A Multiple Choice Testing Program Coupled with a Year-long Elective Experience is Associated with Improved Performance on the Internal Medicine In-Training Examination

Bradley R. Mathis, MD[1], Eric J. Warm, MD[1], Daniel P. Schauer, MD, MSc[1], Eric Holmboe, MD[2], and Gregory W. Rouan, MD[1]

[1]Department of Internal Medicine, University of Cincinnati College of Medicine, Cincinnati, OH, USA; [2]American Board of Internal Medicine, Philadelphia, PA, USA.

BACKGROUND: The Internal Medicine In-Training Exam (IM-ITE) assesses the content knowledge of internal medicine trainees. Many programs use the IM-ITE to counsel residents, to create individual remediation plans, and to make fundamental programmatic and curricular modifications.

OBJECTIVE: To assess the association between a multiple-choice testing program administered during 12 consecutive months of ambulatory and inpatient elective experience and IM-ITE percentile scores in third post-graduate year (PGY-3) categorical residents.

DESIGN: Retrospective cohort study.

PARTICIPANTS: One hundred and four categorical internal medicine residents. Forty-five residents in the 2008 and 2009 classes participated in the study group, and the 59 residents in the three classes that preceded the use of the testing program, 2005–2007, served as controls.

INTERVENTION: A comprehensive, elective rotation specific, multiple-choice testing program and a separate board review program, both administered during a continuous long-block elective experience during the twelve months between the second post-graduate year (PGY-2) and PGY-3 in-training examinations.

MEASURES: We analyzed the change in median individual percent correct and percentile scores between the PGY-1 and PGY-2 IM-ITE and between the PGY-2 and PGY-3 IM-ITE in both control and study cohorts. For our main outcome measure, we compared the change in median individual percentile rank between the control and study cohorts between the PGY-2 and the PGY-3 IM-ITE testing opportunities.

RESULTS: After experiencing the educational intervention, the study group demonstrated a significant increase in median individual IM-ITE percentile score between PGY-2 and PGY-3 examinations of 8.5 percentile points (p<0.01). This is significantly better than the increase of 1.0 percentile point seen in the control group between its PGY-2 and PGY-3 examination (p<0.01).

CONCLUSION: A comprehensive multiple-choice testing program aimed at PGY-2 residents during a 12-month continuous long-block elective experience is associated with improved PGY-3 IM-ITE performance.

INTRODUCTION

The Internal Medicine In-Training Exam (IM-ITE), given annually in October, is a validated examination that assesses internal medicine resident knowledge acquisition during training.[1] The IM-ITE is targeted at the second post-graduate year (PGY-2) residents but is available to those in all levels of training. Individuals receive two scores, percent correct and a percentile score comparing their performance to that of peers nationally. In addition to scores, the individual report also contains an outline of deficiencies to guide further study. An individual's performance on the IM-ITE correlates with performance on the American Board of Internal Medicine-Certifying Exam (ABIM-CE).[2–5] Performance on standardized testing also correlates with the quality of care provided in several domains.[6–8] Residency program directors receive feedback from the IM-ITE regarding their program's performance in comparison to programs across the country.[1] Programs often use the IM-ITE to counsel trainees, create individual remediation plans, and change programmatic curricular emphasis.

Similar to national trends, University of Cincinnati categorical residents improve their annual percent correct scores as they advanced through the program.[9] However, despite the increase in percent correct, IM-ITE scores for University of Cincinnati categorical residents have declined in percentile rank compared to national scores as trainees progressed from PGY-1 to the PGY-3 level. The judgment of the University of Cincinnati's Internal Medicine Graduate Medical Education Committee was that this decline in percentile illustrated a deficit in medical knowledge acquisition and a weakness in our curriculum, and may have left our residents at a competitive disadvantage nationally not only in successful ABIM certification but also in clinical competence.[8] Deliberate educational interventions have been shown to improve medical knowledge acquisition and ITE scores in internal medicine and surgical residents.[10–12] We designed a rotation specific multiple-choice testing program and a separate board review testing program administered during twelve consecutive months of ambulatory

and inpatient elective experience with the intent of improving medical knowledge acquisition. The purpose of this study was to determine the association between these educational interventions and IM-ITE percentile scores in third post-graduate year (PGY-3) categorical residents.

