Title

Assessing the Effects of the 2003 Resident Duty Hours Reform on Internal Medicine Board Scores

Permalink

https://escholarship.org/uc/item/01k3h2b9

Journal

Academic Medicine, 89(4)

ISSN

1040-2446

Authors

Silber, Jeffrey H
Romano, Patrick S
Itani, Kamal MF
et al.

Publication Date

2014-04-01

DOI

10.1097/acm.0000000000000193

Peer reviewed

Assessing the Effects of the 2003 Resident Duty Hours Reform on Internal Medicine Board Scores

Jeffrey H. Silber, MD, PhD, Patrick S. Romano, MD, MPH, Kamal M.F. Itani, MD, Amy K. Rosen, PhD, Dylan Small, PhD, Rebecca S. Lipner, PhD, Charles L. Bosk, PhD, Yanli Wang, MS, Michael J. Halenar, MPH, Sophia Korovaichuk, Orit Even-Shoshan, MS, and Kevin G. Volpp, MD, PhD

Abstract

Purpose

To determine whether the 2003 Accreditation Council for Graduate Medical Education (ACGME) duty hours reform affected medical knowledge as reflected by written board scores for internal medicine (IM) residents.

Method

The authors conducted a retrospective cohort analysis of postgraduate year 1 (PGY-1) Internal Medicine residents who started training before and after the 2003 duty hour reform using a merged data set of American Board of Internal Medicine (ABIM) Board examination and the National Board of Medical Examiners (NMBE) United States Medical Licensing Examination (USMLE) Step 2 Clinical Knowledge test scores. Specifically, using four regression models, the authors compared IM residents beginning PGY-1 training in 2000 and completing training unexposed to the 2003 duty hours reform (PGY-1 2000 cohort, n = 5,475) to PGY-1 cohorts starting in 2001 through 2005 (n = 28,008), all with some exposure to the reform.

Results

The mean ABIM board score for the unexposed PGY-1 2000 cohort (n = 5,475) was 491, SD = 85. Adjusting for demographics, program, and USMLE Step 2 exam score, the mean differences (95% CI) in ABIM board scores between the PGY-1 2001, 2002, 2003, 2004 and 2005 cohorts minus the PGY-1 2000 cohort were −5.43 (−7.63, −3.23), −3.44 (−5.65, −1.24), 2.58 (0.36, 4.79), 11.10 (8.88, 13.33) and 11.28 (8.98, 13.58) points respectively. None of these differences exceeded one-fifth of an SD in ABIM board scores.

Conclusions

The duty hours reforms of 2003 did not meaningfully affect medical knowledge as measured by scores on the ABIM board examinations.

In July 2003, the Accreditation Council for Graduate Medical Education (ACGME) introduced the 80-hour maximum work week for residents training in U.S. residency programs.[1–4] This duty hours rule represented a profound change in resident education, by reducing the work week, which before the reform had averaged between 90 and 120 hours for most specialties, and limiting the number of consecutive work hours to 24 hours with an additional 6 hours for education and transfer of care.[1,3] Although the regulations were motivated by a desire to reduce medical errors, they could also have a significant impact on resident quality of life and educational performance, as others have recently summarized.[5]

The duty hours change ushered in debate over its consequences for patient safety, since an increase in patient handovers[3,4,6–10] may offset the benefits of less fatigued residents,[11–13] and for the quality of educational experiences.[14–16] Previous studies have shown that patient outcomes two years after the change did not worsen and may have improved for medical patients in Veterans Affairs (VA) hospitals.[3,4,8–10,17–22] Prior research has generally examined the reform's effect on resident learning as measured by educational performance at single institutions.[5,23–28] In one exception, a national study reported no significant change in American Board of Neurosurgery exam scores taken for credit between 2002 through 2006, though concerns were raised about neurosurgery resident performance on other metrics such as dwindling presentation activity at national conferences.[29]

In addition to the sleep-safety literature that helped motivate duty hours reforms,[30–32] there is substantial research on the connection between sleep and learning.[33–37] This work has led to opposing hypotheses concerning the effects of duty hours reform on learning. On the one hand, some researchers would argue that alert residents may retain more information than sleep-deprived residents, and hence residents' specialty board examination scores may improve following the duty hours changes. On the other hand, others would argue that reduced exposure to clinical situations may result in a knowledge gap for residents following the change in duty hours, even if the residents were less sleep deprived.[38–42]

The duty hours reform of 2003 therefore presented a unique opportunity to study the impact on educational performance, as measured by board exam scores in a large population of residents. Some of these residents were fully exposed (because the reform occurred prior to starting postgraduate year 1 [PGY-1]) or partially exposed to duty hour regulations (because the reform was implemented in PGY-2 or PGY-3) while others were unexposed, having completed residency prior to 2003.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Dr. Silber, Center for Outcomes Research, The Children's Hospital of Philadelphia, 3535 Market St., Suite 1029, Philadelphia, PA 19104; telephone: (215) 590-5635; e-mail: silberj@wharton.upenn.edu.

Acad Med. 2014;89:644–651.
First published online February 19, 2014
doi: 10.1097/ACM.0000000000000193
Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A190.

