Concurrent Validity of the Osteopathic General Surgery In-Service Examination

John L. Falcone, MD, MS
Marc E. Rosen, DO

From Owensboro Health Surgical Specialists in Kentucky (Dr Falcone) and the American College of Osteopathic Surgeons in Alexandria, Virginia (Dr Rosen).

Financial Disclosures: None reported.

Address correspondence to John L. Falcone, MD, MS, Owensboro Health Surgical Specialists, Ridgecrest Medical Park, 2801 New Hartford Rd, Owensboro, KY 42303-1320.

E-mail: John.FalconeMD@owensborohealth.org

Submitted May 30, 2013; final revision received September 14, 2013; accepted October 28, 2013.

Context: Performance on the Osteopathic General Surgery In-Service Examination (ISE) has been shown to improve over time for osteopathic general surgery residents. The training level–specific concurrent validity of the ISE, however, has not been evaluated.

Objective: To investigate whether residents' scores will improve as they move from level 1 through level 5 of the ISE.

Methods: In this retrospective study, performance on the ISE was obtained from the American College of Osteopathic Surgeons for all of the osteopathic general surgery residency programs from 2008 through 2012. The weighted raw score and standardized score performance mean and standard deviation were determined across training levels. One-way *t* tests were performed between residency years and ISE scores. Parametric statistics were calculated with α set to .05.

Results: The authors evaluated 1952 examinations during the study period. Of the 49 programs screened, 33 (67.3%) met inclusion criteria for the present study. Analysis of variance tests showed that there was significant variation in raw and standardized outcomes between residency levels (both $P<.001$). One-tailed *t* tests for both raw and standardized outcomes showed that all scores' differences between examinee levels were statistically significant ($P<.001$), with the exception of raw scores between level 4 and level 5 examinees ($P=.20$).

Conclusion: There is near-uniform concurrent validity of the ISE by osteopathic general surgery training level. This psychometric characteristic supports the construct validity of this standardized test.

J Am Osteopath Assoc. 2014;114(4):267-272
doi:10.7556/jaoa.2014.052

Candidates seeking to be accredited in osteopathic general surgery must achieve a satisfactory level of performance on the written American Osteopathic Board of Surgery (AOBS) Certifying Examination.[1] The American College of Osteopathic Surgeons' general surgery In-Service Examination (ISE), similar to the AOBS Certifying Examination, is offered to all residents in American Osteopathic Association–approved, general surgery training programs.[2] The ISE—a computer-based examination composed of 300 type-A multiple choice, criterion-referenced items—is used to assess a candidate's knowledge of the tenets of general surgery–patient evaluation and management. The examination is given during all years of residency training. Individual achievement is independent of the performance of others, with no set score for passing.[2]

There is a scarcity of published literature on ISE performance. The psychometric properties of the examination make it a reliable test for measuring educational progress during general surgery training. The examination represents a valid construct, and in a previous study, Shen[3] observed growth trajectories over several cohorts of residents, noting that scores improved over time. Available since the 2007-2008 academic year, the ISE presented examinees with a new, more standardized curriculum.[4] Influences of this curriculum on examination performance have not, to our knowledge, been studied.

Although the ISE is a reliable construct, the concurrent validity of the ISE is unknown with regard to resident level. Concurrent validity should allow investigators to assess the ability of an operationalized construct, which in turn would enable them to distinguish differences among groups.[5] In addition, certification boards have the responsibility to ensure that examinations are both as reliable and as valid as possible.[6] The purpose of the present study was to evaluate the concurrent validity on ISE performance as it occurs across residency training levels. We hypothesized that a graduated performance pattern would be observed between residency postgraduate groups. Further, we hypothesized that residents' scores from level 1 to level 5 will demonstrably improve with each successive year.

Methods

For the present retrospective cohort study, we obtained the raw performance scores and standardized performance scores on the ISE for all US osteopathic general surgery residency programs from 2008 through 2012 for all training levels. These data were obtained from the director of ACOS Postdoctoral Training Standards and Evaluation. Permission to perform this study was granted by the chairman of the General Surgery In-Service Examination Committee. (Because of its retrospective design, the present study was considered exempt from Institutional Review Board approval.) Data reported by the ACOS included the number of examinees at each residency level, the mean raw scores, and the mean standardized scores for each year.

