Does Simulator-Based Clinical Performance Correlate With Actual Hospital Behavior? The Effect of Extended Work Hours on Patient Care Provided by Medical Interns

James A. Gordon, MD, MPA, Erik K. Alexander, MD, Steven W. Lockley, PhD, Erin Flynn-Evans, RPSGT, Suresh K. Venkatan, MBBS, Christopher P. Landrigan, MD, MPH, and Charles A. Czeisler, PhD, MD, for the Harvard Work Hours, Health, and Safety Group (Boston, Massachusetts)

Abstract

Purpose
The correlation between simulator-based medical performance and real-world behavior remains unclear. This study explored whether the effects of extended work hours on clinical performance, as reported in prior hospital-based studies, could be observed in a simulator-based testing environment.

Method
Intern volunteers reported to the simulator laboratory in a rested state and again in a sleep-deprived state (after a traditional 24- to 30-hour overnight shift [n=17]). A subset also presented after a shortened overnight shift (16 scheduled hours [n=8]). During each laboratory visit, participants managed two critically ill patients. An on-site physician scored each case, as did a blinded rater later watching videotapes of the performances (score=1 [worst] to 8 [best]; average of both cases=session score).

Results
Among all participants, the average simulator session score was 6.0 (95% CI: 5.6–6.4) in the rested state and declined to 5.0 (95% CI: 4.6–5.4) after the traditional overnight shift ($P<.001$). Among those who completed the shortened overnight shift, the average postshift simulator session score was 5.8 (95% CI: 5.0–6.6) compared with 4.3 (95% CI: 3.8–4.9) after a traditional extended shift ($P<.001$).

Conclusions
In a clinical simulation test, medical interns performed significantly better after working a shortened overnight shift compared with a traditional extended shift. These findings are consistent with real-time hospital studies using the same shift schedule. Such an independent correlation not only confirms the detrimental impact of extended work hours on medical performance but also supports the validity of simulation as a clinical performance assessment tool.

Simulation has rapidly emerged as a standard component of training in health care, much as it has in other high-risk fields such as aviation, the military, and the nuclear power industry. In addition to its established role in training for enhanced performance,[1,2] realistic simulation also provides a robust platform for assessing clinical performance.[3] Such assessments promise to enhance patient safety by providing objective criteria to ensure that provider skills match patient care assignments at each stage of training and practice. Although regulatory and certification bodies[4,5] have already begun to explore the role of simulation as an assessment tool based on common sense and face validity, little empirical evidence correlates simulator-based performance with real-world behavior.

We designed this study both to explore whether physician performance, prospectively assessed in a simulator-based environment, was affected by work schedule, and to investigate whether any such impact would correlate with real-world performance under identical conditions previously assessed by hospital observers.[6] Our work emerged as a natural experiment embedded within the Harvard Intern Sleep and Patient Safety Study, which documented more sleep, fewer attentional failures, and fewer serious medical errors when 24- to 30-hour extended on-call shifts were abolished for interns in an intensive care unit (ICU) setting.[6,7] We tested the performance of interns in the simulator laboratory while rested and after overnight duty, hypothesizing that performance in the simulated environment would mirror the performance we observed in the original hospital studies. The work aimed to serve two purposes: (1) to explore the validity of simulation as an evaluation tool and (2) to validate independently the original sleep study findings regarding the effect of work hours on medical resident performance. This report concentrates on the former but also documents the latter.

Method

Design
We conducted a prospective trial designed to evaluate clinical performance in a simulator laboratory under varied sleep conditions. We hypothesized that interns would perform worse when relatively sleep deprived (tested after working a 24- to 30-hour extended on-call shift) as compared with when more rested (tested during a standard noncall clinical rotation or after a modified 16-hour overnight scheduled shift); such a finding would correlate with previously

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

Correspondence should be addressed to Dr. Gordon, Division of Medical Simulation, Department of Emergency Medicine, Massachusetts General Hospital, Zero Emerson Place, Suite 3B, Boston, MA 02114; telephone: (617) 726-7622; fax: (617) 724-0917; e-mail: jgordon3@partners.org.

