APM Association of Professors of Medicine

APM Perspectives

The Association of Professors of Medicine (APM) is the national organization of departments of internal medicine at the US medical schools and numerous affiliated teaching hospitals as represented by chairs and appointed leaders. As the official sponsor of The American Journal of Medicine, the association invites authors to publish commentaries on issues concerning academic internal medicine.

For the latest information about departments of internal medicine, please visit APM's website at www.im.org/APM.

The Predictive Validity of the Internal Medicine In-Training Examination

Stewart F. Babbott, MD,[a] B. W. Beasley, MD,[b] K. T. Hinchey, MD,[c] J. W. Blotzer, MD,[d] E. S. Holmboe, MD[e]*

[a]University of Kansas Medical Center, Kansas City, Kan; [b]University of Missouri-Kansas City, Kansas City, Mo; [c]Baystate Medical Center, Springfield, Mass; [d]York Hospital, York, Pa; [e]Yale University, New Haven, Conn.

KEYWORDS: Board certification; Certifying examination; In-training examination; Medical education; Medical knowledge; Prediction; Predictive validity; Residents

The Internal Medicine In-Training Examination (IM-ITE) was developed by the American College of Physicians, the Association of Professors of Medicine, and the Association of Program Directors in Internal Medicine as a formative assessment of medical knowledge during residency training.[1] Designed as a low-stakes examination, the IM-ITE specifically targets a postgraduate year PGY-2 resident's level of knowledge. Most programs offer the IM-ITE yearly to all residents; on average, more than 92% of residents in internal medicine programs complete this examination each year (Sean McKinney, personal communication, 2006). However, an individual resident may or may not take the examination in all 3 years. Residents receive a score report that provides the total percent of questions answered correctly, their percentile rank compared with their peer group at the national level, and feedback on performance in specific areas. Residents use the results of this examination as a guide to their progress and to define areas for more intensive study. Program directors use the results of this examination to assess and counsel residents, create remediation plans for residents who perform poorly on the examination, and address areas of strength and weakness in the residency curriculum.

The American Board of Internal Medicine Certifying Examination (ABIM-CE) is a high-stakes examination taken after graduation. The American Board of Internal Medicine (ABIM) has no role in the IM-ITE, nor does the IM-ITE Committee have a role in the ABIM-CE (the IM-ITE Committee is made up of members from American College of Physicians, Association of Professors of Medicine, and Association of Program Directors in Internal Medicine). However, prior studies have found strong correlations between IM-ITE scores of PGY-2 residents and outcomes of the ABIM-CE. One study found that if a PGY-2 resident's score was above the 35th percentile, then the positive predictive value (PPV) of passing the ABIM-CE was 89%.[2] Another study found almost identical results in a different group of PGY-2 residents; scores above the 35th percentile had a PPV of 88%, and scores above the 20th percentile had a PPV of 81%.[3] An analysis of all residents' IM-ITE results over a 12-year period found that test performance improved with each year of training, demonstrating the construct validity of IM-ITE.[4] In addition, the amount of time spent on internal medicine rotations before the examination and the time permitted to complete the IM-ITE also were predictors of performance on the ABIM-CE.[4] Another analyisis developed

*Dr. Holmboe is now with the American Board of Internal Medicine.

Presented at the Society of General Internal Medicine New England Region and National Meetings, 2004.

Funding Source: none.

Requests for reprints should be addressed to Stewart Babbott, MD, Division of General and Geriatric Medicine, University of Kansas Medical Center, 3901 Rainbow Boulevard, 5026 Wescoe, MS 1020, Kansas City, KS 66160.

E-mail address: sbabbott@kumc.edu.

