Cognitive fatigue influences students' performance on standardized tests

Hans Henrik Sievertsen[a], Francesca Gino[b,1], and Marco Piovesan[c]

[a]The Danish National Centre for Social Research, 1052 Copenhagen, Denmark; [b]Harvard Business School, Harvard University, Boston, MA 02163; and [c]Department of Economics, University of Copenhagen, 1353 Copenhagen, Denmark

Edited by Pamela Davis-Kean, University of Michigan, Ann Arbor, MI, and accepted by the Editorial Board January 15, 2016 (received for review August 25, 2015)

Using test data for all children attending Danish public schools between school years 2009/10 and 2012/13, we examine how the time of the test affects performance. Test time is determined by the weekly class schedule and computer availability at the school. We find that, for every hour later in the day, test performance decreases by 0.9% of an SD (95% CI, 0.7–1.0%). However, a 20- to 30-minute break improves average test performance by 1.7% of an SD (95% CI, 1.2–2.2%). These findings have two important policy implications: First, cognitive fatigue should be taken into consideration when deciding on the length of the school day and the frequency and duration of breaks throughout the day. Second, school accountability systems should control for the influence of external factors on test scores.

cognitive fatigue | time of day | breaks | standardized tests | education

Education plays an important role in societies across the globe. The knowledge and skills children acquire as they progress through school often constitute the basis of their success later in life. To evaluate the effectiveness of schooling on children and to provide data to better manage school systems and develop education curriculum, legislators and administrators across societies have used standardized tests as their primary tool as they commonly believe test data are a reliable indicator of student ability (1, 2). In fact, these tests have become an integral part of the education process and are often used in drafting education policy, such as the *No Child Left Behind Act* and *Race to the Top* in the United States. As a result, students, teachers, principals, and superintendents are increasingly being evaluated (and compensated) based on test results (2).

A typical standardized test assesses a student's knowledge base in an academic domain, such as science, reading, or mathematics. When taking a standardized test, the substance of the test, its administration, and scoring procedures are the same for all takers (3). Identical tests, with identical degrees of difficulty and identical grading methods, are propagated as the most fair, objective, and unbiased means of assessing how a student is progressing in her learning.

The widespread use of standardized testing is based on two fundamental assumptions (3): that standardized tests are designed objectively, without bias, and that they accurately assess a student's academic knowledge. Despite these goals in the creation of standardized tests, in this paper we identify one potential source of bias that drives test results and that is predictable based on psychological theory: the time at which students take the test. We use data from a context in which the timing of the test depends on the weekly class schedule and computer availability at the school and thus is random to the individual. These factors are common conditions of standardized testing. We suggest, and find, that the time at which students take tests affects their performance. Specifically, we argue that time of day influences students' test performance because, over the course of a regular day, students' mental resources get taxed. Thus, as the day wears on, students become increasingly fatigued and consequently more likely to underperform on a standardized test. We also suggest, and find, that breaks allow students to recharge their mental resources, with benefits for their test scores.

We base these predictions on psychological research on cognitive fatigue, an increasingly common human condition that results from sustained cognitive engagement that taxes people's mental resources (4). Persistent cognitive fatigue has been shown to lead to burnout at work, lower motivation, increased distractibility, and poor information processing (5–12). In addition, cognitive fatigue is detrimental to individuals' judgments and decisions, even those of experts. For instance, in the context of repeated judicial judgments, judges are more likely to deny a prisoner's request and accept the status quo outcome as they advance through the sequence of cases without breaks on a given day (13). Evidence for the same type of decision fatigue has been found in other contexts, including consumers making choices among various alternatives (14) and physicians prescribing unnecessary antibiotics (15). Across these contexts, the overall demand of multiple decisions people face throughout the day on their cognitive resources erodes their ability to resist making easier and potentially inappropriate or bad decisions.

At the same time, research has highlighted the beneficial effects of breaks. Breaks help people recover physiologically from fatigue and thus serve a rejuvenating function (16, 17). For instance, workers who stretch physically during short breaks from data entry tasks have been found to perform better than those who do not take breaks (16). Breaks can also create the slack time necessary to identify new ideas or simply reflect (18–20), with benefit for performance.

