The Predictive Validity of the ABFM's In-Training Examination

Thomas R. O'Neill, PhD; Zijia Li, MS; Michael R. Peabody, PhD; Melanie Lybarger, MS; Kenneth Royal, PhD; James C. Puffer, MD

BACKGROUND AND OBJECTIVES: Our objective was to examine the predictive validity of the American Board of Family Medicine's (ABFM) In-Training Examination (ITE) with regard to predicting outcomes on the ABFM certification examination.

METHODS: This study used a repeated measures design across three levels of medical training (PGY1–PGY2, PGY2–PGY3, and PGY3–initial certification) with three different cohorts (2010–2011, 2011–2012, and 2012–2013) to examine: (1) how well the residents' ITE scores correlated with their test scores in the following year, (2) what the typical score increase was across training years, and (3) what was the sensitivity, specificity, positive predictive value, and negative predictive value of the PGY3 scores with regard to predicting future results on the MC-FP Examination.

RESULTS: ITE scores generally correlate at about .7 with the following year's ITE or with the following year's certification examination. The mean growth from PGY1 to PGY2 was 52 points, from PGY2 to PGY3 was 34 points, and from PGY3 to initial certification was 27 points. The sensitivity, specificity, positive predictive value, and negative predictive value were .91, .47, .96, and .27, respectively.

CONCLUSION: The ITE is a useful predictor of future ITE and initial certification examination performance.

(Fam Med 2015;47(5):349-56.)

The American Board of Family Medicine (ABFM) offers residents enrolled in Accreditation Council for Graduate Medical Education (ACGME)-accredited residency programs the opportunity to take the ABFM In-Training Examination (ITE). The purposes of the ABFM's ITE are (1) to provide each resident with a low-cost, low-stakes opportunity to become familiar with the general format and item writing style that will be used on the Maintenance of Certification for Family Physicians (MC-FP) Examination and (2) to provide each resident and his or her program director with an opportunity to assess how well the resident is progressing toward eventually passing the MC-FP Examination. Given these purposes, it is important that (1) the ITE content be similar to that of the MC-FP Examination, (2) the ITE results are a good approximation of how a resident would perform on the MC-FP Examination at that point in time, and (3) the ITE be predictive of an examinee's future results on the MC-FP Examination. The ABFM's ITE is designed as a low-stakes examination and, accordingly, the ABFM advises program directors that the results should not be used to make important decisions related to the promotion or advancement of the residents taking the exam.

The ABFM asserts that the ITE is a good predictor of a resident's performance on the MC-FP examination. The purpose of this study is to describe the extent to which ITE results can be used to predict future examination performance, either ITE or MC-FP, and how confident can one be in those predictions. It also examines the average score growth across each year of residency as a factor in that prediction.

Background

The ABFM's ITE was specifically designed to have a high degree of concurrent and predictive validity, both of which are forms of criterion-related validity. The important criterion in both of these cases is performance on the MC-FP Examination. The concurrent validity claim for the ABFM's ITE is that it is intended to produce scores that would be predictions of how an examinee would perform on the MC-FP Examination if he or she had taken it

From the American Board of Family Medicine, Lexington, KY (Dr O'Neill, Ms Li, Dr Peabody, Ms Lybarger, and Dr Puffer) and North Carolina State University (Dr Royal).

instead of the ITE at that point in time. To achieve this, each form of the ITE is built to the same specifications as the core questions portion of the MC-FP Examination, and ITE scores are equated onto the MC-FP scale. Concurrent validity with the MC-FP Examination is partially established through the regular quality checks that ensure the ITE is developed with the correct content specifications and that the equating was successful. Because the concurrent validity seems quite high, the ABFM has never conducted an experiment in which both tests were administered to examinees on consecutive days; however, the ABFM does administer the MC-FP Examination a few months after PGY3 residents take the ITE, which is how the ABFM usually establishes the predictive validity of the ITE.[1]

In general, the literature on how well ITEs predict success on the corresponding certification examination is positive; however, the methods used and the level of detail with which the results were reported varied noticeably across studies. A number of studies examining the predictive power of ITEs with respect to the outcome on their corresponding certification examinations have been performed by numerous American Board of Medical Specialties (ABMS) member boards, including the American Board of Neurological Surgery,[2,3] the American Board of Surgery,[4-8] the American Board of Internal Medicine,[9-12] the American Board of Psychiatry and Neurology,[13-15] the American Board of Radiology,[16,17] the American Board of Pediatrics,[18,19] the American Board of Obstetrics and Gynecology,[20,21] the American Board of Anesthesiology,[22] the American Board of Orthopedic Surgery,[23] and the American Board of Pathology.[24]

