Concurrent Validity Between a Shared Curriculum, the Internal Medicine In-Training Examination, and the American Board of Internal Medicine Certifying Examination

Stephen D. Sisson, MD
Amanda Bertram, MS
Hsin-Chieh Yeh, PhD

Abstract

Background A core objective of residency education is to facilitate learning, and programs need more curricula and assessment tools with demonstrated validity evidence.

Objective We sought to demonstrate concurrent validity between performance on a widely shared, ambulatory curriculum (the Johns Hopkins Internal Medicine Curriculum), the Internal Medicine In-Training Examination (IM-ITE), and the American Board of Internal Medicine Certifying Examination (ABIM-CE).

Methods A cohort study of 443 postgraduate year (PGY)-3 residents at 22 academic and community hospital internal medicine residency programs using the curriculum through the Johns Hopkins Internet Learning Center (ILC). Total and percentile rank scores on ILC didactic modules were compared with total and percentile rank scores on the IM-ITE and total scores on the ABIM-CE.

Results The average score on didactic modules was 80.1%; the percentile rank was 53.8. The average IM-ITE score was 64.1% with a percentile rank of 54.8. The average score on the ABIM-CE was 464. Scores on the didactic modules, IM-ITE, and ABIM-CE correlated with each other ($P < .05$). Residents completing greater numbers of didactic modules, regardless of scores, had higher IM-ITE total and percentile rank scores ($P < .05$). Resident performance on modules covering back pain, hypertension, preoperative evaluation, and upper respiratory tract infection was associated with IM-ITE percentile rank.

Conclusions Performance on a widely shared ambulatory curriculum is associated with performance on the IM-ITE and the ABIM-CE.

Introduction

Evaluating residents is best done with methods that are both feasible and psychometrically robust.[1] The Accreditation Council for Graduate Medical Education (ACGME) has provided a toolbox of methods for program directors to incorporate into their evaluation processes.[2] Advocates of the toolbox call for the development of additional measures of resident assessment.[3,4] Ideally, these tools should have validity evidence of their assessment results.

One tool of resident assessment with validity evidence is the Internal Medicine In-Training Examination (IM-ITE). The IM-ITE assesses knowledge among internal medicine residents.[5–8] Validity evidence on the IM-ITE includes high performance on tests of internal consistency, improvement in scores among advanced trainees, and association with American Board of Internal Medicine Certifying Examination (ABIM-CE) performance. As a result, the IM-ITE is a valuable assessment tool for program directors and is used annually in most programs.[5–8]

With the advent of competency-based education and milestone interval evaluations, residency training programs need more assessment tools with validity evidence. One type of validity evidence is the correlation of assessment results with other instruments that have good validity evidence (ie, concurrent validity).[9] The Johns Hopkins Internal Medicine Curriculum is a widely used curriculum on topics in ambulatory care distributed online via the Johns Hopkins Internet Learning Center (ILC).[10,11] Education outcomes on the ILC undergo extensive reliability testing and gathering of validity evidence. Trainees are provided with real-time feedback on their performance,

All authors are at Johns Hopkins University School of Medicine. **Stephen D. Sisson, MD,** is Associate Professor, Division of General Internal Medicine, Department of Medicine; **Amanda Bertram, MS,** is Senior Research Program Coordinator, Division of General Internal Medicine, Department of Medicine; and **Hsin-Chieh Yeh, PhD,** is Associate Professor, Departments of Medicine and Epidemiology.

Funding: The authors report no external funding source for this study.

Conflict of interest: Dr Sisson receives an annual stipend for editorial duties related to the Johns Hopkins Internet Learning Center.

Corresponding author: Stephen D. Sisson, MD, 601 N. Caroline Street, Room 7150G, Baltimore, MD 21287, 410.955.6738, fax 410.614.1195, ssisson@jhmi.edu

Received January 24, 2014; revisions received May 28, 2014, and August 19, 2014; accepted September 15, 2014.