Residency programs were selected for inclusion into this study if data were available for each year of the study period and if each program had data for each training level during the study period. This selection method was chosen to limit the study to established programs. For each program, the raw scores and standardized performance scores by level of training were extracted from the data set. Independent double entry was performed to ensure accuracy of the extracted data. Performance scores by level were collated for each year of the study period. Resident examinees who took the ISE during the study period were evaluated in a longitudinal fashion: for example, we evaluated the scores of examinees who were at level 1 in 2008, then at level 2 in 2009.

The weighted raw score and standardized performance score mean and standard deviation were determined for all levels. An analysis of variance test was first performed to determine whether there were differences in raw scores and standardized scores among residency training levels. A post hoc Bonferroni correction was performed to correct for type I error. To further test concurrent validity, 1-tailed *t* tests were performed for raw scores and standardized scores between residency levels.

Simple linear regression was also performed using the academic year as the independent variable and the overall weighted level-specific score as the dependent variable. This test was performed to identify any trends in global performance across years of the study. In addition, we extrapolated the data for the cohort of examinees that started at level 1 in 2008 (ie, the cohort for which results were available across all 5 years of the study). Simple linear regression was performed using the year as the independent variable and score as the dependent variable to calculate if examinees demonstrated increasing scores during residency training. All statistics were performed using Stata 11.1 statistical software (StataCorp), with α set to .05.

Results

From 2008 to 2012, there were 49 general surgery residency training programs that had data available for review; 33 programs (67.3%) provided data for each year of the study period, and all programs retained complete level-specific data.

A total of 1952 examinations were evaluated from the 33 qualifying programs that met the inclusion criteria. The median (interquartile range) number of examinees per program during the study period was 50, (31-78). The median (interquartile range) number of level 1, 2, 3, 4, and 5 examinees per program was 12 (8-17), 11 (8-17), 11 (6-16), 10 (7-15), and 7 (5-14), respectively.

The *Table* shows the raw scores and standardized scores for all general surgery residency training programs across all levels. Analysis of variance tests showed a significant variation in raw scores and standardized scores across residency levels (both $P<.001$). One-tailed t tests for both raw scores and standardized scores showed that all differences between examinee levels were statistically significant ($P<.001$), with the exception of raw scores between level 4 and level 5 examinees ($P=.20$).

The simple linear regression analysis of performance by level over time is shown in *Figure 1*. There were no significant linear trends in raw scores for examinees, as follows: level 1 ($P=.09$, linear regression correlation=0.82); level 2 ($P=.09$, linear regression correlation=0.82), level 3 ($P=.10$, linear regression correlation=0.81), level 4 ($P=.10$, linear regression correlation=0.81), and level 5 ($P=.20$, linear regression correlation=0.69). There were no significant linear trends seen in standardized scores across all years (all $P>.05$; linear regression correlation range, −0.24 to >0.99).

The longitudinal ISE scores of examinees from 2008 to 2012 are shown in *Figure 2*. The slope of the linear regression line for raw score and standardized score was greater than 0 (both $P<.01$). The linear regression correlation between raw examination score and year was 0.97 (data not shown). The linear regression correlation between standardized examination score and year was also 0.97.

Discussion

The main objective of the present study was to evaluate the concurrent validity of the ISE with regard to training level. We hypothesized that a hierarchy would be observed wherein resident scores would increase between levels every year from level 1 to level 5. The ISE has a near-uniform degree of concurrent validity, as demonstrated by its ability to distinguish resident examinees by training level. The test scores over time show a graduated distribution in mean performance from level 1 to level 5 in raw scores and in standardized scores. These findings directly support our study hypothesis. This characteristic of concurrent validity is very important when considering the psychometric properties of such a standardized test. Of note, although raw scores for level 5 examinees were higher than those for level 4 examinees, there was