Acad Med. 2010;85:1583–1588.
doi: 10.1097/ACM.0b013e3181f073f0

observed differences in actual ICU performance.[6]

Setting and population

We conducted this study in the laboratory of the Gilbert Program in Medical Simulation at Harvard Medical School from July 2003 to June 2004. The Brigham and Women's Hospital/Partners Healthcare Human Research Committee approved this work through expedited review.

The simulated environment consisted of a single "emergency department" patient bay with a full-body adult mannequin simulator (Human Patient Simulator, Medical Education Technologies, Inc., Sarasota, Florida) along with appropriate medical supplies and resuscitation equipment (e.g., oxygen, intravenous fluid, bag-valve-mask, defibrillator). The simulated patient featured a voice (transmitted through a wireless microphone), dynamic physiology (e.g., blinking eyes, palpable pulses, and auscultatory heart and lung sounds), and classic physical findings (e.g., wheezing or bradycardia). A bedside monitor provided routine vital signs (blood pressure, heart rate, pulse oximeter, and cardiac tracing). We presented all participants with two dynamic test cases to manage, and expert raters evaluated their performance (see full description of protocol below).

Our participants were postgraduate year 1 internal medicine interns who had already agreed to participate in the ongoing sleep study at Brigham and Women's Hospital and who had volunteered for the additional simulator arm of the study. Each participant received $100 in compensation per simulation session. All participants provided written informed consent per human research committee–approved protocol.

We followed two cohorts. Cohort 1 (n = 17) presented to the simulator laboratory once during a relatively rested state (during an ambulatory clinic rotation with an approximately 40-hour workweek and no overnight call responsibilities) and once again after a traditional on-call night (24 to 30 hours of continuous responsibility in the medical or cardiac ICU starting at about 7:00 AM during an every-third-night [Q3] on-call schedule). Cohort 2 (n = 8), a subset of the primary cohort, completed two additional simulator laboratory sessions later in the year: (1) at a newly rested baseline state (again during an ambulatory clinic rotation with no overnight call responsibilities) and (2) after a modified night-call (a 16-hour scheduled shift starting at about 9:00 PM, designed as part of an experimental intervention[6,7]). We separated comparison sessions (rested versus postcall) within each pair of simulation laboratory visits by less than one month, and the sessions generally occurred at the same time of day (i.e., in the late morning to early afternoon after the interns had completed their postcall work in the hospital).

We powered this study primarily to examine rested versus traditional postcall performance differences (Cohort 1) because we were initially uncertain of the sensitivity of the simulator-based testing instrument; we wanted to see whether we could detect the largest possible differences using an independent test not used in the previous hospital evaluations. As a secondary analysis, however, we experimented with Cohort 2 to assess the effects of reducing overnight shift duration, which allowed us to explore the sensitivity of the instrument for detecting smaller potential differences due to shift duration.

Protocol

For each pair of laboratory visits, we presented the intern with a warm-up case (designed to ensure familiarity with the environment) followed by two 15-minute standardized test cases: (1) a complex medical case (dynamic cardiac or pulmonary disease) followed by (2) a code (cardiopulmonary arrest—either ventricular fibrillation [VF] or pulseless ventricular tachycardia [VT]). If the intern was working in the cardiac ICU during his or her on-call rotation, we provided pulmonary testing material (asthma or chronic obstructive pulmonary disease [COPD]) in the simulator laboratory, designed to reduce the effects of contemporaneous (cardiac) learning on testing outcome. Similarly, if the intern was working in the medical ICU during his or her on-call rotation (caring for intrinsic lung and respiratory failure patients), we provided cardiac case material (inferior or anterior myocardial infarction [MI]). On return testing for each pair of sessions, the participants received the appropriate correlate case (asthma versus COPD; anterior MI versus inferior MI; VT versus VF) to guard against recall of the prior case.

We expected performance following overnight duty to be worse than performance during routine clinic duty, and we expected all interns to perform better with repetition and time; therefore, we always arranged for participants to test in the rested state before they tested after overnight duty, expecting that any practice or time effects (improvement from session 1 to session 2 over time) would counterbalance any decrement in performance due to being on call (deterioration from session 1 to 2). In this way, we biased the study toward the null hypothesis, hoping to test for truly robust experimental (sleep) effects.