DOI: http://dx.doi.org/10.4300/JGME-D-14-00054.1

including a ranking relative to others at the same level of training.[12,13] We hypothesized that individual performance on the ILC would be an indicator of individual performance on 2 key benchmarks of medical knowledge, the IM-ITE and the ABIM-CE, and compared knowledge outcomes data to determine potential correlations.

What was known and gap

Effective resident education and evaluation require validated curricula and assessment tools.

What is new

Assessment on a widely shared internal medicine ambulatory care curriculum was correlated with performance on the in-training examination and the American Board of Internal Medicine (ABIM) examination.

Limitations

Participation rate was low, raising the potential for selection bias.

Bottom line

Performance on the widely shared ambulatory curriculum was associated with performance on the in-training examination and the ABIM examination.

Methods

We performed a cohort study of postgraduate year (PGY)-3 residents at internal medicine residency training programs who subscribed to the ILC Internal Medicine Curriculum in the 2009–2010 academic year. The ILC Internal Medicine Curriculum consists of 41 modules on topics in ambulatory care, including chronic disease management (eg, diabetes, hypertension, depression); acute symptom management (eg, headache, back pain); and preventive care (eg, cancer screening, immunizations).[13] Modules consist of a pretest, a didactic section, and a posttest, and are disseminated online. On the ILC pretests and posttests, item discrimination is performed on each test item, and Cronbach α is performed on each test using the method of Ferguson and Takane.[14] For the purpose of this study, we analyzed posttest performance. We calculated a 2-digit percentile rank on each PGY-3 posttest score, relative to other PGY-3 residents who have completed that same module, by calculating the mean and standard deviation of scores on the module and determining the standardized score for a particular resident. This was then converted to a 2-digit percentile rank. An overall 2-digit percentile rank is calculated for each resident by determining the mean and standard deviation of all the individual percentile rank scores and comparing the individual resident's average percentile rank with the mean and standard deviation of all percentile rank scores.

In the 2009–2010 academic year, there were 109 internal medicine residency training programs using the ILC curriculum, and 3 attempts were made to contact each program by e-mail or by phone to participate in the study. Of those 109 programs, 38 programs responded, with 22 (20.2%) agreeing to participate and 16 (14.7%) declining to participate. Programs that did not respond were categorized as declining to participate. Programs that agreed to participate were sent a score sheet of each PGY-3 resident in their program, containing his or her ILC module scores and percentile rank scores (or no scores for those residents who had not completed any modules). Programs were asked to enter the IM-ITE total score and percentile rank and the ABIM-CE total score and total percentile rank for each resident, and then to remove the resident name from the score sheets to preserve anonymity. Several programs expressed confusion about which score on the ABIM-CE report represented the total percentile rank, and as a result, the ABIM-CE total percentile rank was deleted from the study. The IM-ITE results are always available to training programs, whereas residents may decline to share their ABIM-CE scores with their training program. As a result, we received fewer ABIM-CE scores than IM-ITE scores.

The study was approved by the Johns Hopkins Medicine Institutional Review Board.

Statistical Analysis

The distribution of resident performance on the IM-ITE, the ABIM-CE, and the ILC modules were summarized as means and standard deviations. Associations between IM-ITE, ABIM-CE, and ILC module performance were examined by the pairwise Pearson correlation. We used a linear regression model to determine whether there was any association between the number of modules completed and performance on the IM-ITE and ABIM-CE. β-Coefficients were calculated to estimate the effect on performance per module increase. The number of ILC modules completed was also grouped into thirds for comparison. Student *t* test was used to compare mean IM-ITE total and rank score as well as ABIM-CE total scores in residents completing 2 to 7 modules and 8 or more modules to residents completing 0 or 1 module. In addition, we used a nonparametric test of trends for the ranks of across-ordered groups (0–1, 2–7, and 8+ modules). The test is an extension of the Wilcoxon rank-sum test. Finally, we investigated whether a resident's performance on a specific topic correlated with IM-ITE rank performance. For this analysis, pairwise Pearson correlations were calculated to assess the strength of associations on modules completed by at least 100 residents. We also performed linear regression with IM-ITE rank as the dependent variable and module performance as