The literature related specifically to predicting the success on the ABFM's certification examination is rather sparse. In 1990, Leigh et al[1] used a repeated measure data collection design with regression on a national sample of ABFM ITE scores and ABFM certification examination scores to demonstrate that the ITE was a reasonably good predictor of performance on the certification examination. The correlations between the ITE and the certification examination ranged from .69 to .75. In 2004, Replogle and Johnson[25] used a Monte Carlo study to look at the positive predictive value (PPV) of the ABFM ITE with regard to predicting successful performance on the ABFM certification examination. They concluded that the overall ITE score had a sufficiently high PPV to use it as part of a comprehensive resident evaluation system; however, the PPV for the subtests was too low to warrant their use as performance indicators.

Methods

Participants

The ABFM's ITE is administered to nearly all family medicine residents in ACGME-accredited programs. Each year, approximately 10,000 residents from roughly 450 residency programs take the ITE. The number of residents in each year of residency is fairly evenly distributed.[26] The number of participants reported for the different comparisons in this study is slightly lower because the inclusion criteria required that each physician have a test score from consecutive test administrations. To illustrate, if a physician only had test scores for PGY1, PGY3, and the MC-FP Examination, then only the PGY3 to MC-FP comparison would be included because the PGY1 to PGY2 and the PGY2 to PGY3 comparison would not be available.

Of the possible 10,377 pairs going from PGY1 to PGY2, there were 9,630 matches (93%). Of the possible 9,921 pairs going from PGY2 to PGY3, there were 9,379 matches (95%). Of the 9,523 pairs going from PGY3 to the next administration of the MC-FP Examination, there were 6,152 matches (65%). It is important to note that some PGY3 residents do not take the next available MC-FP Examination, probably for a variety of reasons.

Instrumentation

The ABFM's MC-FP Examination measures physicians' clinical decision-making ability as it relates to family medicine. Passing this examination is one of the requirements for ABFM certification. The exam is administered in examination windows during the months of April and November of each year. The exam consists of a common core of 260 multiple choice questions plus two examinee-selected modules of 45 questions each from a menu of eight modules. These 350 items are scored as right or wrong, and the raw scores are converted to scaled scores that range from 200–800. The MC-FP Examination is scored using the dichotomous Rasch[27] model. In conjunction with a common item equating design, this model is also used to equate examinations across test forms and years of administration. The use of a common scale with a passing standard that is held constant for useful periods of time has the advantage of providing a more stable target for making predictions related to whether a particular candidate will pass or fail. During the timeframe from which the data was gathered, the minimum passing score was 390. The process used to develop the content specifications for this examination is described in greater detail by Norris et al.[28]

The ABFM's ITE contains 240 multiple-choice items and is built to the same specifications as the core element (non-module portion) of the MC-FP Examination. Each year, there is a different form of the ITE with no items in common from the previous form. In order to equate the ITE across administrations and to make the ITE score represent the examinee's predicted performance on the MC-FP Examination, the ABFM includes a small number of ITE questions as unscored pretest questions on the MC-FP Examination, which are calibrated onto the MC-FP scale. These questions and

their associated calibrations on the MC-FP scale are used to connect each administration of the ITE to the continuously maintained MC-FP scale. Because the ITE has been equated onto the MC-FP scale and built to similar specifications, ITE scores should be highly correlated with the MC-FP scores examinees would have earned had they taken it instead of the ITE.

Procedures

This study used a repeated measures design across three levels of medical training (PGY1 to PGY2, PGY2 to PGY3, and PGY3 to initial certification) with three different cohorts (2010–2011, 2011–2012, and 2012–2013) to examine: (1) how well the residents' ITE scores on PGY1, PGY2, and PGY3 are correlated with their test scores in the following year (PGY 2, PGY3, and MC-FP, respectively), (2) what the typical score increase was from PGY1 to PGY2, PGY2 to PGY3, and PGY3 to initial certification, and (3) what was the sensitivity, specificity,[12,29] positive predictive value (PPV) and negative predictive value (NPV) of the ITE with regard to predicting results on the MCFP Examination. This study was deemed exempt by the American Academy of Family Physicians Institutional Review Board.