Testing and outcome measures

The test cases represented classic presentations animated by a full-body simulator mannequin in an acute care setting. Each of these cases had been developed for prior research work and exhibited similar testing properties (i.e., comparable length, complexity, difficulty).[3] To guard against ordering effects within the topic domains, we varied the order in which we presented the test cases. For example, if two participants were scheduled to rotate through the cardiac care unit, Participant A would receive an asthma case followed by a VT case on his or her initial (rested) visit, and Participant B would receive a COPD case followed by VF; then, during return testing (postcall), each would receive the opposite case correlates.

We used a clinical performance evaluation tool based on an instrument previously validated for oral certification examinations in emergency medicine,[8,9] which has shown stable testing properties when used in a simulator-based environment.[3] The evaluation approach rates performance across eight domains: data acquisition, problem solving, patient management, resource utilization, health care provided, interpersonal relations, comprehension of pathophysiology, and overall clinical competence. Each domain is scored on a scale of 1 (poor) to 8 (excellent), and the total case score (also on a scale of 1 to 8) represents the average mark across all eight domains.

A domain score of 4 or less is "unsatisfactory," whereas a domain score of 5 or higher is "acceptable."

Evaluators noted critical actions in each case using a small number of checklist items (three to five) that were based on prior work and agreed on by the primary investigators (J.A.G., E.K.A.); missing a critical action significantly lowers a participant's overall score. For each session, we averaged the score of each of the two paired cases into a "session score." Either one or both of the primary physician investigators scored each session on-site in real time (if both were present, then we averaged their scores). We videotaped all the sessions so that later a third physician reviewer (S.K.V), blinded to the experimental condition, could score them. All three evaluators received uniform training on the application of the scoring rubric to ensure standardization in the assessment process.

Data analysis

We averaged individual session scores, stratified them by experimental condition, and tested for differences (paired *t* test, $P < .05$). Primary comparisons focused on performance after rest versus after traditional ICU call (Cohort 1) and performance after traditional ICU call versus after modified ICU call (Cohort 2). We compared on-site ratings with those of the blinded rater; the interrater correlation coefficient was 0.80.

Results

We conducted 50 simulator sessions (25 rested, 25 postcall) comprising 100 cases (2 per session).

Among the 17 participants in Cohort 1 (34 test sessions, 68 cases), simulator session scores averaged 6.0 (95% confidence interval [CI]: 5.6–6.4) during routine clinic duty and declined to 5.0 (95% CI: 4.6–5.4) after the traditional 24- to 30-hour extended overnight shift ($P < .001$; Figure 1). The performance of 13 of the 17 participants declined, the performance of 2 improved, and the performance of 2 more remained the same. The average change in score for these 17 was 1.0 points (Figure 2A). The proportion of interns with an average session score below 5 during the rested session (4 of 17; 24%) increased to 8 of 17 (47%) following the extended overnight shift (Figure 2A).