0002-9343/$ -see front matter © 2007 The Association of Professors of Medicine. All rights reserved.

doi:10.1016/j.amjmed.2007.05.003

and validated a model to predict ABIM-CE performance based on IM-ITE scores. On the basis of this model, correct classification of performance was between 75% and 79%. This analysis suggested that a score greater than 65% on the IM-ITE would predict certain pass on the ABIM-CE and less than 49% would predict certain fail.[5]

Several events have since occurred that may now affect the predictive performance of the IM-ITE for the ABIM-CE. The first-time pass rate for the certification examination in internal medicine averaged between 68% and 69% in the early 1990s. This number increased to 82% in 1996 and is now more than 90%.[6] Several factors may have influenced these changes. First, the introduction of Maintenance of Certification (MOC) examinations in 1996 led ABIM to formally reevaluate standards for all of its examinations for the first time since 1988 (L. Grosso, personal communication, 2006). During the evaluation, ABIM established common standards for certification and MOC examinations.

PERSPECTIVES VIEWPOINTS

- The study re-examines the predictive validity of the IM-ITE for the ABIM-CE.
- The results show the percentile rank cut-points to be significantly lower than in previous studies.
- The study extends predictors to cover all years of residency training.

Second, the importance of certification seems to be growing. Although board certification has always been a voluntary process, market forces have led some hospitals and health plans to require board certification for their physicians.[7,8] Third, in 1999 the timing of the IM-ITE administration changed from the spring to the fall of each academic year.

Finally, ABIM-CE pass rates for first-time takers is considered a measure of program effectiveness, and these results are now available to the public. In July 2003, the Residency Review Committee for Internal Medicine required an overall ABIM-CE pass rate of 70% as 1 of the program requirements for accreditation. As the importance of passing ABIM-CE increases, perhaps both residents and program leaders have devoted more time, energy, and focus to preparing for both the IM-ITE and the ABIM-CE.

In the context of these events, program directors and residents need guidance on how to accurately identify and advise those individuals at greatest risk for failing the ABIM-CE. To determine more current predictive values of the IM-ITE, a study of residents participating in the IM-ITE at 4 internal medicine residency training programs was conducted. The primary objective of this study was to determine the optimal cut scores based on sensitivity and specificity for PPV and negative predictive value (NPV).

METHODS

Setting and Subjects

Four internal medicine programs with IM-ITE pass rates of less than 100% were identified and invited to participate in this study (Baystate Medical Center, University of Missouri, Kansas City, York Hospital, and Yale Waterbury Primary Care Program). Programs with pass rates of 100% were excluded because the outcomes of interest included failure on the ABIM-CE. Programs with pass rates less than 100% are actually the norm; approximately one quarter of all programs achieve a 100% pass rate each year (L. Grosso, personal communication, 2006). Residents surveyed included those from the graduating classes of 2000, 2001, and 2002 who took an IM-ITE in any year of their residency training and sat for the ABIM-CE. Each program obtained institutional review board approval and recorded the following demographic characteristics for each resident: gender; US medical school graduate or international medical school graduate status; all scores of the IM-ITE percentile and percent correct for each IM-ITE taken; and the ABIM-CE results, year taken, and year eligible. Authors at the respective institutions entered the data; all data were reviewed by the first author.