In this paper, we build on this work by examining how cognitive fatigue influences students' performance on standardized tests. We use data on the full population of children in Danish public schools from school years between 2009/10 and 2012/13 (i.e.,

Significance

We identify one potential source of bias that influences children's performance on standardized tests and that is predictable based on psychological theory: the time at which students take the test. Using test data for all children attending Danish public schools between school years 2009/10 and 2012/13, we find that, for every hour later in the day, test scores decrease by 0.9% of an SD. In addition, a 20- to 30-minute break improves average test scores. Time of day affects students' test performance because, over the course of a regular day, students' mental resources get taxed. Thus, as the day wears on, students become increasingly fatigued and consequently more likely to underperform on a standardized test.

Author contributions: H.H.S., F.G., and M.P. designed research; H.H.S., F.G., and M.P. performed research; H.H.S. analyzed data; and H.H.S., F.G., and M.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. P.D.-K. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: fgino@hbs.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1516947113/-/DCSupplemental.

children aged 8–15) and focus on the effects of both time of the test and breaks—factors that directly relate to students' cognitive fatigue.

The Study

In Denmark, compulsory schooling begins in August of the calendar year the child turns 6 and ends after 10 years of schooling. Approximately 80% of children attend public school (14% attend private schools and 6% attend boarding schools and other types of schools). With the purpose of contributing to the continuous evaluation and improvement of the public school system, in 2010, the Danish Government introduced a yearly national testing program called The National Tests. This program consists of 10 mandatory tests: a reading test every second year (grades 2, 4, 6, and 8), a math test in grades 3 and 6, and other tests on different topics (geography, physics, chemistry, and biology) in grades 7 and 8. Each test consists of three parts, presented in random order. (Importantly, there is no ordering of the subtests. The subareas are not tested after each other; rather, a student might first get a question to subarea 1, then to subarea 2, then to subarea 1 again, then to subarea 3, and so on.) For instance, the math test is divided into Numbers and Algebra, Geometry, and Applied Math. In our analyses, we take the simple average across these three parts and standardize the score by subject, test year, and grade (with mean 0 and SD 1). This approach enables us to interpret effects in terms of SD.

These tests are adaptive: the test system chooses the questions based on the student's level of proficiency as displayed during the test and calculates the test results automatically.

Our dataset comprises all two million tests taken in Denmark between school years 2009/2010 and 2012/2013. Data are provided by the Ministry for Education and linked to administrative registers from Statistics Denmark, a government agency. The administrative data give us information about sex, age, parental background (education and income), and birth weight. The parental characteristics are measured in the calendar year prior to the test year. Our sample consists of 2,034,964 observations from 2,105 schools and 570,376 students. We excluded 17,863 observations (0.9% of the initial sample) to ensure that only normal tests (i.e., tests that were not taken under special circumstances) were included (see *SI Text* for details). We made no other sample selection.

Two characteristics of these tests should be noted. First, the main purpose of these tests is for teachers covering specific topics (e.g., geography) to gain insight into each student's achievements for the creation of individually targeted teaching plans. Teachers have no obvious incentive to manipulate students' performance, and parents are presented with the test results on a simple five-point scale.

Second, these tests are computer based: to test the students, the teacher covering a specific topic has to prebook a test session within the test period (January–April of each year). Therefore, the test time is an exogenous variable because it depends on the availability of a computer room and students' class schedules. Our analysis confirms that students are allocated to different times randomly. In fact, covariates are balanced across test time, and our results are robust to using within-student variation (i.e., variation in test time across years within the same subject for the same student, as shown in *SI Text*). In short, our data represent a natural experiment and thus a unique opportunity to test the effects of time of day and breaks on test scores.

During the school day, students have two larger breaks during which they can eat, play, and chat. Usually these breaks are scheduled around 10:00 AM and 12:00 PM and last about 20–30 minutes. As we use a large sample of 2,105 schools, and each school can organize its schedule independently, we contacted 10% of the schools by phone and asked them about their breaks schedule. We received responses from 95 schools (a 45% response rate). Our interviews revealed that 83% of the schools' first break starts between 9:20 AM and 10:00 AM and that 68% have a second break starting between 11:20 AM and 12:00 PM. Finally, we asked if test days follow a different schedule. Eighty-four percent of the schools we interviewed confirmed they follow the usual break schedule on test days. (Results using only the schools that we contacted confirm those reported below and are shown in *SI Text*.)

To test our main predictions, we first focus on the effect of test time. The upper panel of Fig. 1 shows the hour-to-hour difference in the average test score by test time. We created this graph by estimating a linear model of test score on indicators for test hour using ordinary least squares (OLS). In the model, we control for school, grade, subject, day of the week, and test-year fixed effects, as well as for parental education, parental income, birth weight, sex, spring child, and origin. As the graph shows, time of day influences test performance in a nonlinear way: although the average test score deteriorates from 8:00 to 9:00 AM, it improves from 9:00 to 10:00 AM. This alternating pattern of improvements and deterioration continues during the day (see *SI Text* for details).

