Published in final edited form as:

J Am Coll Surg. 2017 February ; 224(2): 137–142. doi:10.1016/j.jamcollsurg.2016.10.042.

Association between Flexible Duty Hour Policies and General Surgery Resident Examination Performance: A Flexibility in Duty hour Requirement for Surgical Trainees (FIRST) Trial Analysis

Eddie Blay Jr, MD^a, D Brock Hewitt, MD, MPH^a, Jeanette W Chung, PhD^a, Thomas Biester, MS^b, James F Fiore^b, Allison R Dahlke, MPH^a, Christopher M Quinn, MS^a, Frank R Lewis, MD, FACS^b, and Karl Y Bilimoria, MD, MS, FACS^{a,c}

^aSurgical Outcomes and Quality Improvement Center, Department of Surgery and Center for Healthcare Studies, Feinberg School of Medicine, Northwestern University, Chicago, IL

^bAmerican Board of Surgery, Philadelphia, PA

^cAmerican College of Surgeons, Chicago, IL

Abstract

Background—Concerns persist about the effect of current duty hour reforms on resident educational outcomes. We investigated whether a flexible, less-restrictive duty hour policy (Flexible Policy) was associated with differential general surgery examination performance compared to current Accreditation Council for Graduate Medical Education (ACGME) duty hour policy (Standard Policy).

Study Design—We obtained examination scores on the American Board of Surgery (ABS) In-Training Examination (ABSITE), Qualifying Examination (QE, written boards), and Certifying Examination (CE, oral boards) for residents in 117 general surgery residency programs that participated in the Flexibility In Duty Hour Requirements for Surgical Trainees (FIRST) Trial. Using bivariate analyses and regression models, we compared resident examination performance across study arms (Flexible vs. Standard Policy) for 2015 and 2016, and one year of the QE and CE. Adjusted analyses accounted for program-level factors, including the stratification variable for randomization.

Results—In 2016, 4363 general surgery residents were participants in the FIRST Trial. Mean ABSITE scores for PGY residents were not significantly different between study groups (Flexible Policy vs Standard Policy) overall (mean [SD]: Flexible: 502.6 [100.9] vs. Standard: 502.7 [98.6], $p=0.98$) or for any individual PGY levels. There was no difference in pass rates between study

Correspondence address: Karl Y. Bilimoria, MD, MS, Surgical Outcomes and Quality Improvement Center (SOQIC), Department of Surgery, Feinberg School of Medicine, Northwestern Medicine, Northwestern University, 633 N. St. Clair Street, 20th Floor, Chicago, IL 60611, k-bilimoria@northwestern.edu, Office: (312) 695-4853.

Disclosure Information: Nothing to disclose.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

arms for either the QE (Flexible: 90.4% vs Standard: 90.5%, $p=0.99$) or CE (Flexible: 86.3% vs Standard: 88.6%, $p=0.24$). Results from adjusted analyses were consistent with these findings.

Conclusions—Flexible, less-restrictive duty hour policies were not associated with differences in general surgery resident performance on examinations during the FIRST Trial. However, additional years under flexible duty hour policies may be needed to observe an effect.

Keywords

FIRST Trial; ABSITE; Qualifying Examination (QE); Certifying Examination (CE); ACGME

Introduction

In 2003, the Accreditation Council for Graduate Medical Education (ACGME) introduced resident duty hour requirements and added regulations in 2011, in response to public concerns over patient safety and resident well-being.^{1, 2} However, there have been concerns regarding the quality of resident education and training following these reforms.^{3, 4} This has been particularly debated in the surgical community where limitations on duty hours were thought to adversely affect clinical experience and technical skills. Numerous studies have analyzed the impact of duty hour reform on surrogate measures of resident competence such as operative volume, clinical acumen, and written and oral surgery board examination performance, but these studies have had methodological limitations and produced conflicting results.^{5, 6}

The Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) Trial was a prospective cluster-randomized trial which compared standard ACGME duty hour policies (Standard Policy) to flexible, less restrictive duty-hour policies (Flexible Policy) where daily duty hour limits and time off between shift requirements were waived. The FIRST Trial surveyed residents and found that residents under flexible duty hour policies were less likely to perceive negative effects of duty hour policies on various aspects of resident education (e.g., resident autonomy, clinical skills acquisition, operative volume and participation in educational conferences) compared to residents working under standard policies.⁷

Little is known about how different duty hour policies impact resident education, particularly with respect to American Board of Surgery In-Training Examination (ABSITE), the American Board of Surgery (ABS) Qualifying Examination (QE, written boards) and ABS Certifying Examination (CE, oral boards). Our objective was to examine whether flexible duty hour policies were associated with a difference in examination scores compared to standard policies.