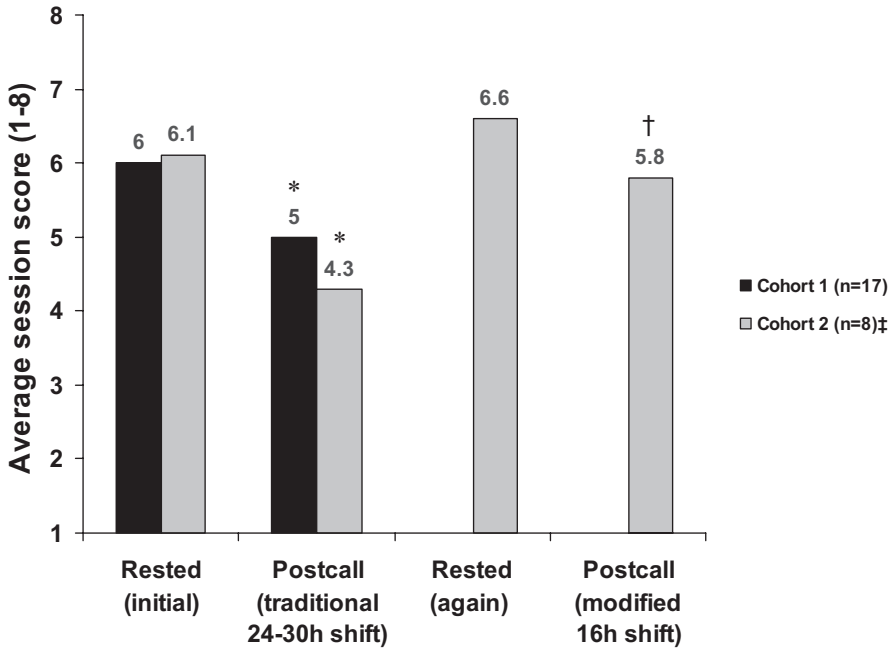

Figure 1 Average performance of medical interns in simulated acute care sessions, in both rested and postcall conditions. Postcall sessions occurred after either a traditional 24- to 30-hour extended on-call shift or a modified 16-hour overnight scheduled shift.
*$P < .001$ for comparison of initially rested vs. traditional postcall (paired *t* test, both cohorts).
†$P < .001$ for comparison of traditional vs. modified postcall (paired *t* test, cohort 2).
‡Cohort 2 is a subset of Cohort 1 that progressed through all four testing cycles.

Among the subset of 8 participants (Cohort 2) who progressed to the 16-hour overnight shift (16 additional sessions, 32 additional cases), simulator session scores averaged 6.6 (95% CI: 6.1–7.1) in a second baseline rested state, compared with 5.8 (95% CI: 5.0–6.6) after the shortened call night ($P = .036$).

Examining the average performance of the Cohort 2 subset only, across all four testing cycles (initially rested, traditional on-call, rested again, modified on-call; 32 sessions, 64 cases), the difference between initially rested performance (average score = 6.1 [95% CI: 5.6–6.6]) and traditional on-call performance (average score = 4.3 [95% CI: 3.8–4.9]) was even more pronounced than that seen in the larger group (Cohort 1) ($P < .001$). Performance after a modified 16-hour night shift (average 5.8) was significantly better than performance after a traditional 24- to 30-hour extended shift (average 4.3 [95% CI: 3.8–4.9], $P < .001$; Figure 1 and 2B); a higher proportion of interns earned a score below 5 following the 24- to 30-hour extended shift (6 of 8; 75%) as compared with the 16-hour night shift (3 of 8; 38%; Figure 2B). We observed marked variation in performance under rested and postcall conditions between individuals.

Discussion

The effect of extended work hour shifts

Our data show that average intern performance, as assessed using a high-fidelity patient simulator, is worse following an extended-duration 24- to 30-hour shift as compared with performance during a standard noncall clinical rotation. Medical simulator performance was significantly better when interns worked a 16-hour overnight shift as compared with a 24- to -30-hour shift, although their performance following the shorter night shift was still not at the level seen in their baseline rested condition. Within the cohort that completed all four conditions, 75% of the interns earned a score below 5 (the minimum for an acceptable performance) following the 24- to 30-hour extended shift; this percentage is double that which the same interns earned after working a modified 16-hour night shift. These findings mirror the difference in medical errors observed in the actual ICU setting under the same