With the support of the American Board of Internal Medicine (ABIM), the National Board of Medical Examiners (NBME), and the Educational Commission for Foreign Medical Graduates (ECFMG), we analyzed ABIM board exam scores for six cohorts of physicians. We used a linked dataset that included each candidate's United States Medical Licensing Examination (USMLE) Step 2 Clinical Knowledge score, which allowed us to adjust for pre-residency knowledge level and test-taking skills, thereby greatly improving the predictive ability of the reported model[43–48] and accounting for differences in specialty selectivity over time.

The USMLE Step 2 exam is taken by all U.S. medical students in their final year of medical school.[49] International medical students and graduates may take the USMLE Step exams, although students must have completed at least two years of medical school to be eligible.[50] The exam "is intended to assess whether an individual can apply medical knowledge, skills, and an understanding of clinical science essential for the provision of safe and effective patient care under supervision" and has been shown to be predictive of other features of performance.[49]

The ABIM board exam is a written exam "designed to evaluate the extent of the candidate's knowledge and clinical judgment in the areas in which an internist should demonstrate a high level of competence."[51] Little is available in the literature linking ABIM board scores to other performance variables; however, several studies have found positive and significant correlations between in-training written examination scores and board exam scores for some specialties.[48,52,53]

Method

Data sources

This research protocol was reviewed and approved by the Institutional Review Boards of The Children's Hospital of Philadelphia and the University of Pennsylvania. From the ABIM, we obtained a data set of internal medicine residents who took their first specialty board examination in Internal Medicine between 2003 and 2008 (representing a cohort of residents who were generally interns in 2000 through 2005). These data included demographic characteristics of the test-taker and an identifier for his or her residency program. The NBME appended Step 2 scores taken near the end of medical school. All personal identifiers were removed before we received the data. We excluded residents who postponed their graduation beyond three years due to research programs or other leaves of absence. We also obtained permission from the Educational Commission for Foreign Medical Graduates (ECFMG) to use the information on international medical graduates who took the Step 2 exam. A total of 41,679 residents took the ABIM board exam for the first time between 2003 and 2008. Of these physicians, we excluded 4,286 because they were not in 3-year programs and 3,910 due to missing NBME data (i.e. 3,342 missing all NBME information including Step 2 scores and demographics, 322 missing Step 2 score, and 246 missing demographics), leaving 33,483 residents with data available for analysis. The difference in average ABIM board score for the excluded residents (missing Step 2 score or unlinked to ABIM board exam) versus included residents was only 6.8 points, or an effect size of 0.04 SDs.

USMLE Step 2 exam and ABIM board exam scores

Beginning in 1999 (for the residency cohort starting in 2000), the Step 2 exam was administered using an electronic format rather than paper and pencil.[54] The exam goes through extensive analysis and scores are "equated" such that one year's test score can generally be compared to the next. Equating is accomplished using responses on questions that did not change year-to-year.[55] Linear equating is a statistical method that adjusts for small differences in difficulty among forms of a test that are built to be as similar in difficulty and content as possible.[56] Common-item equating uses questions that are the same on the different forms of the test to develop a formula for adjusting the overall score on one form of the test to be comparable to the overall score on another form of the test, where the overall score for a form includes both the common items on the different forms of the test and the questions that are unique to that form of the test. The Tucker linear method was used to do the equating.[55] For a setting with two populations each taking a different form of a test, the Tucker method assumes that the regression of overall score on the set of questions that are common to the different forms has the same coefficients and same variance for the two populations. The mean Step 2 score given in 1999 for the PGY-1 cohort that began in 2000 was 210.7, with a standard deviation of 22.6.

There was a change in format for the ABIM board exam in 2006, when the exam began to be administered using an electronic format. The new exam process was shortened to one day from two days, resulting in fewer questions asked per test, and a new standard and equating chain was established in 2006. The new electronic exam was still similar in content and difficulty to the paper and pencil test administered in previous years. As in the case of the Step 2 examinations, the ABIM also reports equated scores to allow for meaningful year-to-year comparisons. Even though in practice a new equating chain began with the new electronic format in 2006, scores on the ABIM board examination were made statistically comparable to prior examination years by using the Tucker linear equating process.[55] The mean ABIM board score in 2003 (taken by the PGY-1 cohort that started in 2000) was 490.9, with a standard deviation of 85.4.

Statistical methods

We used ordinary least squares (OLS) regression to adjust for specific characteristics of the residents, their USMLE Step 2 scores prior to starting their PGY-1 year, and the residency program in which they trained. We chose OLS because the dependent variable, ABIM board score, is continuous and residuals were approximately normally distributed. Because residents were clustered by residency program, and residency program may affect board scores, our preferred models included a fixed-effect term for each specific residency program. We conducted statistical analyses using SAS software, version 9.2 (SAS Institute Inc., Cary, North Carolina).

The model we used to fit ABIM board scores was based on a development process using data from 1997 through 1999, that is, data prior to the baseline year 2000 used for this study. The reference group was the PGY-1 2000 cohort, which graduated before the change in the duty hours rules. We chose 2000 as the reference year for the unexposed cohort because the PGY-1 2000 cohort was the first to take the electronic version of the Step 2 exam, limiting the differences in format between the unexposed and exposed cohorts. As further reassurance, we conducted an additional analysis using 1999 as the reference year and this yielded very similar results (see Supplemental Digital Tables 1–3 at http://links.lww.com/ACADMED/A190).

