluation score, 92.24; median evaluation score, 92.65), as shown in the histogram (Figure 1A) with all annual evaluation scores in the top quartile, while ABSITE scores varied more.

Finally, one can argue that the ABSITE and the faculty evaluations measure different domains, not unlike the qualifying examination and the certifying examination for the American Board of Surgery, and that they complement one another rather than necessarily predict or correlate with one another. Some irascible residents ace the ABSITE, while other residents with excellent professional demeanor and people skills may struggle with multiple-choice questions. Book smarts and street smarts do not have an r value of 1, nor do multiple-choice question skills and the soft competencies of interpersonal and communication skills and professionalism. Should we expect an examination that measures 1 portion of 1 domain, medical knowledge, to correlate with a subjective evaluation measuring 6 or 7 competency domains?

This article challenges us to improve our evaluations of resident performance, a worthy quest of an elusive target. The next major hurdle is evaluation of technical and operative skills.