Methods

Data and Sample

Data for this study came from the FIRST Trial. The study sample included general surgery residents in 117 general surgery residency programs in the U.S. and 151 affiliated hospitals that participated in the FIRST Trial.^{7, 8} Programs were randomly assigned to either “Standard Policy,” which required adherence to all current ACGME duty hour requirements,

or “Flexible Policy” in which the daily duty hour limits (e.g., 16 hours for interns and 28 hours for senior residents) and minimum time off between shift requirements (8 hours between daily shifts and 14 hours off after a 24-hour call) were waived through a formal waiver from the ACGME. All programs, regardless of study arm assignment, were required to adhere to the requirements regarding the 80-hour work week, one day off in seven, and call no more frequently than every third night.

We obtained data for residents in the trial on their performance on the ABSITE, QE (written boards) and CE (oral boards). The ABSITE is a standardized examination that is offered nationally every year to general surgery residents to test the acquisition of surgical knowledge and management of clinical problems related to surgery. It also provides data for program improvement in educational activities. It is standardized to a mean of 500 and standard deviation of 100. The QE, is offered to residents who have graduated from accredited general surgery residencies. The QE is a multiple choice examination testing clinical knowledge and judgment. Passing the QE is required in order to take the CE. The CE is an oral examination focused on assessing clinical judgment and decision making. Passing of both QE and CE is required for board certification by the ABS.

De-identified ABSITE, QE, and CE scores were obtained from the ABS, along with limited program-level information. ABSITE data included the January 2015 and January 2016 examinations. QE and CE data were from 2015, as there is a lag and variation in when these examinations are taken after residency graduation during the 2015-2016 academic year. Thus, we only examined QE and CE performance for the first time a resident took the examination (“first-takers”).

Statistical Analyses

Unadjusted Analyses—Mean ABSITE scores across study arms were compared using t-tests, stratified by post-graduate year (PGY). Pass rates on QE and CE were compared between study arms using chi-square tests. All tests were performed with program-level clustered standard errors to account for clustering of residents within programs.

Adjusted Analyses—A mixed linear regression model with random program-level intercepts was developed to estimate the association between study arm and mean ABSITE performance, adjusting for program-level prior year's tertile of death or serious morbidity (stratification variable in the FIRST Trial).⁸ The unit of analysis was at the resident level. We explored the inclusion of various resident and program level characteristics in the model such as gender, geography, and academic vs non-academic program status. Results from these models were qualitatively similar to those from a more parsimonious model that only adjusted for program-level tertile of death and serious morbidity. Thus we report estimates from the more parsimonious model only.

To assess the association between study arm and board examination performance (pass/fail), we estimated logistic regression models with program-level clustered standard errors and adjustment for the trial stratification variable. The unit of analysis was at the resident level. Separate models were estimated for QE and CE performance. Again, we explored the inclusion of various resident and program-level characteristics in the model, but given

similar results, we present only the parsimonious model. We did not use hierarchical logistic regression techniques to model the association between study arm and board performance because of the much smaller sample size for these data.

Analyses were performed using Stata version 13.1 (StataCorp, College Station, TX). All tests were two-sided, with a significance level of 0.05. The initial trial protocol was reviewed by the Northwestern University institutional review board office, which determined the trial to be non-human-subjects research.

Results

In 2015, 4330 residents took the ABSITE, and 4363 residents took the ABSITE in 2016. During the 2015-2016 academic year, 576 general surgery residency graduates took the QE for the first time (first-takers), and 533 took the CE for the first time.

There were no significant differences in mean ABSITE scores between study arms (flexible policy vs. standard policy) in either 2015 (503.3 [97.5] and 502.3 [100.2], $p=0.85$) or 2016 (502.6 [100.9] and 502.7 [98.6], $p=0.98$) overall or for any of the individual PGY levels (Table 1). In 2015, the mean scores for residents in the Flexible Policy group vs. Standard Policy group were as follows: PGY1 (406.7 [89.9] and 406.0 [96.6], $p=0.97$), PGY2 (495.2 [75.9] and 495.0 [77.7], $p=0.99$), PGY3 (544.6 [64.3] and 546.1 [62.4], $p=0.81$), PGY4 (563.1 [57.7] and 569.1 [58.8], $p=0.33$) and PGY5 (569.9 [72.9] and 577.6 [53.9], $p=0.31$) (Table 1). Correspondingly, there were no significant differences between the Flexible Policy and Standard Policy for each PGY in 2016. The mean scores for residents in the Flexible Policy group vs. the Standard Policy group were as follows: PGY 1 (395.0 [100.5] and 406.5 [94.1], $p=0.23$), PGY2 (498.2 [72.5] and 495.7 [78.8], $p=0.69$), PGY3 (550.2 [63.3] and 544.3 [65.4], $p=0.33$), PGY4 (562.7 [53.4] and 568.7 [54.6], $p=0.28$) and PGY5 (572.8 [54.6] and 568.6 [56.5], $p=0.44$) (Table 1).