Results

Across years of training, the ITE correlated at .69 for PGY1 to PGY2, .70 for PGY2 to PGY3, and .71 for PGY3 to MC-FP (Table 1, Figure 1). These correlations were all positive and statistically significant. The correlations were very similar across cohorts and years of medical training. After disattenuating for the unreliability of the examination, the correlations ranged from .81 to .85.

With regard to resident performance over time, the results indicate that exam scores tend to increase with each successive year of residency; however, the average increase was smaller in each successive year. The average increase from PGY1 to PGY2 was the largest at 52 points, followed by PGY2 to PGY3 with 34 points, and finally 27 points from PGY3 to MC-FP (Table 1, Figure 2).

Using a minimum passing score (MPS) of 390 for both the ITE (PGY3s only) and the MC-FP Examination, the sensitivity, specificity, PPV, and NPV were computed (Table 2). The sensitivity, the proportion of actual MC-FP passers who were also predicted to pass was .91. The specificity, the proportion of actual MC-FP failers who were also predicted to fail was .47. The PPV, the proportion of people who were predicted to pass the MC-FP Examination based on their ITE score and actually passed was .96. The NPV, the proportion of people who were predicted to fail the MC-FP Examination based on their ITE score and actually failed was .27. Additionally, Figure 3 shows the trade-off between

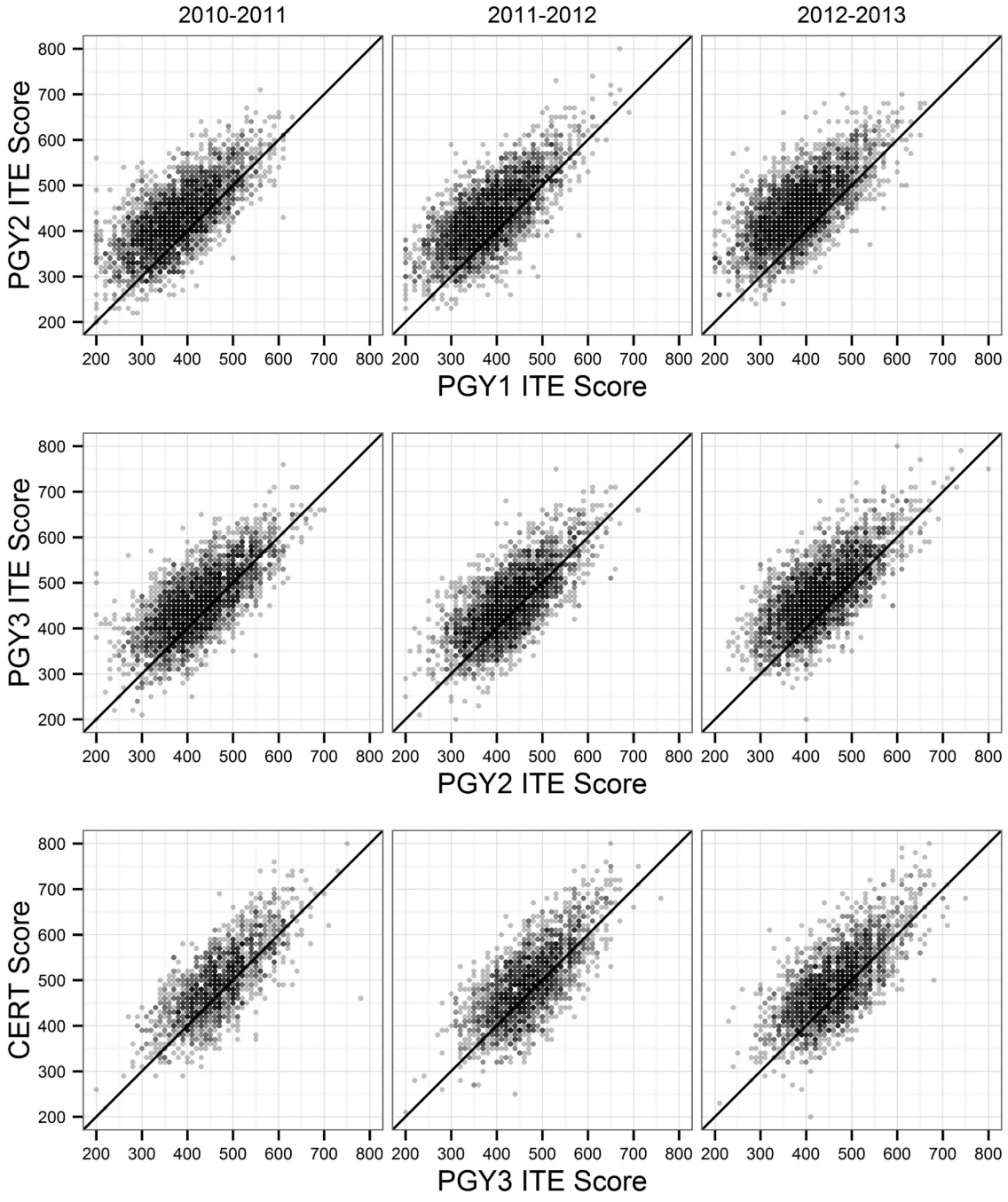
Table 1: Summary Statistics of Comparisons