METHODS

Setting

The University Hospital/University of Cincinnati internal medicine residency is one of the 17 initial sites chosen to participate in the Residency Review Committee for Internal Medicine's Educational Innovation Project (EIP) in July 2006. This recognition allowed us to create innovative ways to train physicians without the need to adhere to many of the traditional guidelines of the Accreditation Council of Graduate Medical Education (ACGME).[13] The main focus of our EIP has been the creation of the long-block.[14] The long-block is 12 consecutive months of residency (November of the PGY-2 year through October of the PGY-3 year) consisting of ambulatory care, inpatient and outpatient electives, and research experiences with minimal overnight call. Each elective experience has a defined set of goals and objectives created by the internal medicine key education faculty and the subspecialty specific educational coordinators.

Educational Interventions

For each elective rotation during the long-block, we developed pre- and post-rotation multiple choice tests using examinations of 30–50 questions each. We created the exams from the American College of Physician's Medical Knowledge Self Assessment Program (MKSAP)® 12 and 13. We chose questions based on the previously defined set of goals and objectives for each rotation. Topics included cardiology, gastroenterology, hepatology, nephrology, hematology-oncology, endocrinology, pulmonary medicine and rheumatology. Residents took the pre-test during the first week of their rotation and received confidential feedback regarding their performance shortly thereafter. The residents were encouraged to discuss areas of weakness with their elective rotation attending to create a personalized learning plan for the month. During the last week of the rotation, each resident received and completed a post-test that used different questions than the pretest, but covered similar curricular areas. We then asked residents to reflect on progress made during the elective rotation, and to create further study plans for persistently deficient areas.

Additionally, in collaboration with the American Board of Internal Medicine (ABIM), we created six additional, 60 question multiple-choice examinations based upon the self-assessment (Self Evaluation Program-SEP) modules in the Maintenance of Certification (MOC) program. These questions were not originally designed as preparation for the certification exam, but rather to improve knowledge and encourage learners to investigate items about which they had questions. Each test comprised of questions taken from the ABIM self-assessment SEP question bank and pertained to the internal medicine subspecialties above as well as infectious disease and general internal medicine topics. We administered these modules as closed-book examinations at two-month intervals in a proctored environment with a two-hour time limit for completion. The tests were again graded confidentially and the answers were returned to the trainees during a two-hour learner driven debriefing discussion the following week. The debriefing sessions were used to address gaps in knowledge and multiple-choice test taking strategies, and to give advice regarding timing and pace.

Both the pre- and post-rotation multiple choice tests and the scores on the ABIM preparatory exams were tracked over time and used for formative and summative feedback as part of a larger long-block multisource evaluation.[15] Residents identified by the program administration as being deficient in medical knowledge were required to meet with the administration to develop a personalized learning plan. These plans varied based on learning styles but included scheduled text reading and review, review of exams and test taking skill, peer guided study groups and introduction to alternate learning methods such as video and audio taped reviews. Residents met with administration at least quarterly but had the opportunity to meet more often.

The long-block and entire testing program occurred between the administration of the PGY-2 and PGY-3 IM-ITE.

Participants

The study group consisted of categorical residents in the first two classes to complete the long-block experience (classes of 2008 and 2009). This group was exposed to the pre- and post-rotation multiple-choice tests, the ABIM preparatory exams, the 12-month long-block experience, the IM-ITE itself and personalized learning plans as needed. The control group consisted of the categorical residents in the three resident classes that preceded the initiation of the long-block and the testing program (classes of 2005 through 2007). This group was exposed only to the IM-ITE and personalized learning plans as needed. The study was granted exemption by the University of Cincinnati IRB.

Statistical Analysis

We used descriptive statistics, including means and medians, to summarize the data. To characterize our sample, comparisons between the study group and control group were made using the t-test and Chi-square test where appropriate. We calculated the individual change in both percent correct score and national percentile score between the PGY-1 and PGY-2 testing opportunities and between the PGY-2 and PGY-3 testing opportunities in both the control and study cohorts. To compare changes between the PGY-1 and PGY-2 testing opportunities and between the PGY-2 and PGY-3 testing opportunities both within and between the study and control cohorts, we used the Wilcoxon signed-rank test for both the percent correct score and national percentile score. For our main outcome measure, assessing the effectiveness of the testing program administered during the twelve month long-

block experience, we compared the change in the IM-ITE percentile from the PGY-2 to the PGY-3 years between the study and control cohorts using the Wilcoxon rank-sum test. SAS version 9.2 (Cary, NC) was used for all analyses.