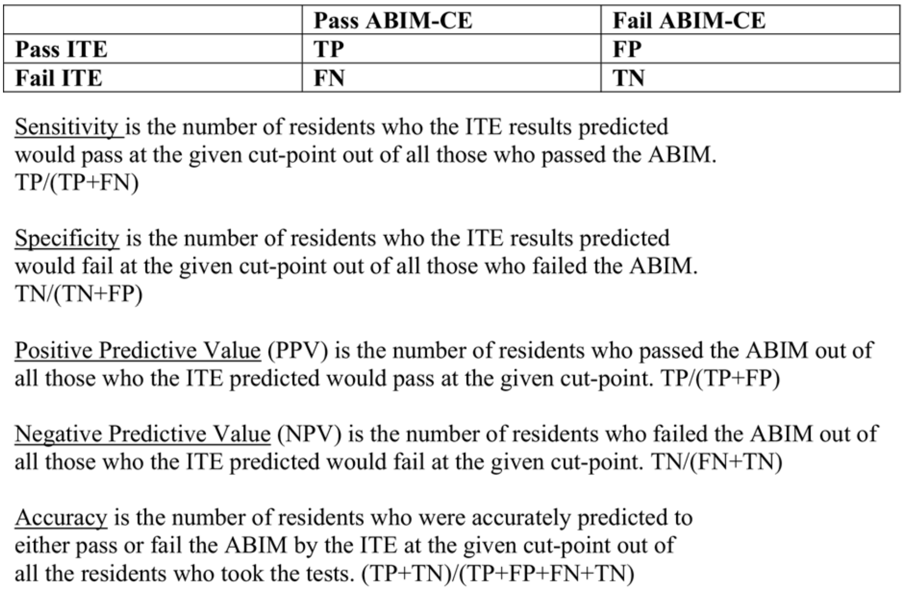
Analysis

Analysis was performed in 4 steps. First, the authors examined frequency distributions and descriptive statistics for evidence of skewing, outliers, and non-normality for continuous variables. To examine differences between the 4 programs, *t* tests and chi-square tests were used. For bivariate analyses, *t* tests, analysis of variance, and Pearson's *r* were used as appropriate to the level of measurement.

Second, the authors used logistic regression models for each IM-ITE taken to calculate the likelihood of passing the ABIM-CE. The fit of the models was assessed using Hosmer-Lemeshow Goodness of Fit and C-statistics; SPSSTM 12.0.1 for Windows TM (SPSS Inc, Chicago, Ill) was used for statistical analysis.

Third, the authors used receiver operating characteristic (ROC) curves to determine the cut-points on IM-ITE scores that optimize the sensitivity, specificity, PPV, and NPV for passing the ABIM-CE.

Finally, the authors performed the multivariable analysis in 2 steps because of the observation that not all residents take the IM-ITE each year. This broader analysis meets the goal of extending prior analyses from just the PGY-2 year to all years of training. The



	Pass ABIM-CE	Fail ABIM-CE
Pass ITE	TP	FP
Fail ITE	FN	TN

Sensitivity is the number of residents who the ITE results predicted would pass at the given cut-point out of all those who passed the ABIM. TP/(TP+FN)

Specificity is the number of residents who the ITE results predicted would fail at the given cut-point out of all those who failed the ABIM. TN/(TN+FP)

Positive Predictive Value (PPV) is the number of residents who passed the ABIM out of all those who the ITE predicted would pass at the given cut-point. TP/(TP+FP)

Negative Predictive Value (NPV) is the number of residents who failed the ABIM out of all those who the ITE predicted would fail at the given cut-point. TN/(FN+TN)

Accuracy is the number of residents who were accurately predicted to either pass or fail the ABIM by the ITE at the given cut-point out of all the residents who took the tests. (TP+TN)/(TP+FP+FN+TN)

Figure Definitions of test characteristics. ABIM-CE = American Board of Internal Medicine Certifying Examination; ITE = In-Training Examination.

authors first analyzed only those individuals who completed the IM-ITE each year of their residency and attempted the ABIM-CE. They next examined the scores of residents who completed at least 1 IM-ITE in any year of residency and attempted the ABIM-CE.

The Figure defines the sensitivity, specificity, PPV, NPV, and accuracy for this study.

RESULTS

The characteristics of participating programs are shown in Table 1. The programs did not differ by the proportion of US medical school graduate/international medical school graduate residents or by ABIM-CE pass rates. One residency program had a greater proportion of graduates who were female ($P < .05$). Between 2000 and 2002, a total of 170 residents graduated from the 4 programs. The database was complete and accurate for all testing data on all residents. Any missing score for a specific IM-ITE or ABIM-CE meant that the resident did not take the test. Ninety-seven residents (61%) completed the IM-ITE each year of their residency and the ABIM-CE, and 158 residents (93%) completed at least 1 IM-ITE during residency and the ABIM-CE.