There were no significant differences in pass rates between Flexible Policy and Standard Policy for the QE (written board; Flexible: 90.4% and Standard: 90.5%, $p=0.99$) and CE (oral boards; Flexible: 86.3% vs. Standard: 88.6%, $p=0.24$) (Table 2).

On adjusted analyses, there were no significant differences between study arms in ABSITE scores in 2015 and 2016. Non-significant differences in QE or CE performance were also found between the two study arms in 2015 on adjusted analyses. (Table 3).

Discussion

There remains considerable interest in assessing resident training after duty hour reforms. While this can be challenging to assess over time, we were able to compare different duty hour policies simultaneously in the FIRST Trial. We found that resident ABSITE, written board examination, and oral board examination performances were not associated with the duty hour policies compared in the FIRST Trial. To our knowledge this is the only study comparing surgical resident performance on standardized examination after a large-scale prospective randomized intervention on resident duty hour policies.

Objectively measuring surgical education quality is difficult given the need to assess technical skills, knowledge base, appropriate clinical decision-making, professionalism, and safe surgical judgement. Nationally, resident clinical knowledge is assessed on an annual basis by the ABS through the ABSITE. Two single institution studies examining performance before and after the 2003 ACGME duty hour changes showed conflicting results regarding ABSITE performance and cannot be generalized to all surgical training programs in the country.^{5, 6} In 2009, a multi-institutional study including both community and university programs examined ABSITE scores before and after implementation of the 2003 ACGME duty hour restrictions and found no significant difference.⁹ A national retrospective study by our group found no difference in ABSITE scores before and after implementation of the 2011 ACGME duty-hour restrictions.¹⁰ In the current study, we found that there was no difference in ABSITE scores between the Flexible Policy and Standard Policy groups of the FIRST Trial for any of the individual PGY levels over 2 years, and these findings are consistent with prior national studies. Thus, flexibility in duty hour policies does not adversely affect ABSITE performance for general surgery residents.

Pass rates on the QE (written boards) have remained consistent between 75-80% over the last 20 years.^{11, 12} Conversely, performance on the CE has varied significantly over the last 20 years; however, over the last 4 years, pass rates have stabilized.^{11,12} Previous studies have not shown a difference in QE and CE examination performance after implementation of duty hour reform.^{5, 9} A single institution study in 2006 found no difference in QE and CE pass rates when comparing performance before and after implementation of 2003 duty hour restrictions.⁵ In a national, retrospective observational study, our group found no difference in first-time performance on QE and CE pass rates before and after implementation of the 2011 ACGME duty-hour restrictions.¹⁰ In the current study, we did not find a significant difference between the Flexible Policy and Standard Policy groups with regard to performance on the QE and CE. Again, these results suggest that flexible duty hour policies are not associated with a decline in written and oral board pass rates.

It could be hypothesized that flexible duty hour policies could improve examination scores by offering residents better exposure to clinical encounters (e.g., see patients through their critical points or operations). However, it could also be suggested that flexible duty hour policies may worsen examination scores given less predictability in time to study and prepare. Given the relatively short amount of time under flexible policies in the FIRST Trial (ABSITE: one and two years; QE and CE: one year), we did not expect to see a large difference in examination performance. It likely takes several years after a change in duty hour policy before significant differences in resident examination performance may be observed. This may be particularly true for the QE and the CE.

Our study has some limitations. First, our analysis was limited to only those programs (and residents) participating in the FIRST Trial. As noted in the original FIRST Trial results publication, the trial cohort had higher board examination scores than programs that did not participate in the trial.⁷ This may limit the generalizability of our results. Second, the focus of our study was on examination performance for general surgery residents. Our findings regarding the effect of alternative duty hour policies on examination performance may be generalizable to other surgical specialties, but the impact of duty hour policies on test

performance for non-surgical specialties remains unknown. Finally, these tests examine limited aspects of surgeon training and performance. The ABSITE and QE generally assess a residency graduate's fund of knowledge, and the CE can assess aspects of clinical judgment and professionalism. However, determining whether a flexible duty hours impacts surgeon knowledge, judgment, technical skills, and professionalism is likely something that is not well reflected in these examinations.