Next, we focus on the effect of having the test after a typical break. By typical break, we mean a break that commonly occurs at the same time throughout the week, across schools. The dashed line in the lower part of Fig. 1 shows the breaks time. Breaks typically end just before 10:00 AM and 12:00 PM. Together, the

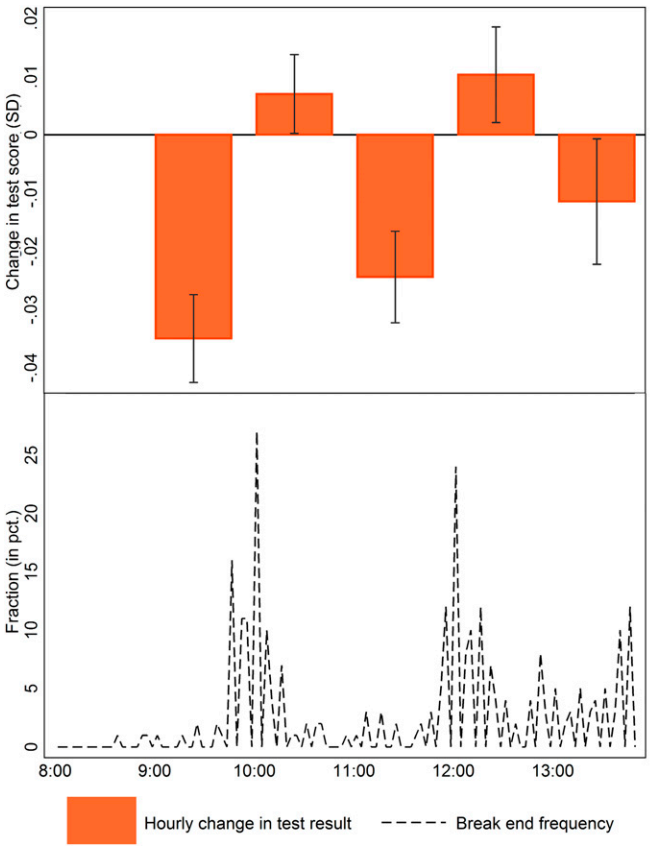

Fig. 1. Hour-to-hour effect on test scores and break patterns. Effects are estimated based on administrative data from Statistics Denmark. (*Upper*) How the average test score changes from hour to hour. (*Lower*) Distribution of when breaks end, based on a survey conducted on 10% of the schools. The hourly effect is estimated in a linear model controlling for unobserved time invariant fixed effects on grade, day of the week, and school level. We also control for test year fixed effects, as well as parental income, parental education, nonwestern origin, sex, spring child, and birth weight. The details on the model and estimation procedure are shown in *SI Text*, along with a table with regression results.

hour-to-hour changes and the break pattern show that test performance declines during the day but improves at test hours just after a break. Breaks, it appears, recharge students' cognitive energy, thus leading to better test scores.

Next, to provide further support for our hypotheses, we explicitly model the effects of time of day and breaks by estimating the linear relationship between test score, test hour, and breaks. The model is estimated by OLS and also includes the individual characteristics and the fixed effects described above. The point estimates on break and test hour are shown in Fig. 2. Fig. 2*A* shows these point estimates for various specifications and subsamples. The first two bars show that for the full sample, the test score is reduced by 0.9% (95% CI, 0.7–1.0%) of an SD for every hour (the red bar), but a break improves the test score by 1.7% (95% CI, 1.2–2.2%) of an SD (the blue bar). We then conduct the same analyses for various subsamples but find limited evidence of heterogeneous effect of breaks across subject (i.e., mathematics vs. reading) and age (i.e., young vs. old). For hour of the day, the effect is more pronounced for tests in mathematics and older children. The last four bars show that the results are robust to two important robustness checks: using only data on the subsample of schools we included in the break survey and using only within students' variation in test hour (i.e., including individual fixed effects). In this individual fixed effects specification, we remove any individual time-and-subject-invariant unobserved effect, but still find the same pattern of improvements during breaks and deterioration for every hour later in the day the test is taken. These effects, therefore, are not driven by selection of students into specific times of the day.