Table.
Longitudinal Raw Scores and Standardized Scores on the Osteopathic General Surgery In-Service Examination From 2008 to 2012, mean (standard deviation)

Level					
Score Type	1 (n=424)	2 (n=457)	3 (n=393)	4 (n=545)	5 (n=555)
Raw	157.3 (14.0)	165.3 (12.4)	176.1 (12.5)	181.8 (12.9)	182.4 (10.8)
Standardized	435.7 (3.8)	477.3 (5.9)	517.0 (6.3)	545.1 (6.0)	555.2 (6.7)

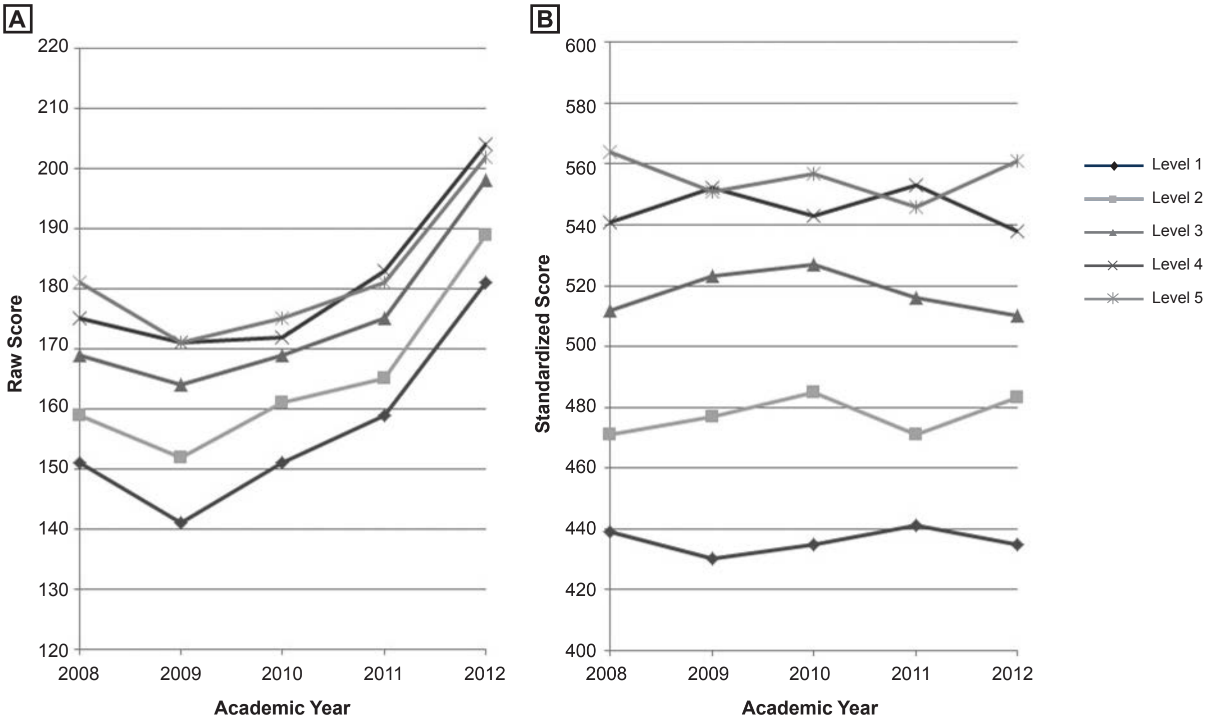

Figure 1.
Longitudinal raw scores (A) and standardized scores (B) on the Osteopathic General Surgery In-Service Examination by training level from 2008 to 2012 for 33 general surgery residency training programs.

no statistically significant difference in raw scores between these 2 groups. This finding contrasts with the results of the standardized scores. This discordance between raw and standardized outcome measures suggests that concurrent validity of the ISE may not occur in the last 2 years of residency training. This finding is consistent with a knowledge plateau toward the end of general surgery residency training. The present study is the first to our knowledge to specifically evaluate the concurrent validity of the ISE as reflected by training level of osteopathic general surgery residents.