To estimate the effect of duty hours reform, we examined a baseline data set of variables considered educationally relevant for predicting board scores: residency program affiliation, Step 2 score, training type, English as native language ("language"), citizen of the United States or Canada ("citizen"), graduated from medical school in the United States or Canada ("school"), age, sex, race/ethnicity (black, Asian Pacific, Hispanic, other, and white as reference), language × citizen, language × school, citizen × school. Training type refers to the residency program designation used to distinguish between program tracks. Training type consists of "internal medicine" and "categorical" as one group (90.5% of residents), "primary care" (6.7%), pediatrics (0.2%), and various smaller other training types capturing residents who completed stacked dual programs in sequence or completed internal medicine preliminary to some other specialty, such as neurology (2.6%).

To these pre-specified variables, we added all pair-wise interaction terms that were significant after correcting for multiple comparisons using the Bonferroni procedure (i.e. an upper bound for type I error of 0.05 was divided by 25 interactions to give us a required $P < .0002$ for each comparison to reach significance). The resulting model included the following interaction terms: training type × age, race × Step 2 score, race × language, race × citizen, race × school, race × age, training type × race, Step 2 score × language, Step 2 score × citizen, Step 2 score × school, Step 2 score × age, language × age, citizen × age, school × sex. These interactions identified test-takers with notable combinations of characteristics, such as U.S. and Canadian students who chose to attend schools elsewhere, and students from English-speaking countries outside the United States and Canada.

Stability analyses

Since international medical graduates (IMGs) comprise almost half of Internal Medicine residents, we first asked if our results were stable across IMG (n = 15,156) and non-IMG groups (n = 18,327). In a second stability analysis we examined a more homogeneous definition of residents (a strict definition), retaining Internal Medicine candidates only (n = 30,319) who did not switch into their program from another track or specialty.

Results

The population of residents used in this study taking the ABIM board examination for their first time is displayed in Table 1. These residents came from 418 residency programs with a mean of 13.6 trainees per program taking the ABIM exam each year.

Table 2 shows the unadjusted scores for both the USMLE Step 2 and the ABIM board examinations. We have arranged the scores by the cohort associated with both exams. For example, the PGY-1 2000 cohort of residents would have taken their Step 2 exam in 1999 and would have been on track to take their ABIM board exam in 2003. They would have completed all residency training prior to the duty hours reform that started in July 2003. Note that there was an increase in Step 2 scores over the study time period, creating about a 10-point difference (one-half an SD) in scores between the PGY-1 2000 and 2005 cohorts. There also was an increase of about one-third of an SD increase in board exam scores over this study period.

To better account for variation at the individual level, we developed models to adjust for specific characteristics of the residents, their Step 2 scores, and their residency programs. In Table 3 we display four models, of increasing complexity, for predicting the difference in ABIM board scores between PGY-1 2000 minus PGY-1 in 2001, 2002, 2003, 2004, and 2005.

- Model 1 includes only resident demographic information;
- Model 2 includes demographic information and residency program affiliation;
- Model 3 includes demographic information and Step 2 score; and
- Model 4 includes demographic information, Step 2 score, and residency program affiliation.

Adding the Step 2 scores greatly improved prediction of board exam scores, although

Table 1

Description of Residents Taking the American Board of Internal Medicine Board Examination for the First Time, 2003–2008, and Their Residency Programs

Characteristic	Sample
Residents (N = 33,348)	
Mean age (SD)	32.3 (4.1)
Female, no. (%)	13,976 (41.7)
Race and ethnicity,[a] no. (%)	
White	13,961 (41.7)
Asian/Pacific Islander	11,818 (35.3)
Other	2,683 (8.0)
Hispanic	2,017 (6.0)
Black	1,882 (5.6)
Missing	1,122 (3.4)
International medical graduate, no. (%)	15,156 (45.3)
English not native language, no. (%)	14,323 (42.8)
Residency programs (N = 418)	
Number of PGY-1 medical residents per program	
Mean (SD)	13.6 (10.4)
Median	10.2
5th, 25th, 75th, 95th percentiles	3.3, 6.3, 17.5, 38.7

[a] Race is self-identified. "Other" includes Native American, multiple, and other.

Table 2

Unadjusted United States Medical Licensing Examination (USMLE) Step 2 Clinical Knowledge and American Board of Internal Medicinal (ABIM) Board Score Distributions by Year

	Resident PGY-1 cohort					
Exam	2000[a] (n = 5,475)	2001 (n = 5,672)	2002 (n = 5731)	2003 (n = 5,726)	2004 (n = 5,716)	2005 (n = 5,163)
USMLE Step 2						
Year of exam	1999	2000	2001	2002	2003	2004
Score, mean (SD)	210.7 (22.6)	213.0 (23.0)	214.9 (22.4)	216.2 (22.0)	218.7 (21.2)	220.8 (21.4)
Score, median (IQR)	209 (193–228)	213 (195–230)	215 (197–231)	216 (199–232)	218 (202–234)	221 (204–237)
ABIM board						
Year of exam	2003	2004	2005	2006	2007	2008
Score, mean (SD)	490.9 (85.4)	489.6 (83.8)	494.7 (80.0)	503.9 (74.2)	517.3 (69.4)	522.2 (70.3)
Score, median (IQR)	502 (439–552)	498 (435–550)	502 (443–553)	511 (457–557)	524 (474–566.5)	528 (478–574)

[a] This group represents the "unexposed" cohort, postgraduate year 1 (PGY-1) residents who began their internship in 2000, and were on track to take the ABIM Board exam in 2003, thus completing all residency training prior to the duty hours reform that began in July 2003.

residency program affiliation was also helpful. The R^2 statistic, representing the proportion of total variation explained by the model, rose from 13.1% using just demographic variables (Model 1) to 20.8% when program affiliation was included (Model 2), to 39.8% when Step 2 score was included but not program affiliation (Model 3), and to 43.9% in the full model (Model 4).