	Gains by Year						Correlations	
	n	Mean	SD	Min	Max	SE	Pearson r	Disattenuated Correlations
PGY1 to PGY2								
2010–2011	3,102	48	58	-180	360	1.0	.69**	.82
2011–2012	3,242	47	56	-190	290	1.0	.72**	.88
2012–2013	3,286	61	59	-170	330	1.0	.67**	.82
Overall	9,630	52	58	-190	360	0.6	.69**	n/a
PGY2 to PGY3								
2010–2011	3,067	31	57	-210	320	1.0	.71**	.84
2011–2012	3,078	26	56	-150	260	1.0	.70**	.85
2012–2013	3,234	43	57	-200	250	1.0	.70**	.85
Overall	9,379	34	57	-210	320	0.6	.70**	n/a
PGY3 to MC-FP								
2010–2011	1,617	27	55	-320	220	1.4	.75**	.84
2011–2012	2,104	23	59	-190	220	1.3	.71**	.81
2012–2013	2,431	29	58	-210	290	1.2	.70**	.81
Overall	6,152	27	58	-320	290	0.7	.71**	n/a

* $P<.05$, ** $P<.01$

Note: Disattenuated correlations could not be calculated for the overall results because the disattenuation process removes the degree of unreliability from the pair of test forms. The degree of unreliability of each test could not be easily combined.

Figure 1: Scatterplot of ITE Performance and Subsequent Exam Performance by Year of Training and Cohort



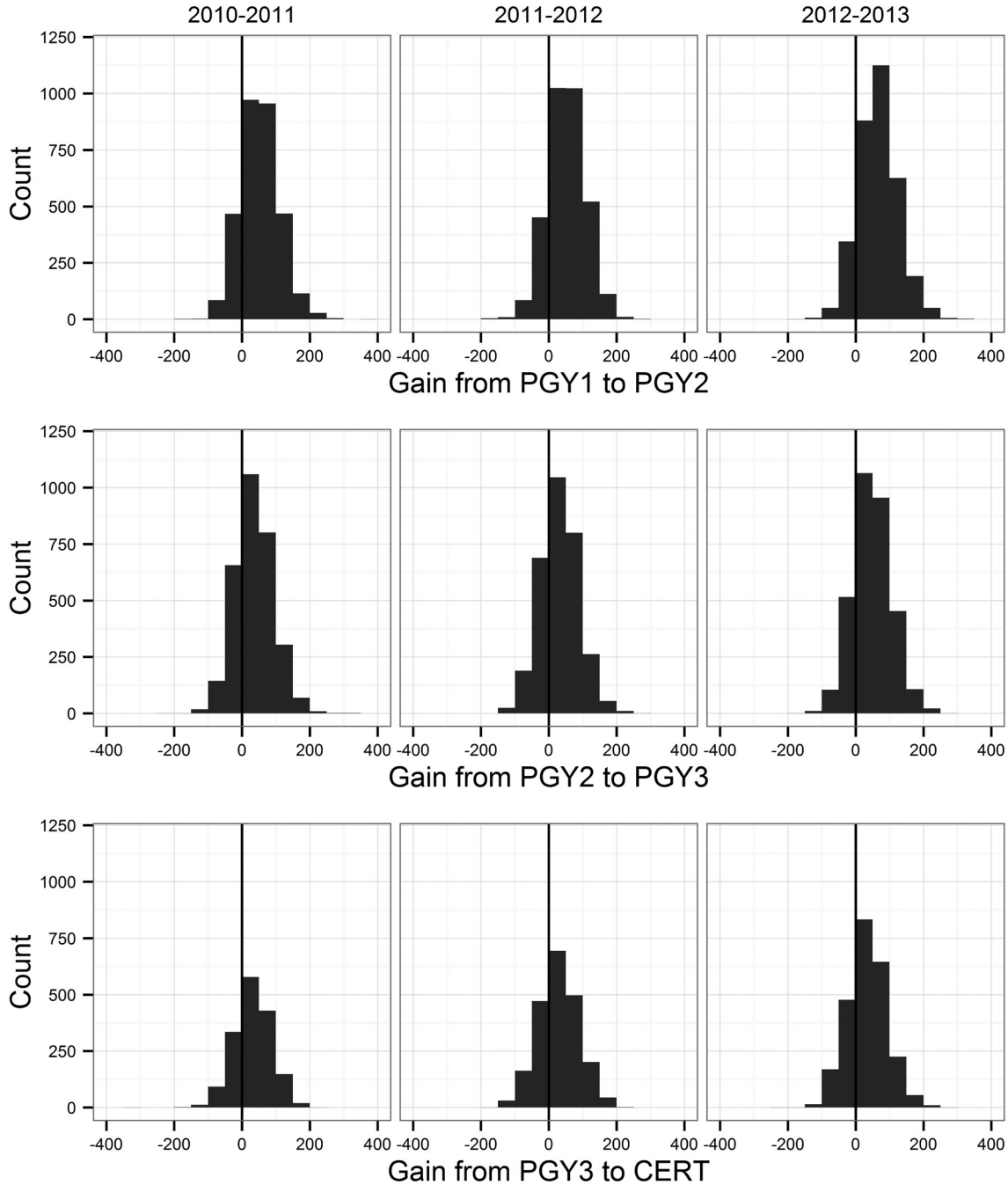
sensitivity and specificity using different ITE prediction thresholds.