Of the 97 residents who completed the IM-ITE each year, 72 (74%) passed the ABIM-CE on the first attempt. Fifty-four (56%) were men and 43 (44%) were women. Of the residents who completed the test at least once during their residency (n = 158, Table 1), the overall pass rate was 72% (similar to the pass rate of the cohort who took the examination each year). Women comprised 53% of this group (vs 44% of the cohort who took the examination each year). ABIM-CE pass rates

Table 1 Demographics of Participating Residency Programs

	BMC*	UMKC	YORK	YALE	Total
No. of categoric positions per year	18	16	6	20	
No. of graduating residents during study	51	44	14	61	170
Completed at least 1 ITE and ABIM-CE, No. (%)	47 (92)	39 (89)	13 (92)	59 (97)	158 (93)
Completed ITE every year and ABIM-CE, No. (%)	36 (77)	33 (85)	8 (62)	20 (34)	97 (61)
Female, No. (%)	24 (48)	19 (43)	12 (86)†	34 (56)	89 (53)
International medical graduates, No. (%)	2 (4)	11 (25)	3 (21)	6 (10)	22 (13)
Passed ABIM on first attempt, No. (%)	32 (68)	31 (80)	10 (77)	40 (68)	113 (72)
Passed ABIM on any attempt, No. (%)	36 (77)	31 (80)	10 (77)	48 (81)	125 (79)

BMC = Baystate Medical Center, Tufts University School of Medicine; UMKC = University of Missouri, Kansas City, School of Medicine; YORK = York Hospital, University of Maryland School of Medicine/Pennsylvania State University School of Medicine; YALE = Yale Waterbury Primary Care Program, Yale University School of Medicine; ITE = In-Training Examination; ABIM-CE = American Board of Internal Medicine Certifying Examination.

No. (%) = number of residents (percent of item compared with total for the site in each of the site specific columns; for the total column, the denominator is 170 and subtotals are a percent of that number).

†<0.05.

Table 2 Test Characteristics for Internal Medicine In-Training Examination Predicting American Board of Internal Medicine Certifying Examination Scores for Residents Taking Internal Medicine In-Training Examination Every Year (N = 97)

Absolute IM-ITE Scores (Percent Questions Answered Correctly)						
PGY	Cutoff IM-ITE (Percent Items Correct)	Sensitivity	Specificity	PPV	NPV	Accuracy
1	48	79%	80%	92%	57%	79%
2	54	81%	80%	92%	59%	80%
3	58	86%	92%	98%	70%	87%
Relative IM-ITE Scores (Percentile Rank Relative to PGY Peer Group)						
PGY	Cutoff IM-ITE (Percentile Rank)	Sensitivity	Specificity	PPV	NPV	Accuracy
1	23	82%	84%	94%	62%	83%
2	23	86%	88%	95%	69%	87%
3	21	89%	92%	97%	74%	90%

IM-ITE = Internal Medicine In-Training Examination; PGY = postgraduate year; PPV = positive predictive value; NPV = negative predictive value.

P value for each area under the curve is less than .0001.

were not significantly different in regard to the 4 study sites, resident gender ($P = .47$), or international medical school graduate status ($P = .55$). The Hosmer-Lemeshow Goodness of Fit and C-statistics for all models were not significant ($P > .9$), indicating that the observed and predicted numbers of graduates passing the ABIM-CE did not differ significantly.

Table 2 shows the cutoff scores in the ROC analysis at which sensitivity and specificity were maximized for residents who took the IM-ITE each year. The results show a greater than 90% PPV for passing the ABIM-CE for each year the resident took the IM-ITE. As expected, the absolute cutoff for percent questions answered correctly on the IM-ITE increased modestly with each PGY year, consistent with prior analysis.[4] However, when using the percentile rank to compare the PGY peer group, the percentile rank cutoff ranged between the 21st and 23rd percentile.