In Table 3 we also show all four adjustment models comparing each PGY-1 cohort that was exposed to the duty hours reform to the unexposed reference PGY-1 2000 cohort (n = 5,475) who started PGY-1 in the year 2000 and completed residency prior to the July 2003 duty hours reforms. For example, the variable "PGY-1 2001 vs. 2000" provides the mean difference in ABIM board scores between partially-exposed residents who started residency in 2001 (exposed to the duty hours regulations for only one year), and the unexposed reference cohort. The value −5.43 in Model 4 suggests that the mean ABIM board score of the partially exposed PGY-1 2001 cohort was 5.43 (95% confidence interval, −7.63 to −3.2) points less than that of the unexposed PGY-1 2000 cohort, adjusting for demographics, residency programs, and Step 2 scores. The variable "PGY-1 2002 versus 2000" compares those who experienced duty hours reform during both their last two residency years to the unexposed cohort, with similar results. We also compare variables PGY-1 2003, PGY-1 2004, and PGY-1 2005—the three fully exposed cohorts that experienced all three years of residency under the new duty hours regime—to the unexposed reference cohort.

According to the most complete adjustment model (Model 4), partially exposed residency cohorts (PGY-1 2001 and 2002) performed significantly worse on the ABIM board exam than did the unexposed cohort (PGY-1 2000), although this effect was only about 5.5 points relative to a mean of 490.9 and less than a tenth of an SD in effect size (5.5/85.4) (Table 2),[57] but with a highly statistically significant *P*-value due to the very large sample size. The first fully-exposed cohort (PGY-1 2003) displayed about a 2.6-point increase in mean board exam scores ($P < .04$), and the PGY-1 2004 and 2005 cohorts also showed significant improvements in their board scores of about 11 points ($P < .0001$). This is a small change relative to the mean of the ABIM board exam and represents only slightly more than one-tenth of an SD.

Stability analyses

Running the full models with resident characteristics on subsets of the population—IMGs (n = 15,156) and strictly-defined Internal Medicine program residents (n = 30,319)—yielded results that were very similar to the main results from the total resident population results shown in Table 3. For each of the stability analyses, we used the most complete model (Model 4), which includes resident characteristics, USMLE Step 2 scores, and residency program affiliations. Across all models and comparison years shown in Table 4, the effect difference between models in any year did not exceed 12.6 points, or 0.15 SDs in the ABIM board score—a clinically insignificant change. Inside each subset model, the difference in scores between the PGY-1 2000 baseline cohort and any other cohort was at most 14.9 points, or less than 0.17 SDs in the board score. Again, this change is not clinically significant.

Discussion

The ACGME took a bold step in reducing the working hours of residents in 2003. Many believed such a large change in resident duty hours would adversely affect patient outcomes and resident learning.[6,7,38–42] Others argued that such a change would improve outcomes and improve a resident's ability to learn, as they would be less sleep deprived.[11–13] It is only in retrospect that we can assess the effects of this natural experiment

Table 3

Difference in Adjusted American Board of Internal Medicine (ABIM) Board Exam Scores of Postgraduate Year 1 (PGY-1) Residents From 2001–2005 (N = 33,483), Compared to the Unexposed Cohort (n = 5,475)[a]

	Adjusted Score Differences						
Model[b]	2001 PGY-1 vs. 2000	2002 PGY-1 vs. 2000	2003 PGY-1 vs. 2000	2004 PGY-1 vs. 2000	2005 PGY-1 vs. 2000	F-test (*df*)	R^2
Model 1 (demographics)						93.28 (54)	13.1%
Mean (95% CI)	−0.47 (−3.19, 2.25)	4.78** (2.06, 7.50)	14.26*** (11.54, 16.98)	27.15*** (24.43, 29.88)	31.43*** (28.64, 34.23)		
Effect size in SD	−0.06	0.06	0.17	0.32	0.37		
Model 2 (demographics + program affiliation)						18.38 (471)	20.8%
Mean (95% CI)	−0.49 (−3.10, 2.12)	4.67** (2.06, 7.27)	12.93*** (10.32, 15.55)	26.10*** (23.48, 28.72)	30.10*** (27.40, 32.79)		
Effect size in SD	−0.01	0.05	0.15	0.31	0.35		
Model 3 (demographics + Step 2 score)						350.59 (63)	39.8%
Mean (95% CI)	−5.17*** (−7.44, −5.17)	−3.17** (−5.44, −0.90)	3.35** (1.07, 5.63)	11.71*** (9.42, 14.00)	12.20*** (9.84, 14.56)		
Effect size in SD	−0.06	−0.04	0.04	0.14	0.14		
Model 4 (demographics + Step 2 score + program affiliation)						53.81 (480)	43.9%
Mean (95% CI)	−5.43*** (−7.63, −3.23)	−3.44** (−5.65, −1.24)	2.58* (0.36, 4.79)	11.10*** (8.88, 13.33)	11.28*** (8.98, 13.58)		
Effect size in SD	−0.06	−0.04	0.03	0.13	0.13		

[a] The "unexposed" cohort and reference year is PGY-1 residents who began their internship in 2000 and were on track to take the ABIM Board exam in 2003, thus completing all residency training prior to the duty hours reform that began in July 2003.