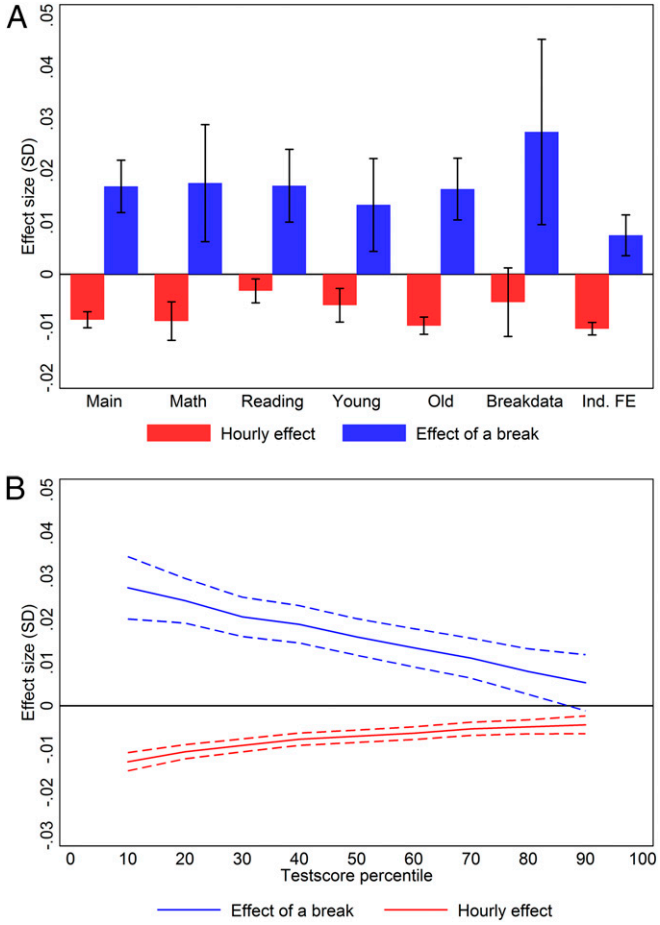

Fig. 2. Effect of time of day and breaks. Effects are estimated based on administrative data from Statistics Denmark. The figures show the parameter estimates for break and test hour from estimating a linear model of test score on test hour and break and controlling for test year fixed effects, as well as parental income, parental education, nonwestern origin, sex, spring child, and birth weight. We also control for school, grade, subject, and day of the week fixed effects. The details on the model and estimation procedure are shown in *SI Text*, along with a table with regression results. (*A*) Main effect and the effect by subgroups. (*B*) Results from quantile regression at the 10th, 20th, 30th, 40th,. . .,90th percentiles using Canay's plugin fixed effect estimator (21). The graph shows the effect of breaks and test hour over the test score distribution.

Fig. 2*B* shows the heterogeneous effects of test hour and breaks on different percentiles of the test score distributions. The graph was created based on quantile regressions and shows the effect of breaks and test hour for different percentiles of the test score distribution. This analysis shows that both breaks and time of day affect the lower end of the distribution, i.e., the low performing students, significantly more than the upper end of the test score distribution, i.e., the high performing students. For the 10th percentile, a break causes 2.7% (95% CI, 2.0–3.5%) of an SD improvement in test score, and for every hour later in the day, the test performance worsens by 1.3% (95% CI, 1.0–1.5%) of an SD. At the upper end of the distribution, there is no effect of breaks on performance, and for every hour, the test score declines by only 0.4% (95% CI, 0.2–0.6%) of an SD.

Overall, the results of our analyses provide support for our hypotheses that taking tests later in the day worsens performance and taking tests after a break improves performance.

Conclusion

Standardized testing is commonly used to assess student knowledge across countries and often drives education policy. Despite its implications for students' development and future, it is not without bias. In this paper, we examined the influence of the time at which students take tests and of breaks on test performance. In Denmark, as in many other places across the globe, test time is determined by the weekly class schedule and computer availability at schools. We find that, for every hour later in the day, test scores decrease by 0.9% SDs. In addition, a 20- to 30-minute break improves average test scores. Importantly, a break causes an improvement in test scores that is larger than the hourly deterioration. Therefore, if there was a break after every hour, test scores would actually improve over the day. However, if, like in the Danish system, there is only a break every other hour, the total effect is negative. Our results also show that low-performing students are those who suffer more from fatigue and benefit more from breaks. Thus, having breaks before testing is especially important in schools with students who are struggling and performing at low levels.