The results by PGY level for any resident who took at least 1 IM-ITE during their residency are shown in Table 3. The results are nearly identical to the results of the 97 residents who took the IM-ITE every year.

For a comparison with the previous work by Rollins and colleagues,[5] the authors studied the cut-point at which the PPV of the IM-ITE became 100% or when passing the ABIM-CE was a certainty. In the dataset, PGY-2 residents who scored at or greater than 61% of the items correct all passed the ABIM-CE. At this cut-point, the sensitivity was 41% and the specificity was 100%. In contrast, however, the PGY-2 resident

Table 3 Test Characteristics for the Internal Medicine In-Training Examination Predicting American Board of Internal Medicine Certifying Examination Scores for Residents Taking at Least One Internal Medicine In-Training Examination*

Absolute IM-ITE Scores (Percent Questions Answered Correctly)							
PGY	No. of Residents Each Year	Cutoff ITE (Percent Items Correct)	Sensitivity	Specificity	PPV	NPV	Accuracy
1	128	48	79%	79%	90%	61%	79%
2	136	54	78%	76%	90%	56%	77%
3	134	57	81%	82%	93%	59%	81%
Relative IM-ITE Scores (Percentile Rank Relative to PGY Peer Group)							
PGY	No. of Residents Each Year	Cutoff ITE (Percent Items Correct)	Sensitivity	Specificity	PPV	NPV	Accuracy
1	128	24	81%	82%	91%	64%	81%
2	136	23	83%	84%	93%	65%	83%
3	134	20	84%	82%	93%	63%	84%

PGY = postgraduate year; ITE = In-Training Examination; PPV = positive predictive value; NPV = negative predictive value.

P value for each area under the curve is less than .0001.

*Subjects in this analysis had taken the ITE in any year of their training. This includes the 97 subjects who took all 3 years, and their classmates who took the examination in 1 or 2 of their 3 years of residency.

scoring the lowest on the IM-ITE in this study's dataset also passed the ABIM-CE.

DISCUSSION

These results show that the percentile rank cut-points are significantly lower than in previous studies that reported the 35th percentile at the PGY-2 level as an important predictor for passing the ABIM-CE.[2,3] For years, program directors have used this figure as the "gold standard." The results of this study also confirm the construct validity of the IM-ITE with residents, on average, improving their absolute scores over 3 years of training. Last, these results continue to show the predictive validity of the IM-ITE relative to the ABIM-CE.

This study, however, extended the prior analyses by looking at any and all years of IM-ITE completion (in contrast with the prior studies that addressed PGY-2 IM-ITE scores only) and by addressing the progression across years in program-specific data rather than in aggregate national data. These broadened analyses could help meet a practical need for program directors who must counsel residents on the basis of any available IM-ITE data.

Although it is not clear why the cut-point has changed, several explanations are possible. First, it is possible that the ABIM-CE, the IM-ITE, or both may have changed. As noted earlier, ABIM evaluated standards for all of its examinations with the introduction of the MOC examination in 1996. In this evaluation, ABIM established common standards for certification and MOC examinations. ABIM rigorously reviews the examination each year, and over time the mean-equated scores for each examination have remained the same, although first-time performance on the examination has been improving (L. Grosso, personal communication, 2006). Factors such as examinee ability, motivation, and preparation for the examination may all influence pass rates.

The reliability of the IM-ITE is high and hovers consistently at approximately KR20 = 0.93 across administrations. This level of internal consistency implies that candidates taking the test have a precise measurement of their proficiency level. The process by which the examination is developed and the close quality control procedures applied at every step of its development and scoring attest to the content validity of the examination. Scores on the IM-ITE have been equated (Rasch model) and placed on a research scale since 1988. These scores have been found to demonstrate consistent statistical behavior across the years (R. Subhioyah, PhD, personal communication, 2007).