[b] See the text for demographic information and interaction terms. Step 2 score indicates United States Medical Licensing Examination Step 2 Clinical Knowledge score. The effect size can be calculated by dividing the parameter estimate by a score of 85.4, the SD of the ABIM board examination for the reference year 2000.

*P < .05;

**P < .01;

***P < .0001; no designation implies not significant at the P > .05 level.

on medical knowledge, as measured by ABIM board scores. One of our study's strengths is the linking of USMLE Step 2 scores with ABIM board scores to analyze those effects.

We found that any effect from the duty hours regulations on medical knowledge was small and inconsistent over time. In the first two years following the reform, board scores dropped slightly (i.e., 3.4–5.4 points or less than one-tenth of a standard deviation), reflecting minimal change in learning.[57] In the third year after reform, for the first cohort of residents who were fully exposed to the new rules for all three years of residency (PGY-1 2003), we found no difference in board scores as compared with the unexposed PGY-1 2000 cohort. Finally, in the two most recent cohorts of fully exposed residents, we found small but statistically significant improvements in board scores, with effect sizes of about 0.13 SDs (e.g. for a PGY-1 2005 cohort board score difference of 11.3 points, using the PGY-1 2000 cohort board score SD of 85.4, we get an effect size of 0.13 SDs). Small effect size and the lack of a consistent effect across the three fully exposed cohorts suggests that the observed association is unlikely to be causal.

Although we made efforts to equate test difficulty over time, the written ABIM board examination changed from a paper and pencil to an electronic format starting with the PGY-1 2003 cohort, and the number of questions was reduced. These changes may confound our results, with respect to the PGY-1 2003-2005 cohorts, but it appears that if there was any effect of the ACGME regulations on ABIM board score performance, it was a small one. Three alternative explanations for this finding include the following: (1) the 2003 regulations may have been too limited to reduce residents' sleep deprivation, thereby limiting their effect on learning; (2) the 2003 regulations may have had opposing effects that roughly canceled each other, as suggested by our alternative hypotheses—less sleep deprived residents learned better, but reduced exposure to clinical situations created a knowledge gap; or (3) the frequency and severity of the sleep deprivation experienced by Internal Medicine residents, before or after 2003, may not have been sufficient to affect learning over the course of a three year training program. The ACGME adopted stricter duty hours limits in 2011, which

Table 4

Stability Analysis Comparing American Board of Internal Medicine Board Scores of International Medical Graduates (IMGs), non-IMGs, and Strictly Defined Internal Medicine (IM) Residents to Scores of the Total Resident Population, Postgraduate Year 1 (PGY-1) 2001–2005 Cohorts vs. PGY-1 2000 Cohort[a]

	Adjusted Score Differences[b]						
Population	2001 PGY-1 vs. 2000	2002 PGY-1 vs. 2000	2003 PGY-1 vs. 2000	2004 PGY-1 vs. 2000	2005 PGY-1 vs. 2000	F-test (*df*)	R^2
IMG (n = 15,156)						19.89 (445)	37.6%
Mean (95% CI)	−11.90***	−10.76***	0.73	13.34***	14.89***		
	(−15.60, −8.20)	(−14.46, −7.06)	(−2.96, 4.42)	(9.61, 17.07)	(11.06, 18.73)		
Effect size as SD	−0.14	−0.13	0.01	0.16	0.17		
Non-IMG (n = 18,327)						48.28 (404)	52.1%
Mean (95% CI)	−0.90	1.87	4.11**	8.95***	7.80***		
	(−3.53, 1.72)	(−0.77, 4.50)	(1.43, 6.80)	(6.25, 11.64)	(5.01, 10.59)		
Effect size as SD	−0.01	0.02	0.05	0.10	0.09		
Strict IM[c] IMG and non-IMG (n = 30,319)						40.91 (457)	44.1%
Mean (95% CI)	−5.39***	−3.26*	1.98	10.28***	10.16***		
	(−7.70, −3.07)	(−5.58, −0.94)	(−0.34, 4.31)	(7.94, 12.61)	(7.76, 12.57)		
Effect size as SD	−0.06	−0.04	0.02	0.12	0.12		
Total resident population (N = 33,483)						53.81 (480)	43.9%
Mean (95% CI)	−5.43***	−3.44**	2.58*	11.10***	11.28***		
	(−7.63, −3.23)	(−5.65, −1.24)	(0.36, 4.79)	(8.88, 13.33)	(8.98, 13.58)		
Effect size as SD	−0.06	−0.04	0.03	0.13	0.13		

[a] Cohort year (e.g., "PGY-1 2000") indicates the year residents began their residencies.

[b] Each model adjusts for resident demographics, United States Medical Licensing Examination Step 2 Clinical Knowledge score, and residency program affiliation, which are all significant at the $P < .0001$ level. PGY-1 2001-2005 is being compared to PGY-1 2000 because the latter is the "unexposed" cohort—residents who began their internship in 2000 and were on track to take the ABIM Board exam in 2003, thus completing all residency training prior to duty hours reform that began in July 2003.

