

Supplemental Digital Table 1
Unadjusted Test Score Distributions by Year with the Addition of the PGY-1 1999 Cohort
(N=38,802 for all cohorts)*

Step 2 CK Scores	Year of Step 2 CK Exam	1998	1999	2000	2001	2002	2003	2004
	N	5303	5491	5672	5731	5726	5716	5163
	Mean (SD)	210.0 (21.9)	210.7 (22.6)	213.0 (23.0)	214.9 (22.4)	216.2 (22.0)	218.7 (21.2)	220.8 (21.4)
	Median (IQR)	210 (193-226)	209 (193-228)	213 (195-230)	215 (197- 231)	216 (199-232)	218 (202-234)	221 (204-237)
ABIM Board Scores	Year Starting PGY-1	1999	2000	2001	2002	2003	2004	2005
	Year of ABIM Board Exam	2002	2003	2004	2005	2006	2007	2008
	Mean (SD)	491.1 (87.6)	490.9 (85.4)	489.6 (83.8)	494.7 (80.0)	503.9 (74.2)	517.3 (69.4)	522.2 (70.3)
	Median (IQR)	499 (437-554)	502 (439-552)	498 (435-550)	502 (443-553)	511 (457-557)	524 (474-566.5)	528 (478-574)

*Each column represents a cohort of residents who took their USMLE Step 2 Clinical Knowledge (CK) exam, started postgraduate year 1 (PGY-1) in the following year, and completed the ABIM specialty board exam three years later.

Supplemental Digital Table 2

Difference in Adjusted ABIM Board Exam Scores of Resident Cohorts Partially or Fully Exposed to the 2003 Duty Hours Reform (PGY-1 2001, 2002, 2003, and 2004 Cohorts) Compared to the Unexposed PGY-1 2000 Reference Cohort with the Addition of the PGY-1 1999 Cohort (N=38,802)

Model		Adjusted Score Differences						F test (df)	R ²
		PGY-1 1999 vs. 2000	PGY-1 2001 vs. 2000	PGY-1 2002 vs. 2000	PGY-1 2003 vs. 2000	PGY-1 2004 vs. 2000	PGY-1 2005 vs. 2000		
Model 1 <i>Demographics</i>	Mean (95% CI) Effect Size (SD)	0.50 (-2.31, 3.31) 0.01	-0.58 (-3.34, 2.18) -0.01	4.68** (1.92, 7.44) 0.05	14.12*** (11.36, 16.88) 0.17	27.03*** (24.26, 29.79) 0.32	31.30*** (28.46, 34.14) 0.37	108.90 (df=55)	13.4%
Model 2 <i>Demographics+ Program Affiliation</i>	Mean (95% CI) Effect Size (SD)	1.27 (-1.43, 3.96) 0.01	-0.64 (-3.29, 2.00) -0.01	4.51 (1.86, 7.15) 0.05	12.75** (10.10, 15.41) 0.15	25.88*** (23.23, 28.53) 0.30	29.93*** (27.20, 32.66) 0.35	21.56 (df=472)	21.0%
Model 3 <i>Demographics+ Step 2 CK Score</i>	Mean (95% CI) Effect Size (SD)	3.10** (0.77, 5.43) 0.03	-5.37*** (-7.67,-3.09) -0.06	-3.47** (-5.75, -1.18) -0.04	2.94* (0.65, 5.24) 0.03	11.15*** (8.85, 13.46) 0.13	11.50*** (9.12, 13.88) 0.13	416.83 (df=64)	40.8%
Model 4 <i>Demographics+ Step 2 CK Score+ Program Affiliation</i>	Mean (95% CI) Effect Size (SD)	3.33** (1.07, 5.58) 0.04	-5.68*** (-7.90,-3.46) -0.07	-3.74** (-5.96,-1.52) 0.04	2.21 (-0.02, 4.44) 0.13	10.56*** (8.32, 12.95) 0.03	10.63*** (8.32, 12.95) 0.12	64.22 (df=481)	44.6%

*P < 0.05; **P < 0.01; ***P < 0.0001; no designation implies not significant at the P > 0.05 level.

Supplemental Digital Table 3
Difference in Adjusted ABIM Board Exam Scores Resident Cohorts PGY-1 2000–2005 Compared to the Unexposed PGY-1 1999 Cohort (Reference) (N=83,802)

Model		Adjusted Score Differences						F test (df)	R ²
		PGY-1 2000 vs. 1999	PGY-1 2001 vs. 1999	PGY-1 2002 vs. 1999	PGY-1 2003 vs. 1999	PGY-1 2004 vs. 1999	PGY-1 2005 vs. 1999		
Model 1 <i>Demographics</i>	Mean (95% CI) Effect Size (SD)	-0.50 (-3.31, 2.31) -0.01	-1.09 (-3.88, 1.70) -0.01	4.18** (1.39, 6.96) 0.05	13.61*** (10.82, 16.40) 0.16	26.52*** (23.73, 29.32) 0.30	30.79*** (27.92, 33.66) 0.35	108.90 (df=55)	13.4%
Model 2 <i>Demographics+ Program Affiliation</i>	Mean (95% CI) Effect Size (SD)	-1.27 (-3.96, 1.42) -0.01	-1.91 (-4.59, 0.77) -0.02	3.24* (0.55, 5.91) 0.04	11.48*** (8.80, 14.17) 0.13	24.61*** (21.93, 27.30) 0.28	28.67*** (25.90, 31.43) 0.33	21.56 (df=472)	21.0%
Model 3 <i>Demographics+ Step 2 CK Score</i>	Mean (95% CI) Effect Size (SD)	-3.10** (-5.43, -0.78) -0.04	-8.47*** (-10.78, -6.16) -0.10	-6.57**8 (-8.88, -4.26) -0.08	0.16 (-2.48, 2.16) 0.00	8.05*** (5.72, 10.38) 0.09	8.40*** (6.00, 10.80) 0.10	416.83 (df=64)	40.8%
Model 4 <i>Demographics+ Step 2 CK Score+ Program Affiliation</i>	Mean (95% CI) Effect Size (SD)	-3.33** (-5.58, -1.07) -0.04	-9.01*** (-11.25, -6.77) -0.10	-7.07*** (-9.32, -4.83) -0.08	-1.12 (-3.38, 1.14) 0.01	7.23*** (4.96, 9.65) 0.08	7.31*** (4.97, 9.76) 0.08	64.22 (df=481)	44.6%

*P<0.5; **P<0.01; P<0.0001; no designation implies not significant at the P > 0.05 level