Discussion

Correlation of Exam Scores

The correlation of ITE scores with ITE scores 1 year later or with MC-FP scores 6 months later is typically about 0.7 (Table 1). This indicates that ITE scores can be used as reasonably good predictors of future performance on the ITE and MC-FP Examinations. This correlation is the appropriate correlation for making predictions because it includes both differences in the dimensionality across test forms and the degree of unreliability associated with each test form. The Rasch reliability of the ITE typically runs approximately .81 to .83. The Rasch reliability of the MC-FP Examination is typically .92 or .93. To assess the extent to which two test forms are measuring the same dimension, the correlation must be disattenuated for the degree of unreliability

Figure 2: Histogram of Score Increase Across Exam Administrations by Year of Training and Cohort



associated with both test forms.[30-32] The disattenuated correlation across test forms ranged from .81 to .88 (Table 1), which demonstrates that the construct across these pairs of test forms is quite similar but not perfectly identical. It should be noted that the MC-FP Examination has two, examinee-selected, 45-item modules that the ITE does not. As expected, the disattenuated correlations for PGY1 to PGY2 and for PGY2 to PGY3 are slightly better than the correlation from PGY3 to MC-FP. This suggests that the modules, which account for 26% of the MC-FP Examination, tap a slightly different dimension. The unadjusted correlations look to be slightly better for the PGY3 to MC-FP than the PGY1 to PGY2 and PGY2 to PGY3 comparisons, but this could be attributed to the higher degree of reliability found with the MC-FP Examination.

In order for the ITE to have a high degree of predictive validity, it

Table 2: Ability of ITE to Predict MC-FP Examination Results

		MC-FP Examination		
		Pass	Fail	
In-Training Exam	Pass	5,188 TP	226 FP	Positive Predictive Value 5,188/5,414= .96 TP/(TP+FP)
	Fail	538 FN	200 TN	Negative Predictive Value 200/738= .27 TN/(TN+FN)
		Sensitivity 5,188/5,726= .91 TP/(TP+FN)	Specificity 200/426= .47 TN/(TN+FP)	

Pass-fail predictions on the ITE and outcomes on the MC-FP Examination were both based upon a score of 390. n=6,152.

TP—True positive, FP—False positive, FN—False negative, TN—True negative

is necessary, but not sufficient for the ITE scores to be highly correlated with the certification examination scores. In addition to being correlated, the scores must also be on a common scale across administrations so that the score can be used to make a prediction about future performance. If each administration of the ITE and MC-FP Examination were scored on unique and unconnected scales, then the formula to convert an ITE score into a prediction about a subsequent test performance would be known only after the subsequent performance. Such a situation could describe past performance but would not qualify as making a prediction.

