Validity and Interrater Reliability of a Regional Mock Oral Board Examination

John L. Falcone, MD,* Daniel J. Gagne, MD,† Kenneth K.W. Lee, MD,* and Giselle G. Hamad, MD,*

*Department of Surgery, University of Pittsburgh School of Medicine, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania; and †Department of Surgery, West Penn Allegheny Health System, Pittsburgh, Pennsylvania

BACKGROUND: There have been decreasing pass rates recently on the American Board of Surgery Certifying Examination (ABSCE). General surgery residents from the University of Pittsburgh Medical Center, the West Penn Allegheny Health System, the Conemaugh Health System, and Mercy Hospital participate in a mock oral board examination, which is similar to the ABSCE. The aims of the study are to compare examinee performance on the mock oral boards with the ABSCE and to evaluate the interrater reliability of examiner pairs.

METHODS: In this retrospective study from 2003 to 2010, outcomes on the mock oral boards and the first attempt of the ABSCE for chief residents were compared for the 4 regional residency programs. Interrater reliability for examiner pairs was evaluated with agreement and kappa statistics. Nonparametric statistics were performed, with $\alpha = 0.05$.

RESULTS: A total of 32 of 38 (84.2%) chief residents passed the mock oral boards. The median score for each of the 3 rooms was 6 (clear pass). A total of 37 of 38 (97.4%) residents passed the ABSCE. The sensitivity of the mock oral boards was 83.8%, with a positive predictive value of 96.9%, and an accuracy of 81.6%. A total of 25 of 47 (53.2%) examiner pairs were from the same residency institution, whereas 22 of 47 (46.8%) were from different institutions. The median agreement was 100% (interquartile range (IQR) [100% - 100%]). The median kappa statistic was 1.00 (IQR [0.38-1.00]). The Mann-Whitney *U* tests showed no difference in agreement or kappa for examiner pairs from the same or from different institutions ($p > 0.05$).

CONCLUSIONS: The mock oral boards have substantial sensitivity and positive predictive value in relation to the ABSCE. There are also very high levels of interrater agreement and interrater reliability. This regional mock oral board examination is valuable for ABSCE preparation. (J Surg 70:402-407. © 2013 Published by Elsevier Inc. on behalf of the Association of Program Directors in Surgery)

KEY WORDS: certification, educational measurement, general surgery, regional medical programs, reliability and validity, specialty boards

COMPETENCIES: Medical Knowledge, Patient Care, Interpersonal Skills and Communication

INTRODUCTION

The American Board of Surgery Certifying Examination (ABSCE) is an oral examination consisting of 3 consecutive sessions conducted by a team of 2 examiners. Its purpose is to evaluate a candidate's clinical skills in organizing the diagnostic evaluation of common surgical problems and determining appropriate therapy. Passing this examination is necessary for becoming certified by the American Board of Surgery.[1]

The first-time pass rates on the ABSCE have now become one of the standards of excellence to evaluate residency programs.[2] Importantly, there has been a significant decline in the pass rate of the ABSCE in recent years. In 2006, 84% of the 1093 examinees passed the ABSCE. In 2012, 72% of the 1356 examinees passed the ABSCE.[3,4] This decline is statistically significant through proportion and linear regression analyses, and the etiology of this important trend in surgeon certification is unclear.[3,5]

One potential method of improving residency program performance on the ABSCE is by giving a mock oral board examination.[6] Administering monthly mock oral board examinations has prospectively been shown to improve first-time pass rates on the ABSCE.[2] Moreover, administering mock oral board examinations with feedback from observers can improve performance on the ABSCE.[2,7] It is subjectively believed by program graduates that the mock oral examination helped in the preparation for the

Correspondence: Inquiries to John L. Falcone, MD, MS, University of Pittsburgh Medical Center, Presbyterian University Hospital, F-675, 200 Lothrop Street, Pittsburgh, PA 15213. Fax: 412-647-4889; e-mail: falconej@upmc.edu

  1931-7204/$30.00 http://dx.doi.org/10.1016/j.jsurg.2012.12.004

ABSCE.[2,8] As a corollary, it has been shown that larger residency programs have higher pass rates than smaller programs on the ABSCE.[9]

With the published literature on mock oral board examinations, the sensitivity, specificity, positive predictive value, and negative predictive value of these mock oral board examinations are unknown. There is a paucity of literature available regarding interrater reliability in that the agreement between paired examiners has not been reported.