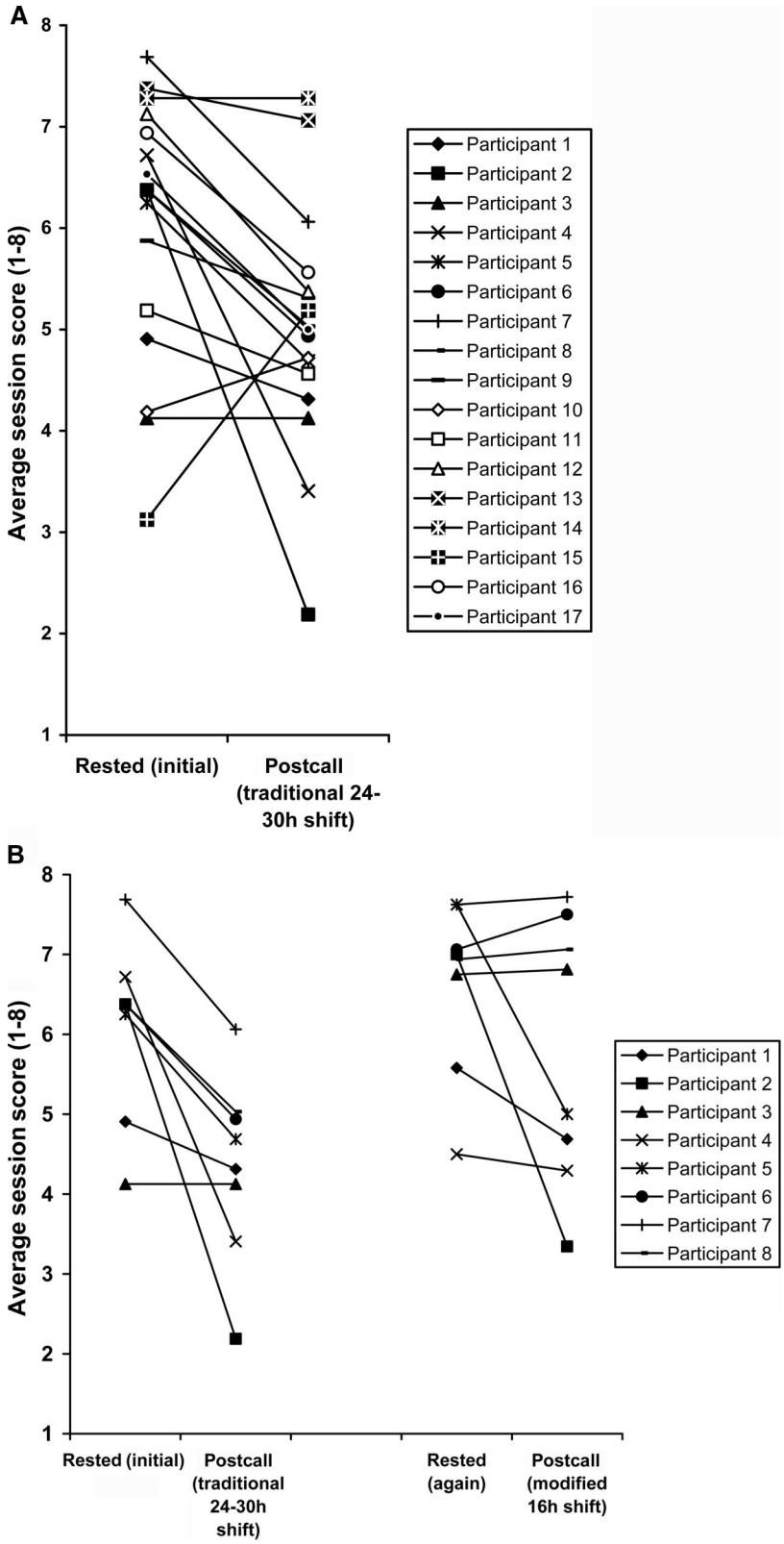

Figure 2 A. Individual performances of the 17 medical interns of Cohort 1 in simulated acute care sessions, in both rested and traditional postcall conditions. Traditional postcall sessions occurred after a 24- to 30-hour extended on-call shift. **B.** Individual performances of the 8 medical interns of Cohort 2 in simulated acute care sessions, in both rested and postcall conditions. Postcall sessions occurred after either a traditional 24- to 30-hour extended on-call shift or a modified 16-hour overnight scheduled shift.

exact scheduling conditions.[6,7] To illustrate, in the Intern Sleep and Patient Safety studies, interns under direct observation made 36% more medical errors that were categorized as serious when working a traditional 24- to 30-hour extended shift as compared with when working a schedule that limited shifts to 16 consecutive hours.[6] In addition, the extended-call interns incurred double the rate of objectively derived attentional failures when on duty overnight (from 11:00 PM until 7:00 AM).[7] In the simulator laboratory, interns scored 5.8 (on a scale of 1–8) after the scheduled 16-hour night-call versus 4.3 after the 24- to 30-hour extended shift; the latter score represents a 26% decline. In the simulation laboratory, we also observed interindividual variation in performance that echoed the interindividual variation in sleepiness observed in the original ICU setting.[7] After being on call, some sleep-deprived participants performed much worse than others under the same conditions, the performance of some participants did not change, and the performance of rare individuals actually improved. This finding is consistent with prior work suggesting inherent differences in personal susceptibility to sleep disturbance.[10–12]