TABLE 1 **Resident Performance on the Internal Medicine In-Training Examination (IM-ITE), the American Board of Internal Medicine Certifying Examination (ABIM-CE), and the Johns Hopkins Internet Learning Center (ILC) Modules**

Test	Residents, No. (%)	Average Result (±SD)
IM-ITE total score	305 (68.8)	64.1% (7.8)
IM-ITE rank score	323 (72.9)	54.8 (34.7)
ABIM-CE total score	182 (41.1)	464 (110.4)
ILC module total score	313 (70.7)	80.0% (11.7)
ILC module rank score	313 (70.7)	53.8 (34.7)

the independent variable. Effect size was calculated using the R^2 in regression as the proportion of shared variability between the 2 variables.[15] All tests of significance were 2 tailed, with an α level of .05. Analyses were performed using Stata/SE version 12.0 (StataCorp LP).

Results

Respondent Characteristics

Reports were received from 22 internal medicine residency training programs, including 16 community hospitals (72.7%) and 6 academic medical centers (27.3%). There were 506 PGY-3 residents at the 22 programs, and we received score reports for 443 (87.5%) of them. We received IM-ITE scores on 305 residents (68.8%), and rank IM-ITE scores on 323 (72.9%). We received ABIM-CE scores on 182 (41.1%) residents. Of the 443 residents, 313 (70.7%) completed at least 1 module, and 130 (29.3%) completed no modules. The mean number of modules completed by residents was 7.5. Mean and rank scores are shown in TABLE 1.

Respondents Versus Nonrespondents

We compared module performance between the 22 participating and the 87 nonparticipating programs. The average module score on all modules among participating programs was 80.1% and among nonparticipating programs was 81.3%, a difference that was not statistically significant ($P = .11$). The average resident rank score also did not differ between participating programs and nonparticipating programs (54.0 versus 54.4, $P = .82$). At participating programs, rank score did not differ among residents for whom we had IM-ITE scores and those for whom we did not (53.7 versus 55.2, $P = .61$), nor between those residents for whom we had ABIM-CE scores and for those for whom we did not (52.3 versus 55.2, $P = .27$).

Associations

Associations among IM-ITE, ABIM-CE, and ILC module performance are shown in TABLE 2. In a post hoc analysis using linear regression with ABIM-CE total score as the dependent variable and IM-ITE rank score as the independent variable, R^2 was 23%. Adding the ILC module total score to that model improved the R^2 to 26%, but that result was not statistically significant ($P = .67$).

We next looked for correlations among the number of modules completed (regardless of performance on those modules) and IM-ITE total and percentile rank scores. When we categorized the number of modules completed by tertiles (0 to 1 module completed; 2 to 7 modules completed; 8 or more modules completed), we found that when a resident completed at least 8 modules, scores correlated with the IM-ITE total ($P < .01$) and

TABLE 2 **Pairwise Correlations Among the Internal Medicine In-Training Examination (IM-ITE), the American Board of Internal Medicine Certifying Examination (ABIM-CE), and the Johns Hopkins Internet Learning Center (ILC) Module Performance**

	IM-ITE Total Score	IM-ITE Rank Score	ABIM-CE Total Score	ILC Module Total Score	ILC Module Rank Score
IM-ITE total score	1.000				
IM-ITE rank score	0.93[a]	1.00			
ABIM-CE total score	0.49[a]	0.50[a]	1.000		
ILC module total score	0.26[a]	0.26[a]	0.16[b]	1.00	
ILC module rank score	0.25[a]	0.25[a]	0.18[b]	0.94[a]	1.00

[a] $P < .01$.
[b] $P < .05$.