In Central/Western Pennsylvania, general surgery residents at the postgraduate year 4 and 5 levels participate in a regional mock oral board examination, which is run in a similar fashion to the ABSCE. The examiners and the examinees are from the following centers: the University of Pittsburgh Medical Center (Pittsburgh, PA), the West Penn Allegheny Health System (Pittsburgh, PA), the Conemaugh Health System (Johnstown, PA), and Mercy Hospital (Pittsburgh, PA). To keep a degree of unfamiliarity between the examiners and examinees, residents from one program are examined by a team of surgeon examiners from other institutions. Examiners participate on a voluntary basis, and minimal examiner training is performed before the mock oral board examination. The location of the regional mock oral board examination changes on an annual basis. Chief resident examinees are videotaped, and the video performance can be reviewed for feedback purposes. All examinees receive written scores, case-specific comments, and overall individualized feedback comments on testing style.

With the decline in the ABSCE pass rate, and the potential benefit of mock oral board examinations, we are interested in the mock oral board examination as it relates to the first-time attempt at the ABSCE. We hypothesize that the sensitivity, specificity, positive predictive value, and negative predictive value would be above 80%. We also want to describe the interrater reliability of paired surgeon examiners. We hypothesize that the agreement between examiner pairs would be above 80%.

METHODS

In this retrospective study, resident performance on the regional mock oral board examination and the ABSCE was reviewed from 2003 to 2010. Chief general surgery residents who had data available for all rooms in the regional mock oral board examination and took the ABSCE within a year following graduation were eligible for the study. This collaborative study was approved by the institutional review board at each of the 4 regional residency program sites.

Performance on the mock oral board examination was described for each room and overall. The binary outcome of the mock oral board examination (pass/fail) was determined by adding all of the overall room scores for each of the 6 examiners. Residents passed the mock oral board examination if the aggregate total was ≥ 32; this is comparable to the scale and aggregate total used on the ABSCE.

Using the outcome of the first attempt of the ABSCE as the gold standard, the binary outcome of the regional mock oral board examination and the ABSCE was tabulated to derive a sensitivity, specificity, positive predictive value, and negative predictive value of the regional mock oral board examination.

To check for adequate resident sampling, the number of included resident examinees and their ABSCE outcomes was compared with the population of first-time examinees and their outcomes from 2005 to 2010 with a chi-square test or the Fisher exact test for each institution. First-time examinee outcomes on the ABSCE from 2005 to 2010 were available from the American Board of Surgery website.[10] Permission to use these previously published data was confirmed with a personal communication with the Director of Psychometrics and Data Analysis.

To evaluate interrater reliability of paired surgeon examiners, examinee performance evaluations were used to calculate a level of agreement and a kappa statistic for each examiner pair. Examinee performance in each room for the regional mock oral board examination was evaluated using the following ordinal evaluation scale: 3—critical fail; 4—clear fail; 5—indeterminate; 6—clear pass; and 7—high pass. Rooms where there were 2 clearly named examiners were included in the study. There were subgroups of examiner pairs from the same institution and from different institutions, both of which were different from the examinees tested by the examiner pair. The Mann-Whitney U test was performed to test for differences in agreement and kappa statistics between examiners from the same institution and examiners from different institutions. All statistical tests were performed with Stata 11.1 statistical software (StataCorp, College Station, TX), using an $\alpha = 0.05$.

RESULTS

From 2003 to 2010, there were 38 chief residents (2 women) who had any data available for the regional mock oral board examination. All (100%) of these eligible chief residents satisfied the inclusion criteria for the study. A binomial test showed that there were more male chief residents than female chief residents ($p < 0.001$). There were 22 (58%) residents from the University of Pittsburgh Medical Center, 12 (32%) residents from the West Penn Allegheny Health System, 3 (8%) residents from the Conemaugh Health System, and 1 (3%) resident from Mercy Hospital.

The median score on the mock oral board examination was 34 (IQR [32-36], range [28-40]). A histogram of mock oral board examination performance is shown in Figure 1. A total of 32 of 38 (84.2%) chief residents passed the mock