RESULTS

We included all 45 post-intervention residents in the study group and all 59 residents from the preceding three years as controls in the final analysis. Each categorical resident was given the opportunity to take the IM-ITE during October in each year of training. The study group, classes of 2008 and 2009, completed a total of 127 out of 135 IM-ITE testing opportunities during their three years of residency (94%). The control group, classes of 2005–2007, completed 160 out of 177 IM-ITE testing opportunities during their three years of residency (90%). Participant demographics and United States Medical Licensing Examination (USMLE) 1 and 2 scores were similar between the two groups (Table 1). When included as a covariate in the analysis, prior USMLE scores did not significantly affect the results.

Neither the three residency classes of the control group nor the two residency classes of the study group were exposed to the long-block and its testing program before the administration of the PGY-2 IM-ITE.

The control cohort showed an improvement in median individual percent correct of six points when comparing the raw score results of the PGY-1 IM-ITE and the PGY-2 IM-ITE (p< 0.01) and 5 points when comparing the raw score results of the PGY-2 IM-ITE and the PGY-3 IM-ITE (p<0.01). Similarly, the study cohort showed an improvement in median individual

percent correct of 5 points when comparing the raw score results of PGY-1 IM-ITE and the PGY-2 IM-ITE (p<0.01) and of six points when comparing the raw score results of the PGY-2 IM-ITE and PGY-3 IM-ITE (p<0.01). There were no significant differences between the study and control cohort with respect to change in percent correct (Fig. 1).

Despite the increase in percent correct, the control group demonstrated a decline in median individual percentile rank of 5.5 percentile points (p=0.02) and the study group demonstrated a decline in median individual percentile rank of 3.0 percentile points (p=0.01) when comparing the results of their PGY-1 IM-ITE to the results of their PGY-2 IM-ITE (Fig. 2).

The control group demonstrated an insignificant increase in median individual percentile rank of 1.0 percentile point (p= 0.58) when comparing the results of the PGY-2 IM-ITE to the results of the PGY-3 IM-ITE (Fig. 2).

After exposure to the long-block and its testing program, the study group demonstrated an increase in median individual percentile rank of 8.5 percentile points (p<0.01) when comparing results of the PGY-2 IM-ITE to the results of the PGY-3 IM-ITE (Fig. 2).

Comparing study group to control regarding the median individual change in performance on the PGY-3 IM-ITE there was statistically significant improvement of 7.5 percentile points by the study group (p=0.01).

DISCUSSION

We designed a rotation-based testing and board review program as part of our newly implemented twelve-month long-block. Our

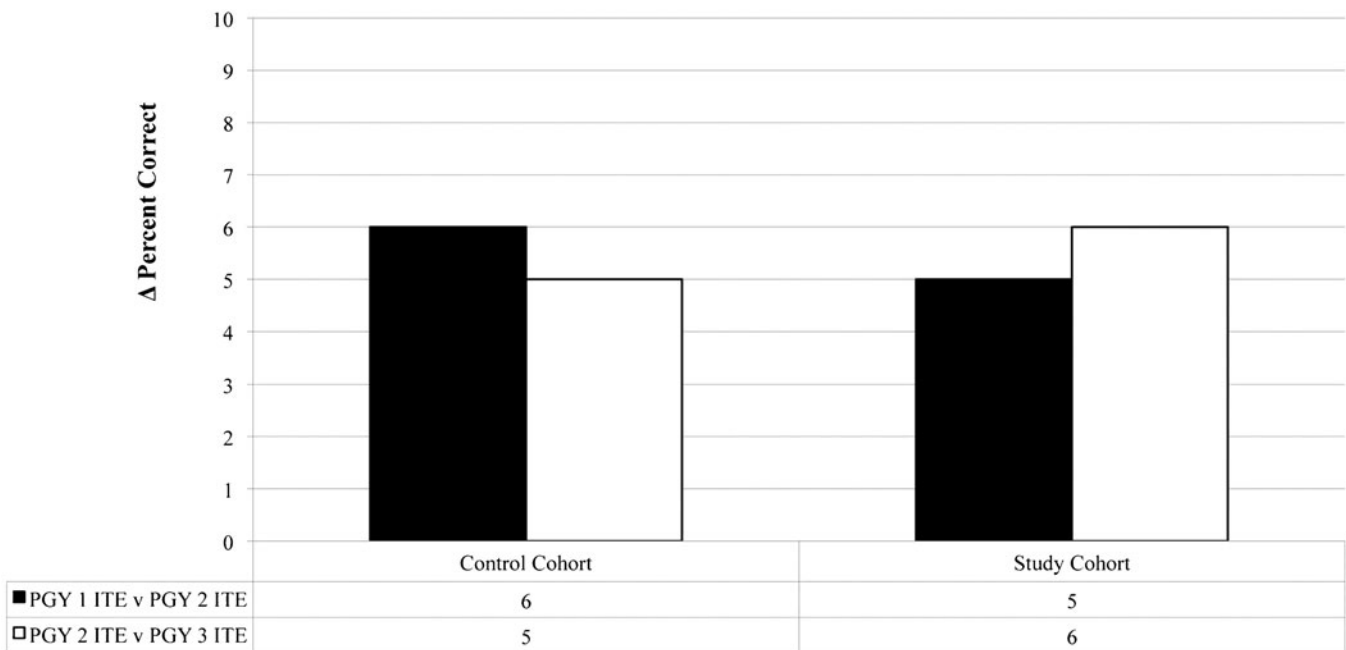|	Control Cohort	Study Cohort
■ PGY 1 ITE v PGY 2 ITE	6	5
□ PGY 2 ITE v PGY 3 ITE	5	6