Conclusion

Despite differences found in satisfaction with education, operative experience and autonomy, by study arm in the initial FIRST Trial results, the results from the current study suggest that flexible duty hour policies do not result in significantly different educational outcomes, as measured by the ABSITE, QE, and CE. However, measuring actual surgeon education and quality of training is more challenging and may not be possible with these examinations. Future FIRST Trial analyses will examine the impact of multiple years of flexible duty hour policies on examination performance.

Acknowledgments

The authors would like to thank all those who contributed to the administration and execution of the trial: Jonathan Fryer, MD, Anne Grace, PhD, Julie K Johnson, PhD, Lindsey J Kreutzer, MPH, Shari Meyerson, MD, Emily S Pavey, MA, Sean Perry, JD, Alfred Rademaker, PhD, and Ravi Rajaram, MD, Larry V Hedges, PhD., Remi Love, BS, David M Mahvi, MD, David D Odell, MD, Jonah J Stulberg, MD, PhD, MPH, Anthony D Yang, MD (Northwestern University); Judy Shea, PhD, Rachel R. Kelz, MD (University of Pennsylvania); Sameera Ali, MPH, Amy Hart, BS, Emma Malloy, BA, Brian Matel, BA, Craig Miller, BSEE, Lynn Modla, MS, Ajit Sachdeva, MD, and Lynn Zhou, PhD, Mark E Cohen, PhD, Clifford Y Ko, MD, David B Hoyt, MD, (American College of Surgeons); James Hebert, MD (University of Vermont); Michael Englesbe, MD, MPH, and Paul Gauger, MD (University of Michigan); Christine V Kinnier, MD (Massachusetts General Hospital); Joseph Cofer, MD (University of Tennessee, Chattanooga); John D Mellinger, MD (Southern Illinois University); Mitchell Posner, MD (University of Chicago); Eugene Foley, MD (University of Wisconsin); Thomas Louis, PhD (Johns Hopkins), and Andrew Jones, PhD (American Board of Surgery); John L Tarpley, MD (Vanderbilt), Rebecca Miller, MS, Thomas Nasca, MD, and John Potts, MD (ACGME); Margaret M Class (Defense Health Agency); all the surgeon champions and surgical clinical reviewers at the 151 participating American College of Surgeons NSQIP hospitals; and all the program directors and program coordinators at the 117 participating general surgery residency programs.

Support: The FIRST trial was funded by the American Board of Surgery, the American College of Surgeons, and the Accreditation Council for Graduate Medical Education; and ClinicalTrials.gov number, NCT02050789. Stipend for Drs Blay and Hewitt are supported by National Institutes of Health Grant #5T32HL094293.

References

1. Nasca TJ, Day SH, Amis ES. ACGME Duty Hour Task Force. The new recommendations on duty hours from the ACGME Task Force. *N Engl J Med*. 2010; 363:e3. [PubMed: 20573917]
2. Philibert I, Friedmann P, Williams WT, et al. New requirements for resident duty hours. *JAMA*. 2002; 288:1112–1114. [PubMed: 12204081]
3. Klingensmith ME, Cogbill TH, Luchette F, et al. Factors influencing the decision of surgery residency graduates to pursue general surgery practice versus fellowship. *Ann Surg*. 2015; 262:449–455. [PubMed: 26258313]
4. Yeo H, Viola K, Berg D, et al. Attitudes, training experiences, and professional expectations of US general surgery residents: a national survey. *JAMA*. 2009; 302:1301–1308. [PubMed: 19773565]
5. de Virgilio C, Yaghoobian A, Lewis RJ, et al. The 80-hour resident workweek does not adversely affect patient outcomes or resident education. *Curr Surg*. 2006; 63:435–439. [PubMed: 17084773]
6. Durkin ET, McDonald R, Munox A, Mahvi D. The impact of work hour restrictions on surgical resident education. *J Surg Educ*. 2008; 65:54–60. [PubMed: 18308282]