Longitudinal Performance

On average, scores increased with each successive year of residency, although the magnitude of the increase diminished with each additional year (Figure 2). Within the change in score performance distribution for each year, a noticeable variability in the amount of change for individuals was apparent. Sometimes the change was negative. It seems unlikely that an additional year of training caused poorer examination performance, so the causes for this were more likely to be resident specific. Due to the noticeable variation in performance over time (SD=58), expectations related to an individual's longitudinal performance should be made cautiously; however, it should be noted that the mean and standard deviation of the score change was fairly stable across cohorts (Table 1).

Sensitivity, Specificity, and Predictive Power of the ITE

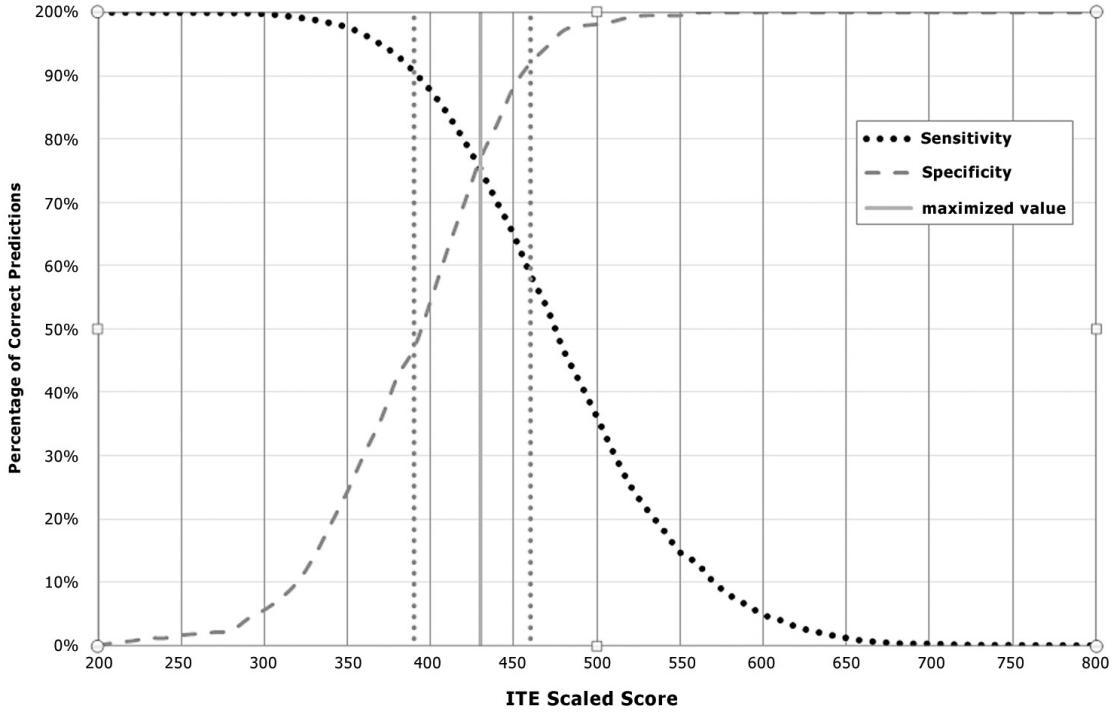
Because the ABFM knows the outcomes from both tests (ITE and MC-FP) when these studies are conducted, the accuracy of the ITE predictions are assessed separately by the MC-FP Examination pass-fail status, or in other words, the sensitivity and specificity of the ITE. For people who pass, the ITE was usually (91%) correct when assuming that an ITE of 390 predicts passing. For people who fail, the ITE was far less predictive (47%). A review of Figure 3 shows that there is a tradeoff between false positives and false negatives depending on where the ITE prediction threshold is set. Lowering the prediction threshold below 390 will not increase the proportion of true positives by very much, but it will significantly drop the number of true negatives. If the ITE prediction threshold was set at 460, it would significantly increase the number of true negatives, but it would also significantly decrease the number of true positives. The optimal compromise point between sensitivity and specificity seems to be 430. A table, similar to Figure 3, examining positive and negative predictive values was not included because the row totals (Table 2) change depending on where the prediction threshold is set. The resulting chart would not have a monotonically increasing (or decreasing) function, and the interpretation would be complicated because the number of observations would change at each condition level.

Limitations

This study was limited to describing the extent to which the ITE predicts future ITE performance or MC-FP Examination performance. One could infer a construct of growth to describe the mean score increases, but this construct would only be descriptive in nature, not causative. It was based upon a national sample that was collected over several years, so it should be fairly generalizeable to family medicine residents but not necessarily for residents in other specialties. The low indices for specificity and NPV indicate that the ITE is not good at predicting failers; however, this suggests that some low scoring third-year residents have been able to successfully remediate prior to taking the certification examination.