[c] "Strict IM" includes only IM candidates who did not switch into their program from another track or specialty.

**P* < .05;

***P* < .01;

****P* < .0001; no designation implies not significant at the *P* > .05 level.

may provide a future opportunity to test the first explanation.[58]

It was important to account for Step 2 scores in our analyses. If we had not done so, we would have mistakenly concluded that there was about a 30-point improvement in mean ABIM board scores in the fully exposed cohorts that was potentially attributable to the duty hours reform. However, when we adjusted for the Step 2 scores, the improvement decreased to about 11 points, suggesting that observed improvements in ABIM board scores were driven primarily by candidates who were better at taking the Step 2 exam. This finding may reflect increasing selectivity of Internal Medicine programs during this study period. Adjustment for previous test-taking ability and other resident factors is a crucial step in any future research studying the influence of changes in working hours, such as those implemented by the 2011 iteration of the duty hours rules, on board exam performance.

Our results were also stable over various subpopulations for which we had data. The 2003 ACGME duty hours reform appears to have had similar effects on the board scores of IMG and non-IMG physicians, as well as those residents who started and completed traditional three year programs.

Although many studies have assessed the effects of sleep deprivation on learning,[33–37] our study does not directly address this relationship, as we do not know how residents changed their study habits after the 2003 duty hours reform. Specifically, we do not know whether residents increased their studying time outside the hospital or improved their learning efficiency given similar study time. We report the net change in ABIM board scores, which is relevant to the policy question at hand—the training and acquired knowledge of medicine residents—but is not as informative for those interested in the pure effect of sleep deprivation on learning. Furthermore, it remains unclear how relevant any exam is at predicting clinical expertise. Clinical experiences involve synthesizing data in context. Typically, board exams ask for nuggets of knowledge that may be decontextualized. Hence, our study is bound by the limitations inherent to any written examination designed to determine how likely a physician is to provide optimal care for future patients.

In conclusion, it appears that the duty hours reform of 2003 had a very small, clinically insignificant but positive effect on ABIM board scores. Our findings suggest that neither the widespread concerns that duty hours restrictions would worsen educational performance, nor the hopes of great improvement in retention of information, have been realized.

Acknowledgments: We thank the National Board of Medical Examiners, Philadelphia, PA and the Educational Commission for Foreign Medical Graduates and the Foundation for Advancement of International Medical Education and Research, Philadelphia, PA and the American Board of Internal Medicine, Philadelphia, PA for their assistance with this research. We also thank Traci Frank from the Center for Outcomes Research for her contribution concerning this research.

Funding/Support: This research was supported by a grant from the National Heart, Lung, and Blood Institute, Grant # R01 HL082637.

Other disclosures: None reported.

Ethical approval: This research protocol was reviewed and approved by the institutional review boards of The Children's Hospital of Philadelphia and the University of Pennsylvania.

Previous presentations: Portions of this research were presented at the Academy Health Annual Meeting, June 2012, Orlando, Florida.

Dr. Silber is professor, Departments of Pediatrics and Anesthesiology & Critical Care, Perelman School of Medicine; professor, Department of Health Care Management, The Wharton School; director, Center for Outcomes Research, The Children's Hospital of Philadelphia; and senior fellow, Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, Pennsylvania.

Dr. Romano is professor of medicine and pediatrics and director, Primary Care Outcomes Research Faculty Development Program, Division of General Medicine and Center for Healthcare Policy and Research, University of California Davis School of Medicine, Sacramento, California.

Dr. Itani is professor, Department of Surgery, Boston University School of Medicine, and chief of surgery, VA Boston Health Care System and Boston University, Boston, Massachusetts.

Dr. Rosen is professor, Department of Health Policy and Management, Boston University School of Public Health, affiliated with the Center for Organization, Leadership and Management Research, VA Boston Healthcare System, Boston, Massachusetts.

Dr. Small is associate professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania.

Dr. Lipner is senior vice president of evaluation, research and development, American Board of Internal Medicine, Philadelphia, Pennsylvania.

Dr. Bosk is professor, Departments of Sociology and Medical Ethics & Health Policy, and senior fellow, Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, Pennsylvania.

Ms. Wang is a statistical programmer, Center for Outcomes Research, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania.

Mr. Halenar is a research assistant, Center for Health Equity Research and Promotion, Veteran's Administration Hospital, Philadelphia, Pennsylvania.

Ms. Korovaichuk is a research assistant, Center for Outcomes Research, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania.

Ms. Even-Shoshan is associate director, Center for Outcomes Research, The Children's Hospital of Philadelphia, and senior fellow, Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, Pennsylvania.

Dr. Volpp is professor, Department of Health Care Management, The Wharton School, and Department of Medicine, Perelman School of Medicine; and senior fellow, Leonard Davis Institute of Health Economics, University of Pennsylvania; and affiliated with the Center for Health Equity Research and Promotion, Veterans Administration Hospital, Philadelphia, Pennsylvania.

References

1 Accreditation Council for Graduate Medical Education. Resident Duty Hours in the Learning and Working Environment. http://www.acgme.org/acgmeweb/Portals/0/PDFs/dh-ComparisonTable2003v2011.pdf. Accessed December 17, 2013.