Simulation and real-world performance

Whereas our findings provide independent validation of the documented effects of extended work hours on medical performance—which have informed the debate on work hours restrictions in medicine[13]—they also suggest that simulator-based performance correlates with real-world performance. Such a correlation is important in validating simulation as a tool for measuring competence and predicting safe practice across health care. Despite educators' and researchers' increasing confidence in the real-world benefits of simulator-based medical training,[1,2] evidence supporting the value of simulation-based assessment remains limited; however, dynamic simulation platforms are increasingly being considered as a means to more accurately test for action skills that traditional testing cannot capture.[4,5] If behavior in a realistic simulation laboratory truly reflects behavior in an actual patient encounter, then educators, credentialing bodies, and the public will have a powerful new tool to ensure that

provider skills match patient care responsibilities at each stage of training and practice. Moreover, targeted use of simulation training may also be helpful in mitigating the "time and chance" variability in real-world case presentations, which some cite as justification for extended-duration shifts over many years of training.[14]

Limitations

We collected the simulator-based performance data reported here from a single site, limiting generalizability; but the data are comparable with the observed ICU performance in another set of interns working on an identical schedule.[6,7] Although the intern classes from this and the previous study are similar—both derive from two subsequent classes of medical interns at Brigham and Women's Hospital—we cannot comment on any inherent differences between the two groups; however, given that both intern classes were selected on the basis of the same criteria, studied in the same program, and were admitted in successive years, we do not suspect important differences. In addition, the in-hospital performance metrics deployed in the detection of actual medical errors in the previous study are totally different from the simulator-based evaluation methods reported here. Although this distinction allows an independent appraisal of the original hospital observations, an exact comparison between the two studies is impossible. Nonetheless, the essential findings are concordant regardless of the metric. We could not blind the participants to the research hypothesis, but their level of engagement appeared uniformly high and did not seem to vary across experimental conditions, suggesting consistency in their approach to each phase of the study.

Conclusions

This simulator-based trial supports previous in-hospital work showing that interns assessed after working a 16-hour overnight shift perform better than the same interns assessed after a 24- to 30-hour extended shift. Performance under either condition was worse than baseline performance level assessed during a routine ambulatory clinic rotation, highlighting the performance decrements inherent in provision of overnight clinical coverage. These independent findings not only reinforce the conclusions of our prior studies on the effects of work hours on medical performance and patient safety, but also suggest a robust correlation between simulator-based assessment and real-world clinical behavior.

Dr. Gordon is director, Gilbert Program in Medical Simulation, associate professor of medicine, Harvard Medical School, and chief, Division of Medical Simulation, Department of Emergency Medicine, Massachusetts General Hospital, Boston, Massachusetts.

Dr. Alexander is assistant professor of medicine, Harvard Medical School, and director of medical student education, Brigham and Women's Hospital, Boston, Massachusetts.

Dr. Lockley is assistant professor of medicine, Division of Sleep Medicine, Department of Medicine, Harvard Medical School, and associate neuroscientist, Division of Sleep Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts.

Ms. Flynn-Evans is research associate, Division of Sleep Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts.

Dr. Venkatan is a teaching associate in surgery, Harvard Medical School, and simulation specialist, Division of Medical Simulation, Department of Emergency Medicine, Massachusetts General Hospital, Boston, Massachusetts.

Dr. Landrigan is assistant professor of pediatrics and medicine, Harvard Medical School, director, Sleep and Patient Safety Program, Division of Sleep Medicine, Department of Medicine, Brigham and Women's Hospital, and research director (inpatient service), Division of General Pediatrics, Department of Medicine, Children's Hospital, Boston, Massachusetts.

Dr. Czeisler is director, Division of Sleep Medicine, Baldino Professor of Medicine, Harvard Medical School, and chief, Division of Sleep Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts.