It should be noted that this study was based only on residents' performance on newly released ITEs. All ABFM diplomates have access to the last 3 years of ITEs for free, and it is not uncommon for diplomates to download and take an ITE to assess how well prepared they are for their next certification examination. The quality of the predictions for diplomates should be comparable to those for residents if the diplomate is encountering the ITE as a "newly released" test; however, there are several reasons why a diplomate might not encounter the ITE as a newly released test, which

Figure 3: Sensitivity and Specificity of the ITE Across Dfferent Prediction Threshold



would inflate the prediction. For example, diplomates sometimes use the old ITEs as a study guide or use the ABFM smartphone application, which contains ITE questions. The ABFM also licenses old ITE questions to the American Academy of Family Physicians. In these cases, if a diplomate were to take an ITE with questions that he or she had seen before, then answering those questions correctly does not represent the diplomate's mastery over the entire body of family medicine but only their ability to recall questions that he or she had seen before. For this reason, it is recommended that looking at ABFM style questions be used by diplomates only for assessing how much preparation they need before retesting.

Conclusions

Neither program directors[33,34] nor the residents themselves[35] are good at predicting residents' exam scores. Although the ITE seems to be a useful tool for predicting successful performance on the MC-FP Examination, there are multiple factors that may influence the accuracy of the prediction. A resident may have been sick, failed to take the test seriously, been called away to perform a clinical task, had a serious family issue, etc. The ITE should be used to help residents identify issues related to their timely progress toward becoming certified by the ABFM. It seems that this is indeed happening. The low NPV (.27) shows that a substantial number of people who were predicted to fail were able to successfully remediate before taking the MC-FP Examination. Studies that examine the predictive validity of any ITE with regard to subsequent performance on a certification examination presuppose that this is at least one of the purposes of the ITE. It would stand to reason that predictive power would be a highly desirable quality as program directors would be able to predict whether a resident has a high likelihood of passing; the pass rates can have an impact on a program's ACGME accreditation status. Residents would also care about this predictive quality because failing the certification examination often means having to retake an expensive examination, having to continue to prepare for it, and potentially missing employment opportunities available on completion of training. When these issues are considered, a strong case could be made that the predictive power of an ITE is its most important quality.

CORRESPONDING AUTHOR: Address correspondence to Dr O'Neill, American Board of Family Medicine, 1648 McGrathiana Parkway, Suite 550, Lexington, KY 40511-1247. 859-269-5626, ext. 1225. toneill@theabfm.org.

References

1. Leigh TM, Johnson TP, Pisacano NJ. Predictive validity of the American Board of Family Practice In-Training Examination. Acad Med 1990;65(7):454-7.
2. Hubbard J, Levit E. The National Board of Medical Examiners: the first seventy years. Philadelphia, PA: National Board of Medical Examiners, 1985.
3. Hubbard JP, Furlow LT, Matson DD. An in-training examination for residents as a guide to learning. N Engl J Med 1967;276(8):448-51.
4. Garvin PJ, Kaminski DL. Significance of the in-training examination in a surgical residency program. Surgery 1984;96(1):109-13.