Figure 1. Change in Median Individual ITE Percent Correct. Both control and study cohort demonstrate improvement in raw scores between the PGY-1 and PGY-2 exams and the PGY-2 and PGY3 exams (p<0.01, Wilcoxon signed-rank test). There were no significant differences between the study and control cohorts using the Wilcoxon rank-sum test.

	Before Testing Program (n=59)	After Testing Program (n=45)	p-value[a]
Mean USMLE-1 score	221 (19.3)	219 (18.6)	0.80
Mean USMLE-2 score	221 (21.8)	222 (19.8)	0.53
Gender (% male)	51%	49%	0.84
IMG (%)	29%	31%	0.80

USMLE=United States Medical Licensing Examination
IMG=International Medical Graduate
[a]*by t-test or Chi-square where appropriate*

goal was to improve medical knowledge acquisition in our categorical residents. Before the testing intervention we had seen a decline in our average individual percentiles when compared year to year despite increasing absolute scores. Our cumulative three-year ABIM-CE pass rates were consistently in the 25th to 50th percentile nationally. We postulated that this decline in IM-ITE percentile and our suboptimal ABIM-CE pass rate was due in part to exposure to only a narrow variety of inpatient diagnoses seen in our traditional residency program thereby producing residents who lacked breadth in knowledge of internal medicine. We chose improvement on the IM-ITE as the measure of our success as the IM-ITE's reliability allows residents and programs across the country to measure individual improvement from year to year.[5]

Although the IM-ITE changes somewhat from year to year and comparing non-concurrent cohorts is imperfect, there is substantial similarity in the testing cohorts as 92% of internal medicine residents participate in the IM-ITE annually.[5] During our study period alone there were two changes in the administration of the examination. In 2003 the total number of questions decreased from 360 to 340 and in 2008 the time to take the test expanded from 7 to 8 hours. However, analysis of annual results undertaken by the administrators of the examination demonstrated no change in the year-to-year reliability of the IM-ITE despite these changes (S. McKinney, personal communication 2011). Additionally, the changes in number of questions and length of testing time may affect the percent correct scores but they should not affect percentile rank data as the entire national cohort is exposed to such changes. We chose to use change in individual percentile to normalize the scores to the national cohort and to allow for annual differences in test characteristics.