2 Accreditation Council for Graduate Medical Education. Report of the Work Group on Resident Duty Hours and the Learning Environment. June 11, 2002. In: The ACGME's Approach to Limit Resident Duty Hours 12 Months After Implementation: A Summary of Achievements, 2004. http://www.acgme.org/acgmeweb/Portals/0/PFAssets/PublicationsPapers/dh_dutyhoursummary2003-04.pdf. Accessed December 6, 2013.

3 Volpp KG, Rosen AK, Rosenbaum PR, et al. Mortality among hospitalized Medicare beneficiaries in the first 2 years following ACGME resident duty hour reform. JAMA. 2007;298:975–983.

4 Volpp KG, Rosen AK, Rosenbaum PR, et al. Mortality among patients in VA hospitals in the first 2 years following ACGME resident duty hour reform. JAMA. 2007;298:984–992.

5 Fletcher KE, Reed DA, Arora VM. Patient safety, resident education and resident well-being following implementation of the 2003 ACGME duty hour rules. J Gen Intern Med. 2011;26:907–919.

6 Drazen JM. Awake and informed. N Engl J Med. 2004;351:1884.

7 Mukherjee S. A precarious exchange. N Engl J Med. 2004;351:1822–1824.

8 Silber JH, Rosenbaum PR, Even-Shoshan O, et al. Length of stay, conditional length of stay, and prolonged stay in pediatric asthma. Health Serv Res. 2003;38:867–886.

9 Rosen AK, Loveland SA, Romano PS, et al. Effects of resident duty hour reform on surgical and procedural patient safety indicators among hospitalized Veterans Health Administration and Medicare patients. Med Care. 2009;47:723–731.

10 Silber JH, Romano PS, Rosen AK, et al. Failure-to-rescue: Comparing definitions to measure quality of care. Med Care. 2007;45:918–925.

11 Gaba DM, Howard SK. Patient safety: Fatigue among clinicians and the safety of patients. N Engl J Med. 2002;347:1249–1255.

12 Landrigan CP, Rothschild JM, Cronin JW, et al. Effect of reducing interns' work hours on serious medical errors in intensive care units. N Engl J Med. 2004;351:1838–1848.

13 Lockley SW, Cronin JW, Evans EE, et al. Effect of reducing interns' weekly work hours on sleep and attentional failures. N Engl J Med. 2004;351:1829–1837.

14 Myers JS, Bellini LM, Morris JB, et al. Internal medicine and general surgery residents' attitudes about the ACGME duty hours regulations: A multicenter study. Acad Med. 2006;81:1052–1058.

15 Jagsi R, Shapiro J, Weissman JS, Dorer DJ, Weinstein DF. The educational impact of ACGME limits on resident and fellow duty hours: A pre–post survey study. Acad Med. 2006;81:1059–1068.

16 Horwitz LI, Krumholz HM, Huot SJ, Green ML. Internal medicine residents' clinical and didactic experiences after work hour regulation: A survey of chief residents. J Gen Intern Med. 2006;21:961–965.

17 Shetty KD, Bhattacharya J. Changes in hospital mortality associated with residency work-hour regulations. Ann Intern Med. 2007;147:73–80.

18 Horwitz LI, Kosiborod M, Lin Z, Krumholz HM. Changes in outcomes for internal medicine inpatients after work-hour regulations. Ann Intern Med. 2007;147:97–103.

19 Prasad M, Iwashyna TJ, Christie JD, et al. Effect of work-hours regulations on intensive care unit mortality in United States teaching hospitals. Crit Care Med. 2009;37:2564–2569.

20 Silber JH, Rosenbaum PR, Rosen AK, et al. Prolonged hospital stay and the resident duty hour rules of 2003. Med Care. 2009;47:1191–1200.

21 Volpp KG, Rosen AK, Rosenbaum PR, et al. Did duty hour reform lead to better outcomes among the highest risk patients? J Gen Intern Med. 2009;24:1149–1155.

22 Press MJ, Silber JH, Rosen AK, et al. The impact of resident duty hour reform on hospital readmission rates among Medicare beneficiaries. J Gen Intern Med. 2011;26:405–411.

23 Hutter MM, Kellogg KC, Ferguson CM, Abbott WM, Warshaw AL. The impact of

the 80-hour resident workweek on surgical residents and attending surgeons. Ann Surg. 2006;243:864–871.

24 Froelich J, Milbrandt JC, Allan DG. Impact of the 80-hour workweek on surgical exposure and national in-training examination scores in an orthopedic residency program. J Surg Educ. 2009;66:85–88.

25 Durkin ET, McDonald R, Munoz A, Mahvi D. The impact of work hour restrictions on surgical resident education. J Surg Educ. 2008;65:54–60.

26 Schneider JR, Coyle JJ, Ryan ER, Bell RH Jr, DaRosa DA. Implementation and evaluation of a new surgical residency model. J Am Coll Surg. 2007;205:393–404.

27 de Virgilio C, Yaghoubian A, Lewis RJ, Stabile BE, Putnam BA. The 80-hour resident workweek does not adversely affect patient outcomes or resident education. Curr Surg. 2006;63:435–439.