Given that both examinations have a high quality of review, it is less likely that either test has changed substantially over time, although specific content would be expected to change given advances in medical knowledge. Although there has not been an analysis between the ABIM-CE and the IM-ITE, the relationship between these tests could be an area for future study.

Second, because board certification has grown in importance over the years, residents may be spending more time preparing for the ABIM-CE as a result of low IM-ITE scores. Historically, residents have not necessarily prepared specifically for the IM-ITE but have prepared for the ABIM-CE.

Finally, residency programs may be providing more structured curriculum and guidance, including specific suggestions for materials and methods for preparation. This study did not address any program or individual measures that enhance or address the IM-ITE scores to improve the ABIM-CE passing rates. Further work in addressing program-specific interventions could assist in understanding which individual or combination of efforts would be most effective in improving test scores. Continuous program improvement is required by the Residency Review Committee for Internal Medicine. Program-wide efforts could help residents in their continual improvement.

Faculty might not accurately predict a resident's knowledge and related performance on the IM-ITE.[9] Identifying residents with insufficient medical knowledge is important for several reasons. First, and perhaps most important, physicians who lack adequate medical knowledge are more likely to make errors in diagnosis and perform less well clinically.[10-12] Second, successful completion of the ABIM-CE is important both for the individual resident and for the program.

The reason for poor IM-ITE performance may be that residents are truly deficient in cognitive skills, lacking sufficient content, effective access to content, or both. However, residents may have an element of poor test-taking skills, either independent of or linked to a medical knowledge deficit. Little guidance exists on how best to help these residents. Test results could help program directors with limited resources decide where to best concentrate remediation efforts among residents with low scores.

The cutoff IM-ITE scores are determined as the point on the ROC curve where both sensitivity and specificity are spontaneously maximized. The analyses show that the accuracy for these cut-points was approximately 80% to 90%. However, there is always a possibility that a person who is more likely to fail can have prepared well for the examination, or conversely a person who is more likely to pass may have a bad test-taking experience. Therefore, these cut-points are not absolutes. Residents need to know that although there is a strong correlation, no single score is fully predictive. Indeed, for the PGY-2 year, the resident with the lowest IM-ITE score passed the ABIM-CE. This observation reinforces the recommendation to use these results as guidelines and to consider the individual and program-wide efforts as well.

Although the IM-ITE has been suggested as a measure of competency in knowledge, it has been designed as a formative test of a level of knowledge. The distinction is important because criteria for competency of medical knowledge include both a level of knowledge and its application.[13] Therefore, the IM-ITE provides a guide but not a full assessment for this competency. Given that faculty prediction of knowledge and IM-ITE performance is not always accurate,[10] further work could focus on additional assessment techniques. Regardless, the IM-ITE is a valuable feedback tool, and faculty need accurate benchmarking data to guide discussions with physicians-in-training about the implications of their IM-ITE results.

This study has several limitations. First, only 4 programs were included in the analysis. However, the number was enough to accomplish the study's primary goal. The total sample of 170 residents is in line with prior studies, which included 109 and 223 residents.[2,3] Because a high proportion of residents pass the ABIM-CE (72% in this study) and a 1 standard deviation increase (25%) above the mean on the IM-ITE (35% for PGY-2) results in such a high pass rate, a sample size of 170 provides more than 99% power to determine how well IM-ITE predicts ABIM-CE pass rates.[14] Consequently, the prediction models were highly significant. More than twice this number of residents would have been required to maintain just 80% power to determine differences in pass rates between the covariate of international medical school graduate status.

Second, we can't assign causality as to why the predictive values have changed as much as this study suggests. A follow-up study looking at a larger group of programs and educational factors could verify these results.

Third, the authors are unaware of any strategies residents in this study used to prepare for the ABIM-CE differently than past residents as a result of their IM-ITE results. This study was retrospective and not designed to answer this question. Further study could be done to understand what educational interventions or instruction in test-taking strategies could be used in an identified subset of residents at risk for failure on the ABIM-CE.