Our study documents that exposure to a testing program based on a broad spectrum of inpatient and outpatient internal medicine topics during a 12-month continuous long-block elective experience is associated with improvement in participants' IM-ITE percentile rank. Although our study design does not allow us to prove direct causality, or determine the relative contributions of test taking skills, medical knowledge acquisition and elective experiences, we postulate the improvement in IM-ITE scores is due to several factors. We believe exposure to a broad range of medical information in a structured learning environment provided the feedback and tools necessary for individual improvement in medical knowledge. Additionally our
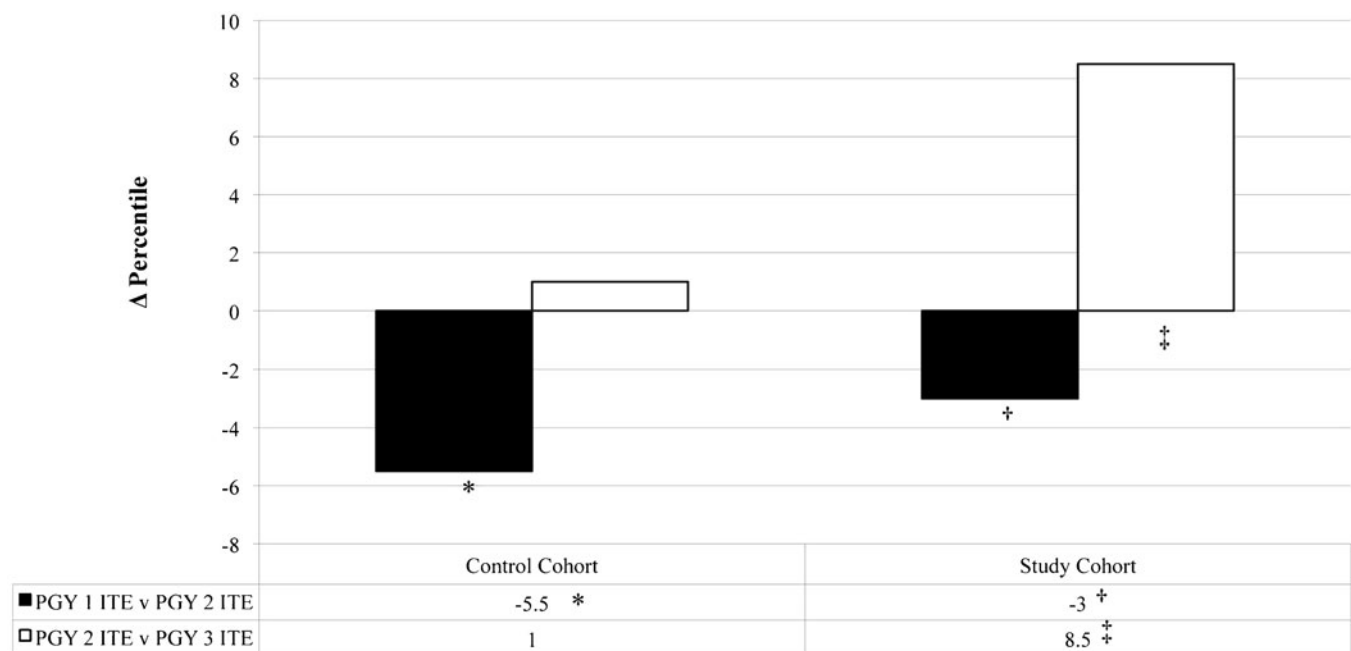

Figure 2. Change in Median Individual ITE Percentile Before and After PGY2 Testing Program Institution. The control cohort demonstrates a decline in IM-ITE percentile scores between the first and second postgraduate years (*p=0.02) and did not change between the second and third postgraduate years (p=0.56). Before exposure to the testing program, the study cohort demonstrates a decline in IM-ITE percentile scores between the first and second postgraduate years (†p=0.01). After exposure to the testing program, the study cohort demonstrates an increase in IM-ITE percentile scores between the second and third postgraduate years (‡p<0.01). All comparisons were made using the Wilcoxon signed-rank test.

testing program provided a high volume of multiple-choice questions in an environment requiring time management and proper pacing. The board review sessions not only discussed question content but also addressed strategies for answering multiple-choice questions. Finally, the testing program was contained within a one-year block during which the residents have exposure to a wide variety of elective experience with minimal inpatient service and on-call responsibilities. This schedule results in more elective exposure and a higher amount of unencumbered study time than previously afforded to our residents.

The improvement seen in standardized testing by our cohort may translate into improvement in both future standardized testing performance and future quality of clinical care.[6–8] Success on the IM-ITE predicts success on the ABIM-CE.[2–5] In addition, higher achievement on standardized tests in internal medicine predicts higher quality in diabetes care, malignancy screening and myocardial infarction therapy.[6,7] The historical "Gold Standard" cut point for predicting passage of the ABIM-CE with a positive predictive value (PPV) of 89% is the 35th percentile on the PGY-2 IM-ITE.[2] A more recent study showed that a PGY-3 resident scoring in the 21st percentile on the IM-ITE predicts passing the ABIM-CE with a PPV of 97%.[5] The improvement in scores demonstrated by our residents may ensure that they surpass such predictive cut-points.