TABLE 3 **Associations Among the Johns Hopkins Internet Learning Center (ILC) Modules Completed, the Internal Medicine In-Training Examination (IM-ITE) Performance, and the American Board of Internal Medicine Certifying Examination (ABIM-CE) Performance**

	Mean (SD)	Mean Difference (95% CI)
IM-ITE rank		
No module	50.7 (24.9)	Reference
Any module	54.2 (27.5)	3.5 (−11.7 to 4.7)
0–1 modules	48.0 (26.6)	Reference
2–7 modules	52.1 (29.2)	4.1 (−4.4 to 12.7)
8+ modules	56.8 (25.8)	8.8 (1.1 to 16.6)[a]
Continuous, per module increase	β-Coefficient = 0.24 (−0.02 to 0.5)	
IM-ITE total		
No module	62.6 (6.4)	Reference
Any module	64.7 (7.5)	1.16 (−4.4 to 0.18)
0–1 modules	61.9 (6.9)	Reference
2–7 modules	63.9 (8.1)	2.0 (−0.6 to 4.5)
8+ modules	65.0 (7.9)	3.1 (0.9 to 5.4)[a]
Continuous, per module increase	β-Coefficient = 0.07 (−0.003 to 0.14)	
ABIM-CE		
No module	446.4 (173.1)	Reference
Any module	465.9 (101.9)	19.5 (−73.7 to 34.6)
0–1 modules	440.6 (151.3)	Reference
2–7 modules	456.2 (97.5)	15.6 (−38.9 to 70.0)
8+ modules	472.5 (105.4)	31.9 (−17.2 to 81.0)
Continuous, per module increase	β-Coefficient = 1.5 (−0.03 to 3.1)	

[a] $P < .05$ as compared with the reference group using Student t test; $P < .05$ in nonparametric test of trends for the ranks of across-ordered groups (0–1, 2–7, and 8+ modules).

percentile rank scores ($P = .03$). The IM-ITE total and percentile rank scores improved with greater numbers of modules completed (TABLE 3). Mean IM-ITE percentile rank scores increased in residents who completed more modules, relative to those who completed fewer (for the trend, $P = .03$). Although ABIM-CE scores also improved with additional modules completed, these differences were not significant ($P = .07$).

Finally, among the 18 modules completed by at least 100 of the 323 residents with IM-ITE rank scores, performance on back pain, hypertension, preoperative evaluation, and upper respiratory tract infection modules was statistically associated with IM-ITE rank scores ($P < .05$; TABLE 4).

Discussion

We showed that evaluative data generated by an interactive ambulatory curriculum have concurrent validity with IM-ITE and ABIM-CE performance. For an individual learner, ILC module performance correlated with IM-ITE performance when at least 8 modules had been completed. Higher numbers of completed modules were associated with better performance on the IM-ITE. The Johns Hopkins Ambulatory Curriculum thus offers evaluative information that may predict performance on the IM-ITE and ABIM-CE.

It is not perfectly explained why performance on an ambulatory curriculum correlates with performance on tests that broadly cover internal medicine. The IM-ITE test blueprint assigns questions to general internal medicine,[8]

TABLE 4 **Correlation Between Module Topic and the Internal Medicine In-Training Examination Performance**

Module Topic	Pearson Correlation	Effect Size[a] Using R^2, %
Upper respiratory tract disorders	0.30[b]	9
Back pain	0.22[c]	5
Preoperative assessment	0.21[c]	4
Hypertension	0.20[c]	4
Cancer screening	0.19[c]	3
Anemia	0.19	3
Chronic kidney disease	0.20	4
Illicit/prescription drug abuse	0.15	2
Gynecology	0.13	2
Headache	0.13	2
Smoking	0.13	2
Hip/knee pain	0.12	2
Osteoporosis	0.10	1
Lipid management	0.03	0.9
Dermatitis	0.02	1
Diabetes	0.02	0.04
Professionalism	−0.02	0.05
Thyroid disorders	−0.08	1

[a] Effect size was calculated using the *R* square in regression as the proportion of shared variability between the 2 variables.
[b] $P < .01$.
[c] $P < .05$.

whereas the ABIM-CE test blueprint does not assign a specific portion of content to general internal medicine.[16] It could be that ILC performance is an indicator of something other than specific knowledge on ambulatory care topics. Residents who perform well on the ILC may have self-directed study habits that cover all topics in internal medicine, and research has shown that self-directed reading is associated with better IM-ITE performance.[17] Our findings are similar: Regardless of module performance, those residents who completed greater numbers of modules performed better on the IM-ITE. The ILC module completion likely serves as a marker of resident study habits.