5. Biester TW. A study of the relationship between a medical certification examination and an in-training examination. Chicago, IL: American Educational Research Association, 1985.
6. Biester TW. The American Board of Surgery in-training examination as a predictor of success on the qualifying examination. Current Surgery 1987;44(3):194-8.
7. 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. Journal of Surgical Education 2015;in press.
8. Shelter PL. Observations on the American Board of Surgery in-training examination, Board results, and conference attendance. Am J Surgery 1982;144(3):292-4.
9. Grossman RS, Fincher R-ME, Layne RC, Seelig CB, Berkowitz LR, Levine MA. Validity of the in-training examination for predicting American Board of Internal Medicine Certifying Examination Scores. J Gen Intern Med 1992;7:63-7.
10. Waxman H, Braunstein G, David D, et al. Performance on the internal medicine second-year residency in-training examination predicts the outcome of the ABIM certifying examination. J Gen Intern Med 1994;9(12):692-4.
11. Rollins LK, Martindale JR, Edmond M, Manser T, Scheld WM. Predicting pass rates on the American Board of Internal Medicine Certifying Examination. J Gen Intern Med 1998;13(6):414-6.
12. Babbott SF, Beasley BW, Hinchey KT, Blotzer JW, Holmboe ES. The predictive validity of the internal medicine in-training examination. Am J Med 2007 Aug;120(8):735-40.
13. Webb LC, Juul D, Reynolds CF, et al. How well does the Psychiatry Residency in-training examination predict performance on the American Board of Psychiatry and Neurology Part I Examination? Am J Psychiatry 1996;153(6):831-2.
14. Goodman JC, Juul D, Westmoreland B, Burns R. RITE performance predicts outcome on the ABPN Part I examination. Neurology 2002;58(8):1144-6.
15. Juul D, Schneidman BS, Sexson SB, et al. Relationship between Resident-In-Training Examination in Psychiatry and subsequent examination performances. Acad Psychiatry 2009;33(5):404-6.
16. Baumgartner BR, Brothers Peterman S. Relationship between American College of Radiology in-training examination scores and American Board of Radiology Written Examination Scores. Acad Radiol 1996;3(10):873-8.
17. Baumgartner BR, Brothers Peterman S. Relationship between American College of Radiology in-training examination scores and American Board of Radiology Written Examination Scores. Acad Radiol 1998;5(5):374-9.
18. McCaskill QE, Kirk JJ, Barata DM, Wludyka PS, Zenni EA, Chiu TT. USMLE step 1 scores as a significant predictor of future board passage in pediatrics. Amb Pediatr 2007;7(2):192-5.
19. Althouse LA, McGuinness GA. The in-training examination: an analysis of its predictive value on performance on the General Pediatrics Certification Examination. J Pediatr 2008;153(3):425-8.
20. Spellacy WN, Carlan SJ, McCarthy JM, Tsibris JC. Prediction of ABOG written examination performance from the third-year CREOG in-training examination results. J Reprod Med 2006;51(8):621-2.
21. Withiam-Leitch M, Olawaiye A. Resident performance on the in-training and board examinations in obstetrics and gynecology: implications for the ACGME Outcome Project. Teach Learn Med 2008 Apr-Jun;20(2):136-42.
22. Kearney RA, Sullivan P, Skakun E. Performance on ABAASA in-training examination predicts success for RCPSC certification. Can J Anesthesia 2000;47(9):914-8.
23. Klein GR, Austin MS, Randolph S, Sharkey PF, Hilibrand AS. Passing the Boards: can USMLE and orthopaedic in-training examination scores predict passage of the ABOS Part-I Examination? J Bone Joint Surg 2004;86(5):1092-5.
24. Rinder HM, Grimes MM, Wagner J, et al, RISE Committee, American Society for Clinical Pathology and the American Board of Pathology certifying examinations. Senior pathology resident in-service examination scores correlate with outcomes of the American Board of Pathology certifying examinations. Am J Clin Pathol 2011 Oct;136(4):499-506.
25. Replogle WH, Johnson WD. Assessing the predictive value of the American Board of Family Practice in-training examination. Fam Med 2004;36(3):185-8.
26. O'Neill TR, Peabody MR. ITE score results handbook. Lexington, KY: American Board of Family Medicine, 2013.
27. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research, 1960.
28. Norris TE, Rovinelli RJ, Puffer JC, Rinaldo J, Price DW. From specialty-based to practice-based: a new blueprint for the American Board of Family Medicine Cognitive Examination. J Am Board Fam Pract 2005;18(6):546-54.
29. Weinstein MC, Fineberg HV. Clinical decision analysis. Philadelphia, PA: W.B. Saunders Company, 1980.
30. Schumaker RE, Muchinsky PM. Disattenuating correlation coefficients. Rasch Measurement Transactions 1996;10(1):479.
31. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15(1):72-101.
32. Zimmerman DW, Williams RH. Properties of the Spearman Correction for Attenuation for Normal and Realistic Non-Normal Distributions. Applied Psychological Measurement 1997;21(3):253-70.
33. Hawkins RE, Sumption KF, Gaglione MM, Holmboe ES. The in-training examination in internal medicine: resident perceptions and lack of correlation between resident scores and faculty predictions of resident performance. Am J Med 1999;106(2):206-10.
34. Taylor C, Lipsky MS. A study of the ability of physician faculty members to predict resident performance. Fam Med 1990 Jul-Aug;22(4):296-8.
35. Parker RW, Alford C, Passmore C. Can family medicine residents predict their performance on the in-training examination? Fam Med 2004;36(10):705-9.