The raw score trends over time are shown in *Figure 1*. There have been overall increasing performance trends across recent years, although these trends did not reach a level of statistical significance. The reasons behind these trends are unclear because the overall difficulty of the examination questions is similar from year to year.

There are a few possible explanations for this trend. First, the study resources and curricula may be improving over time. The new curriculum was introduced and made available to the cohort of osteopathic residents in the 2007-2008 academic year; this new curriculum may have guided study habits and resulted in improved performance over all levels. The new curriculum may have also altered the attitudes and actions of faculty members and may have changed the overall teaching milieu. Second, the cohort of level 1 examinees may have improved in 2009, with a migration effect of raw performance improvement in subsequent years.

Further studies may focus on testing whether a migration effect occurs between general surgery residency applicants and standardized test performance. Additionally, in the 2008-2009 academic year—the year that training was a part of the residency program structure—

level 1 examinees had just completed a residency year and were taking the examination with level 2 residents who had just completed a rotating internship. This effect also may have contributed to the upward trajectory shown in *Figure 1*. Future studies might also evaluate the effect of the introduction of a computerized version of the examination during the 2009-2010 academic year. Finally, the trends may also represent the normal variation in examinee performance over time. These trends should continue to be studied in the future.

Additionally, there is global improvement of resident scores in the cohort taking the level 1 examination in 2008 over time, a finding that is consistent with the educational progress found in a 2000 study by Shen.[3] The linear regression correlation of 0.97 is substantial and suggests that resident knowledge improves during the course of osteopathic general surgery residency training. This finding further supports the construct validity of the ISE.

In allopathic general surgery training, performance on the American Board of Surgery In-Training Examination (ABSITE) is a well-described indicator of performance on the written American Board of Surgery Qualifying Examination required for surgeon certification.[7-10] We can only speculate that there is a similar relationship between the ISE and the AOBS Certifying Examination. A worthwhile future study could explore the predictive value of ISE scores and outcomes on the AOBS Certifying Examination.

There are a number of limitations to the present study. First, its retrospective nature makes it difficult to conclude any cause-and-effect relationship between residency level and ISE scores. However, despite this limitation the concurrent validity of the ISE could be adequately studied. Also, roughly one-third of the residency programs were not included in the analyses. The inclusion criteria, however, allowed us to select the established residency programs that were present during each year of the study period. Residency programs that have closed or that are new may not have well-established curricula, teaching faculty, or learning resources for resident examinees. Another limitation is that studying examination results, arguably, does not matter as much as patient care, real-time decision-making capabilities, or AOBS Certifying Examination results. Studies of the relationship between the ISE and the AOBS Certifying Examination are needed. Such studies should take into account the possibility that examinees improve their scores because they grow accustomed to the examination year after year. For the purposes of the present study, we hypothesized that scores improved as a result of increased exposure in the field and increased knowledge as opposed to examination construct familiarity.

On the other hand, the present study has numerous strengths. The data comprise performance for established