Fourth, although the programs in this study had less than 100% pass rates, data from ABIM indicate that the majority of programs were in a similar situation during the years of this study, adding a broad applicability to the results.

Finally, the authors did not review prior test results, such as medical school testing and United States Medical Licensing Examination scores. However, educators should use the most proximate data for feedback, such as the IM-ITE. Two recent studies address prior testing for predicting future test performance,[15,16] and further work could address a continuum of acquisition of medical knowledge and medical reasoning skills.

The IM-ITE remains a valid predictor of performance on the ABIM-CE and continues to have construct validity for assessing medical knowledge acquired over time in a training program. The detailed results for residents in all years of residency programs presented in this study provide updated information for program directors and residents, and suggest areas for further research.

ACKNOWLEDGMENT

The authors thank Yanmei Xu for statistical analysis in earlier versions of the article and Mr. L. Grosso for ABIM-related information.

References

1. The American College of Physicians. Internal Medicine In-Training Examination. Available at: http://www.acponline.org/catalog/cme/ite.htm. Accessed April 27, 2007.
2. Grossman RS, Fincher RE, Layne RD, et al. Validity of the in-training examination for predicting American Board of Internal Medicine certifying examination scores. *J Gen Intern Med.* 1992;7:63-67.
3. Waxman H, Braunstein G, Dantzker D, et al. Performance in the internal medicine second-year residency in-training examination predicts the outcome of the ABIM certifying examination. *J Gen Intern Med*. 1994;9:692-694.
4. Garibaldi RA, Subhiyah R, Moore ME, Waxman H. The in-training examination in internal medicine: an analysis of resident performance over time. *Ann Intern Med*. 2002;137:505-510.
5. Rollins LK, Martindale JR, Edmond M, et al. Predicting pass rates on the American Board of Internal Medicine certifying examination. *J Gen Intern Med*. 1998;13:414-416.
6. American Board of Internal Medicine. ABIM Internal Medicine Certification Examination: 2002-2006 first-time taker pass rates. Available at: http://www.abim.org/resources/statcert.shtm#1. Accessed January 27, 2006.
7. Freed GL, Uren RL, Hudson EJ, et al. Policies and practices related to the role of board certification and recertification of pediatricians in hospital privileging. *JAMA*. 2006;295:905-912.
8. Freed GL, Singer D, Lakhani I, et al. Use of board certification and recertification of pediatricians in health plan credentialing policies. *JAMA*. 2006;295:913–913.
9. 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;106:206-210.
10. Bordage G. Why did I miss the diagnosis? Some cognitive explanations and educational implications. *Acad Med.* 1999; 74(10 suppl):S138-S143.
11. Tamblyn R, Abrahamowicz M, Dauphinee WD, et al. Association between licensure examination scores and practice in primary care. *JAMA*. 2002;288:3019-3026.
12. Norcini JJ, Lipner RS, Kimball HR. Certifying examination performance and patient outcomes following acute myocardial infarction. *Med Educ*. 2002;36:853-859.
13. Accreditation Council for Graduate Medical Education. Outcome project: general competencies. Available at: http://www.acgme.org/outcome/comp/compFull.asp. Accessed February 1, 2006.
14. Tosteson TD, Buzas JS, Demidenko E, Karagas M. Power and sample size calculations for generalized regression models with covariate measurement error. *Stat Med*. 2003;22:1069-1082.
15. Violato C, Donnon T. Does the medical college admission test predict clinical reasoning skills? A longitudinal study employing the medical council of Canada clinical reasoning examination. *Acad Med*. 2005;80:S14-S16.
16. Andriole DA, Jeffe DB, Hageman HL, Whelan AJ. What predicts USMLE Step 3 performance? *Acad Med*. 2005;80:S21-S24.