We also found that, for some ILC module topics (ie, back pain, hypertension, preoperative assessment, upper respiratory tract infections), individual performance was associated with IM-ITE performance. The reasons are unclear. Preoperative assessment is a general topic that requires comprehension of cardiovascular risk and other comorbidities and might demonstrate broad comprehension of internal medicine. However, this could not be said for knowledge of back pain or upper respiratory tract infections.

If a major thrust of evaluating residents is to determine who is competent to take care of patients, assessment tools must have validity evidence.[1] We developed and provided validity evidence of an ambulatory curriculum that can enhance a program director's ability to evaluate residents, and demonstrated the feasibility of establishing validity evidence of an evaluation instrument by testing correlations between its results with those of the IM-ITE and ABIM-CE.

Our study has several limitations. Most programs using the curriculum declined to participate, introducing possible selection bias. However, module performance did not differ at participating and nonparticipating programs. We did not have access to IM-ITE and ABIM-CE results from nonresponding programs to compare performance on those metrics. We only looked at PGY-3 learners, and it is possible that ILC performance among PGY-2 and PGY-1 learners does not associate with IM-ITE or ABIM-CE performance. We did not assess pretest performance and, thus, could not assess the effect of the curriculum itself on IM-ITE or ABIM-CE performance. We also did not assess association of ILC module performance with clinical outcomes, which would provide very powerful validity evidence of the ILC as an assessment tool.

Conclusion

Our study showed that performance on a widely shared ambulatory curriculum for internal medicine residents was associated with performance on the IM-ITE and the ABIM-CE.

References

1 Swing SR. The ACGME outcome project: retrospective and prospective. *Med Teach*. 2007;29(7):648–654.
2 Accreditation Council for Graduate Medical Education. ACGME outcomes project toolbox of assessment methods. http://www.dconnect.acgme.org/outcome/assess/toolbox.pdf. Accessed November 12, 2014.
3 Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: a systematic review. *Acad Med*. 2009;84(3):301–309.
4 Green ML, Holmboe E. Perspective—the ACGME toolbox: half empty or half full? *Acad Med*. 2010;85(5):787–790.
5 Babbott SF, Beasley BW, Hinchey KT, Blotzer JW, Holmboe ES. The predictive validity of the Internal Medicine In-Training Examination. *Am J Med*. 2007;120(8):735–740.
6 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):206–210.
7 Wang H, Nugent R, Nugent C, Nugent K, Phy M. A commentary on the use of the Internal Medicine In-Training Examination. *Am J Med*. 2009;122(9):879–883.
8 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(6):505–510.

9 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):e7–e16.

10 Sisson SD, Hughes MT, Levine D, Brancati FL. Effect of an Internet-based curriculum on post-graduate education: a multicenter intervention. *J Gen Intern Med*. 2004;19(5, pt 2):503–507.

11 Sisson SD, Rastegar DA, Rice TN, Hughes MT. Multicenter implementation of a shared graduate medical education resource. *Arch Intern Med*. 2007;167(22):2476–2480.

12 Sisson SD, Rastegar DA, Hughes MT, Bertram AK, Yeh HC. Learner feedback and educational outcomes with an Internet-based ambulatory curriculum. *BMC Med Educ*. 2012;12:55. doi:10.1186/1472-6920-12-55.

13 Sisson SD, Dalal D. Internal medicine residency training on topics in ambulatory care: a status report. *Am J Med*. 2011;124(1):86–90.

14 Ferguson GA, Takane Y. *Statistical Analysis in Psychology and Education*. 6th ed. New York, NY: McGraw-Hill Book Company; 1989.

15 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care*. 1989;27(suppl 3):178–189.

16 American Board of Internal Medicine. Internal medicine certification examination blueprint. http://www.abim.org/pdf/blueprint/im_cert.pdf. Accessed November 12, 2014.

17 McDonald FS, Zeger SL, Kolars JC. Factors associated with medical knowledge acquisition during internal medicine residency. *J Gen Intern Med*. 2007;22(7):962–968.