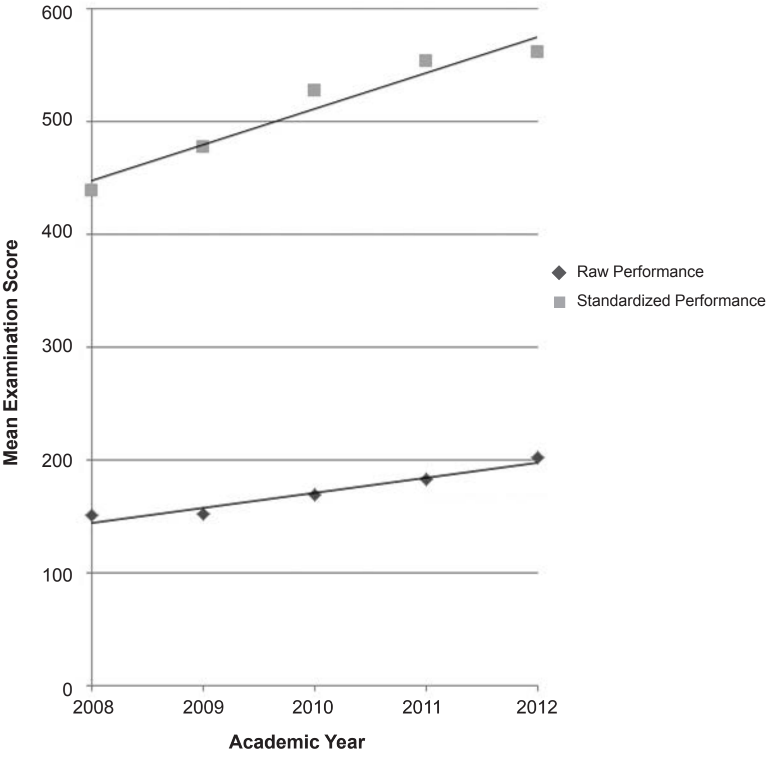

Figure 2.
Longitudinal standardized In-Service Examination performance for residents of 33 general surgery residency training programs who started at level 1 in 2008.

programs over a 5-year study period, establishing longitudinal validity for the study's use of the total population of examinees, as opposed to a small cohort of single-institution or regional examinees. The performance on the ISE is evaluated outright, as opposed to binomial examination outcomes (pass/fail). The straightforward manner in which we conducted the methodology and calculated the statistics could easily be replicated by other researchers. Importantly, we are the first researchers, to our knowledge, to show the relationship between the introduction of the new curriculum and the examination outcomes.

Conclusion

Overall, there is level-specific concurrent validity of the ISE regarding raw scores and standardized scores by residents in osteopathic general surgery programs. This psychometric property supports the construct validity of the ISE. Findings of this study should prove important to the AOBS, osteopathic general surgery program leadership, and future osteopathic general surgery applicants.

Acknowledgments

We sincerely thank Donald B. Kaveny, FACOS, Guy D. Beaumont Jr, and all other members of the ACOS that helped approve and initiate this research project.

References

1. Requirements for Certification. The American Osteopathic Board of Surgery website. Accessed December 5, 2012. http://www.aobs.org/protocol-for-certification#mn-main. Accessed December 5, 2012.
2. In-Service Examination. The American College of Osteopathic Surgeons website. http://www.facos.org/imis15/Public/Education/Residents/Inservice_Exam/Public/Navigation_Area/Education/InService_Exams_Folder/Inservice_Exams_Overview.aspx?hkey=1d694643-571c-4ec6-8991-223333a6ffbe. Accessed December 5, 2012.
3. Shen L. Progress testing for postgraduate medical education: a four-year experiment of American College of Osteopathic Surgeons Resident Examinations. *Adv Health Sci Educ Theory Pract.* 2000;5(2):117-129.
4. General Surgery Residency Curriculum. American College of Osteopathic Surgeons website. http://www.facos.org/imis15/Public/Education/Program_Directors/Curricula/General_Surgery/Public/Navigation_Area/Education/Program_Directors/Curricula/Curricula_GS.aspx?hkey=cc4f3228-46ac-43e8-ab1e-444078ce2a38. Accessed August 16, 2013.
5. Trochim W. *The Research Methods Knowledge Base.* 2nd ed. Cincinnati, OH: Atomic Dog Publishing; 2000.
6. Gerrow JD, Murphy HJ, Boyd MA, Scott DA. Concurrent validity of written and OSCE components of the Canadian dental certification examinations. *J Dent Educ.* 2003;67(8):896-901.
7. Shetler PL. Observations on the American Board of Surgery In-Training examination, board results, and conference attendance. *Am J Surg.* 1982;144(3):292-294.
8. Wade TP, Andrus CH, Kaminski DL. Evaluations of surgery resident performance correlate with success in board examinations. *Surgery.* 1993;113(6):644-648.
9. de Virgilio C, Chan T, Kaji A, Miller K. Weekly assigned reading and examinations during residency, ABSITE performance, and improved pass rates on the American Board of Surgery Examinations. *J Surg Educ.* 2008;65(6):499-503. doi:10.1016/j.jsurg.2008.05.007.
10. 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):852-856. doi:10.1001/archsurg.2010.177.