Several limitations are inherent to our study. First, we are an EIP program and have dedicated an entire 12-month block to inpatient and outpatient elective experience allowing for emphasis on personal and professional improvement. This type of concentrated effort may not be possible in a traditional residency program. However, elements of our program, such as pre- and post-rotation testing are exportable to any program. Second, the study only addresses the competences of medical knowledge and patient care. Other programs may be interested in addressing more than these two competencies when adopting new programs. Third, the self-assessment SEP modules in the ABIM-MOC program were not originally designed for preparation for the certification exam. However, we used them as designed to improve knowledge and encourage learners to investigate items about which they had questions during our post-exam debriefing sessions. Finally we did not control for conference attendance or UpToDate® utilization, both found, in prior analyses, to be predictors of improved IM-ITE scores.[11,12] Any effect on our results should be minimal as access to these resources has been present for a number of years and was neither changed nor limited during the study period.

CONCLUSION

We believe that a comprehensive multiple-choice testing program aimed at PGY-2 residents during our long-block is associated with improved PGY-3 IM-ITE performance. Further research will be needed to demonstrate the relative contribution of test taking skills, medical knowledge acquisition and elective experiences on this outcome, and to determine if this improved performance translates into higher ABIM-CE scores and better long-term clinical performance. Future study and inclusion of other residency sites with a focus on impact on clinical performance may enhance the program's generalizability.

Conflict of Interest: *Dr. Eric Holmboe is employed by the ABIM.*

Corresponding Author: *Bradley R. Mathis, MD; University of Cincinnati College of Medicine, 231 Albert Sabin Way, Cincinnati, OH 45267-0535, USA (e-mail: mathisbr@ucmail.uc.edu).*

REFERENCES

1. The American College of Physicians. Internal Medicine In-Training Examination. Available at: http://www.acponline.org/education_recertification/education/in_training/#3 Accessed March 3, 2011.
2. **Grossman RS, Fincher RE, Layne RD, Seelig CB, Berkowitz LR, Levine MA.** Validity of the In-Training exam 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 on the internal medicine second-year In-training examination predicts the outcome of the ABIM certifying exam. J Gen Intern Med. 1994;9:692–694.
4. **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;13:414–416.
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. **Holmboe E, Wang Y, Meehan T, et al.** Association between maintenance of certification examination scores and quality of care for Medicare beneficiaries. Arch Intern Med. 2008;168(13):1396–1403.
7. **Norcini J, Lipner R, Kimball H.** Certifying examination performance and patient outcomes following acute myocardial infarction. Med Educ. 2002;36(9):853–859.
8. **Sharp L, Bashook P, Lipsky M, Horowitz S, Miller S.** Specialty board certification and clinical outcomes: the missing link. Acad Med. 2002;77:534–542.
9. **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.
10. **Webb TP, Weigelt JA, Redlich PN, Anderson RC, Brasel KJ, Simpson D.** Protected block curriculum enhances learning during general surgery residency training. Arch Surg. 2009;144(2):160–166.
11. **McDonald FS, Zeger SL, Kolars JC.** Factors associated with medical knowledge acquisition during internal medicine residency. J Gen Intern Med. 2007;22:962–968.
12. **McDonald FS, Zeger SL, Kolars JC.** Associations of conference attendance with Internal Medicine In-Training Examination Scores. Mayo Clin Proc. 2008;83(4):449–453.
13. Accreditation Council for Graduate Medical Education. Educational Innovations Project. Available at: http://www.acgme.org/acWebsite/RRC_140/140_EIPindex.asp. Accessed March 3, 2011.
14. **Warm EJ, Schauer DP, Diers T, et al.** The Ambulatory Long-Block: An Accreditation Council for Graduate Medical Education (ACGME) Educational Innovations Project (EIP). J Gen Intern Med. 2008;23(7):921–6.
15. **Warm EJ, Schauer DP, Revis B, Boex JR.** Multi-source feedback in the ambulatory setting. J Grad Med Educ. 2010;2(2):269–277.