28 Barden CB, Specht MC, McCarter MD, Daly JM, Fahey TJ 3rd. Effects of limited work hours on surgical training. J Am Coll Surg. 2002;195:531–538.

29 Jagannathan J, Vates GE, Pouratian N, et al. Impact of the Accreditation Council for Graduate Medical Education work-hour regulations on neurosurgical resident education and productivity. J Neurosurg. 2009;110:820–827.

30 Dawson D, Reid K. Fatigue, alcohol and performance impairment. Nature. 1997;388:235.

31 Van Dongen HP, Maislin G, Mullington JM, Dinges DF. The cumulative cost of additional wakefulness: Dose-response effects on neurobehavioral functions and sleep physiology from chronic sleep restriction and total sleep deprivation. Sleep. 2003;26: 117–126.

32 Belenky G, Wesensten NJ, Thorne DR, et al. Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: A sleep dose-response study. J Sleep Res. 2003;12:1–12.

33 Buysse DJ, Barzansky B, Dinges D, et al. Sleep, fatigue, and medical training: Setting an agenda for optimal learning and patient care. Sleep. 2003;26:218–225.

34 Dinges DF, Pack F, Williams K, et al. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 hours per night. Sleep. 1997;20: 267–277.

35 Doran SM, Van Dongen HP, Dinges DF. Sustained attention performance during sleep deprivation: Evidence of state instability. Arch Ital Biol. 2001;139:253–267.

36 Polzella DJ. Effects of sleep deprivation on short-term recognition memory. J Exp Psychol Hum Learn. 1975;104:194–200.

37 Linde L, Bergström M. The effect of one night without sleep on problem-solving and immediate recall. Psychol Res. 1992;54: 127–136.

38 Weinstein DF. Duty hours for resident physicians—tough choices for teaching hospitals. N Engl J Med. 2002;347:1275–1278.

39 Pennell NA, Liu JF, Mazzini MJ. Interns' work hours. N Engl J Med. 2005;352:726–728.

40 Charap M. Reducing resident work hours: Unproven assumptions and unforeseen outcomes. Ann Intern Med. 2004;140: 814–815.

41 Nash GF, Reddy KM, Bloom IT. A regional survey of emergency surgery: The trainees' perspective. Ann R Coll Surg Engl. 2000;82:95–96.

42 Dickson L, Heymann TD, Culling W. What the SHO saw. J R Coll Physicians Lond. 1994;28:523–526.

43 McCaskill QE, Kirk JJ, Barata DM, et al. USMLE step 1 scores as a significant predictor of future board passage in pediatrics. Ambul Pediatr. 2007;7:192–195.

44 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 Am. 2004;86-A:1092–1095.

45 Sosenko J, Stekel KW, Soto R, Gelbard M. NBME Examination Part I as a predictor of clinical and ABIM certifying examination performances. J Gen Intern Med. 1993;8: 86–88.

46 McDonald FS, Zeger SL, Kolars JC. Associations between United States Medical Licensing Examination (USMLE) and Internal Medicine In-Training Examination (IM-ITE) scores. J Gen Intern Med. 2008;23:1016–1019.

47 Perez JA Jr, Greer S. Correlation of United States Medical Licensing Examination and Internal Medicine In-Training Examination performance. Adv Health Sci Educ Theory Pract. 2009;14:753–758.

48 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.

49 Cuddy MM, Dillon GF, Clauser BE, et al. Assessing the validity of the USMLE step 2 clinical knowledge examination through an evaluation of its clinical relevance. Acad Med. 2004;79:S43–S45.

50 Educational Commission for Foreign Medical Graduates. ECFMG 2011 Information Booklet. 2011. http://www.ecfmg.org/2012ib/2012ib.pdf. Accessed December 6, 2013.

51 American Board of Internal Medicine. Internal Medicine Certification Examination Blueprint. http://www.abim.org/pdf/blueprint/im_cert.pdf. Accessed December 6, 2013.

52 Waxman H, Braunstein G, Dantzker 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;9:692–694.

53 Frederick RC, Hafner JW, Schaefer TJ, Aldag JC. Outcome measures for emergency medicine residency graduates: Do measures of academic and clinical performance during residency training correlate with American Board of Emergency Medicine test performance? Acad Emerg Med. 2011;18(suppl 2):S59–S64.

54 Kies S, Shultz M. Proposed changes to the United States Medical Licensing Examination: Impact on curricula and libraries. J Med Libr Assoc. 2010;98:12–16.

55 Kolen MJ, Brennan RL. Chapter 4: Nonequivalent groups—linear methods; 4.1 Tucker method. In: Test Equating, Scaling and Linking: Methods and Practices. New York, NY: Springer-Verlag; 2004:103–108.

56 Kolen MJ, Brennan RL. Chapter 1: Introduction and concepts; 1.1: Equating and related concepts. In: Test Equating, Scaling and Linking: Methods and Practices. New York, NY: Springer-Verlag; 2004:2.

57 Hedges LV, Olkin I. Chapter 5: Estimation of a single effect size: Parametric and nonparametric methods. In: Statistical Methods for Meta-Analysis. Orlando, Fla: Academic Press; 1985:78.

58 Accreditation Council for Graduate Medical Education. Common Program Requirements. July 2011. http://web.archive.org/web/20120120050353/http://www.acgme.org/acWebsite/home/Common_Program_Requirements_07012011.pdf. Accessed December 6, 2013.