7. Bilimoria KY, Chung JW, Hedges LV, et al. National cluster-randomized trial of duty-hour flexibility in surgical training. *N Engl J Med*. 2016; 374:713–727. [PubMed: 26836220]
8. Bilimoria KY, Chung JW, Hedges LV, et al. Development of the Flexibility in Duty Hour Requirements for Surgical Trainees (FIRST) trial protocol: a national cluster-randomized trial of resident duty hour policies. *JAMA Surg*. 2016; 151:273–281. [PubMed: 26720622]
9. Sneider EB, Larkin AC, Shah SA. Has the 80 hour workweek improved surgical resident education in New England? *J Surg Educ*. 2009; 66:140–145. [PubMed: 19712912]
10. Rajaram R, Chung JW, Jones AT, et al. Association of the 2011 ACGME resident duty hour reform with general surgery patient outcomes and with resident examination performance. *JAMA*. 2014; 312:2374–2384. [PubMed: 25490328]
11. Lewis FR, Klingensmith ME. Issues in general surgery residency training-2012. *Ann Surg*. 2012; 256:553–559. [PubMed: 22964740]
12. American Board of Surgery. Philadelphia, PA, USA: General Surgery Examination Pass Rates. Available at: <http://www.absurgery.org/default.jsp?statgeneral> [Accessed September 13, 2016]

Annual American Board of Surgery In-Training Examination Performance for General Surgery Residents by Postgraduate Year and FIRST Trial Study Arm

Table 1

PGY	2015					2016				
	Flexible policy		Standard policy		p Value	Flexible policy		Standard policy		p Value
	No. of examinees	Mean (SD)	No. of examinees	Mean (SD)		No. of examinees	Mean (SD)	No. of examinees	Mean (SD)	
All	2,110	503.3 (97.5)	2,220	502.3 (100.2)	0.85	2,143	502.6 (100.9)	2,220	502.7 (98.6)	0.98
1	540	406.7 (89.9)	616	406.0 (96.6)	0.97	548	395.0 (100.5)	595	406.5 (94.1)	0.23
2	527	495.2 (75.9)	554	495.0 (77.7)	0.99	534	498.2 (72.5)	536	495.7 (78.8)	0.69
3	434	544.6 (64.3)	438	546.1 (62.4)	0.81	456	550.2 (63.3)	448	544.3 (65.4)	0.33
4	315	563.1 (57.7)	313	569.1 (58.8)	0.33	303	562.7 (53.4)	336	568.7 (54.6)	0.28
5	294	569.9 (72.9)	299	577.6 (53.9)	0.31	302	572.8 (54.6)	305	568.6 (56.5)	0.44

American Board of Surgery In-Training Examination scores are nationally standardized each year to an overall mean of 500 and standard deviation of 100. p Values are from t-tests for differences in group means with adjustment for program-level clustering.

Table 2
American Board of Surgery Qualifying Examination (Written) and Certifying Examination (Oral) Performance by FIRST Trial Study Arm

Examination	Flexible policy		Standard policy		p Value
	No. of examinees	% Passing	No. of examinees	% Passing	
QE (written boards)	293	90.4	283	90.5	0.99
CE (oral boards)	272	86.3	261	88.6	0.24

QE and CE results restricted to first-takers only. p Values are from chi-square tests of association with adjustment for program-level clustering.
QE, American Board of Surgery Qualifying Examination (written boards); CE, American Board of Surgery Certifying Examination (oral boards).

Table 3
Adjusted Analyses Comparing FIRST Trial Study Arms (Standard Policy vs Flexible Policy)

	Coefficient	95% CI	p Value	Odds Ratio	95% CI	p Value
ABSTE 2015 (n=4330) [*]						
Standard policy	1.0 [Ref]					
Flexible policy	-1.64	-11.49 - 8.20	0.74			
ABSTE 2016 (n= 4363) [*]						
Standard policy	1.0 [Ref]					
Flexible policy	-1.49	-11.67 - 8.71	0.77			
QE, written boards (n=576) [†]						
Standard policy				1.0 [Ref]		
Flexible policy				0.94	0.51 - 1.75	0.85
CE, oral boards (n=533) [†]						
Standard policy				1.0 [Ref]		
Flexible policy				0.76	0.46 - 1.24	0.28

^{*} Coefficients, 95% CI and p values were obtained from hierarchical linear regression models with program-level random intercepts.

[†] Odds ratios were exponentiated coefficients obtained from logistic regression models estimated with program-level robust clustered standard errors.

ABSTE, American Board of Surgery In-Training Examination; QE, American Board of Surgery Qualifying Examination (written boards; CE, American Board of Surgery Certifying Examination (oral boards)).