Acknowledgments: The authors would like to thank the study volunteers, and the faculty and staff of the Internal Medicine Residency Program, the Coronary Care Unit, and the Medical Intensive Care Unit at Brigham and Women's Hospital. They are grateful to the support team at the Division of Sleep Medicine, and to the community of colleagues who facilitated the work of the Intern Sleep and Patient Safety Study. They would also like to thank Yuchiao Chang, PhD, of Massachusetts General Hospital for her statistical expertise.

Funding/Support: This study was supported by grants from both the Agency for Healthcare Research and Quality (RO1 HS12032), which afforded data confidentiality protection by federal statute (Public Health Service Act; 42 U.S.C.), and the National Institute of Occupational Safety and Health within the U.S. Centers for Disease Control (RO1 OH07567), which provided a Certificate of Confidentiality for data protection. Grants from the Department of Medicine, Brigham and Women's Hospital, the Division of Sleep Medicine, Harvard Medical School, and the Brigham and Women's Hospital supported this study as well. Grants from the National Center for Research Resources awarded to the Brigham and Women's Hospital General Clinical Research Center (M01 RR02635) and the Harvard Clinical and Translational Science Center (1 UL1 RR025758) also supported this study.

Ms. Flynn-Evans is the recipient of a predoctoral fellowship in the program of training in Sleep, Circadian and Respiratory Neurobiology at Brigham and Women's Hospital (NHLBI; T32 HL079010). Dr. Lockley and Dr. Czeisler are supported in part by the National Space Biomedical Research Institute through the National Aeronautics and Space Administration (NCC 9-58).

Other disclosures: None.

Ethical approval: This work was approved by expedited review through the Brigham and Women's Hospital/Partners Healthcare Human Research Committee.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health.

Previous presentations: This work was previously presented at the Agency for Healthcare Research and Quality Patient Safety and Information Technology Annual Conference, Washington, DC, June 5, 2006; and at the International Meeting on Simulation in Healthcare, Orlando, Florida, January 14–17, 2007.

References

1 Reznick RK, MacRae H. Teaching surgical skills—Changes in the wind. N Engl J Med. 2006;355:2664–2669.
2 Wayne DB, Didwania A, Feinglass J, Fudala MJ, Barsuk JH, McGaghie WC. Simulation-based education improves quality of care during cardiac arrest team responses at an academic teaching hospital: A case–control study. Chest. 2008;133:56–61.
3 Gordon JA, Tancredi DN, Binder WD, Wilkerson WM, Shaffer DW. Assessment of a clinical performance evaluation tool for use in a simulator-based testing environment: A pilot study. Acad Med. 2003;78(10 suppl): S45–S47.
4 Berkenstadt H, Ziv A, Gafni N, Sidi A. Incorporating simulation-based objective structured clinical examination into the Israeli National Board Examination in Anesthesiology. Anesth Analg. 2006;102:853–858.
5 Gallagher AG, Cates CU. Approval of virtual reality training for carotid stenting: What this means for procedural-based medicine. JAMA. 2004;292:3024–3026.

6 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.

7 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.

8 Munger BS, Krome RL, Maatsch JC, Podgorny G. The certification examination in emergency medicine: An update. Ann Emerg Med. 1982;11:91–96.

9 Maatsch JL. Assessment of clinical competence on the Emergency Medicine Specialty Certification Examination: The validity of examiner ratings of simulated clinical encounters. Ann Emerg Med. 1981; 10:504–507.

10 Van Dongen HP, Baynard MD, Maislin G, Dinges DF. Systematic interindividual differences in neurobehavioral impairment from sleep loss: Evidence of trait-like differential vulnerability. Sleep. 2004;27: 423–433.

11 Viola AU, Archer SN, James LM, et al. PER3 polymorphism predicts sleep structure and waking performance. Curr Biol. 2007;17:613–618.

12 Czeisler CA. Medical and genetic differences in the adverse impact of sleep loss on performance: Ethical considerations for the medical profession. Trans Am Clin Climatol Assoc. 2009;120:249–285.

13 Institute of Medicine. Resident Duty Hours: Enhancing Sleep, Supervision, and Safety. Washington, DC: National Academies Press; 2009.

14 Lewis FR. Comment of the American Board of Surgery on the recommendations of the Institute of Medicine report, "Resident duty hours: enhancing sleep, supervision, and safety." Surgery. 2009;146:410–419.