To understand effect sizes, we computed the simple correlations between test score and parent income, parental education, and school days (see *SI Text* for details). We find that an hour later in the day causes a deterioration in test score that corresponds to 1,000 USD lower household income, a month less parental education, or 10 school days. A break causes an improvement in test score that corresponds to about 1,900 USD higher household income, almost 2 months of parental education, or 19 school days. The effect sizes are small but nonnegligible compared to the unconditional influence of individual characteristics.

Importantly, the students in our sample are young children and early adolescents, and older adolescents may fare differently. We hope future research will investigate this possibility. Future work could also examine other forms of potential variation in students' performance on standardized tests, including circadian rhythms (22). In fact, research has shown individuals' cognitive functioning (e.g., memory and attention) is at its peak at their optimal time of day and decreases substantially at their nonoptimal times (23–25).

Our results should not be interpreted as evidence that that the start time of the school day should change to later (thus allowing students to sleep in, as currently debated in the United States) or

that schools tests should be administered earlier in the day. Rather, we believe these results to have two important policy implications: first, cognitive fatigue should be taken into consideration when deciding on the length of the school day and the frequency and duration of breaks. Our results show that longer school days can be justified, if they include an appropriate number of breaks. Second, school accountability systems should control for the influence of external factors on test scores. How can school systems handle such potential biases? One approach would be to adjust the test scores based on the parameters identified in this paper. Based on our results, policy makers should adjust upward test scores by 0.9% of an SD for every hour later in the day the test is taken, and adjust downward tests after breaks with 1.7% of an SD. We recognize that this approach may not always be feasible to implement in practice given that it would require continuous monitoring and adjustments. A more straightforward approach would be to plan tests as closely after breaks as possible. Moreover, as breaks and time of day clearly affect students' test performance, we also expect other external factors like hunger, light conditions, and noise to play a role. These external factors should be accounted for when comparing test scores across children and schools.

Data and Methods

Here we describe how to obtain access to the data analyzed in our paper. For additional methodological detail, full results, and tables, please refer to *SI Text*. The project was carried out under Agreement 2015-57-0083 between The Danish Data Protection Agency and the Danish National Centre for Social Research. Specifically, this study was approved by the research board of the Danish National Centre for Social Research under Project US2280 and approved by the Danish protection agency under Agreement 2015-57-0083. We note that there is no Danish institutional review board for studies that are not randomized controlled trials.

The analyses are based on data from administrative registers on the Danish population provided by Statistics Denmark and the Danish Ministry for Education. All analyses have been conducted on a server hosted by Statistics Denmark and owned by The Danish National Centre for Social Research (SFI server project number 704335). All calculations were done with the software STATA (version 13.0). Given that these data contain personal identifiers and sensitive information for residents, they are confidential under the Danish Administrative Procedures (§27) and the Danish Criminal Code (§152). Therefore, we cannot make the data publicly available. However, independent researchers can apply to Statistics Denmark for access, and we will assist in this process in any way we can. If interested researchers request and obtain access to the data, they can use the stata code included in *SI Appendix* to reproduce the results of the analyses reported in the paper and in the *SI Text*.

Statistics Denmark requires that researchers who access the confidential information receive approval by a Danish Research Institute. The Danish National Centre for Social Research is willing to grant researchers access to this project, given that they satisfy the existing requirements. As of today, the formal requirements involve a test in data policies and a signed agreement. More information on the Danish National Centre for Social Research can be found at www.sfi.dk.

The Danish Ministry for Education granted us access to all test results from the mandatory National Tests in Danish Public Schools between school years 2009/2010 and 2012/2013. The data were sent from the Ministry to Statistics Denmark. Statistics Denmark anonymized the personal identifiers and provided information on each student's birth weight, parental income rank, parental education (years), and sex. Before analyzing the data, we excluded 14,945 tests that were taken at 2:00 PM and 2,918 tests that were taken in grades and subject combinations that are out of schedule. The pattern of results for tests occurring at 2:00 PM is in line with the overall conclusions we draw in our research, but this test time was so uncommon that we excluded it from the sample. In total we excluded 17,863 of 2,052,827 observations (0.9% of the raw sample). All conclusions remain unchanged if we conduct the analyses on the raw sample. The sample selection is done to ensure that the analysis is based on normal tests and not tests that were taken under special circumstances.

ACKNOWLEDGMENTS. We thank seminar participants from the Copenhagen Education Network for comments. We appreciate the helpful comments we received from Ulrik Hvidman, Mike Luca, Alessandro Martinello, and Todd Rogers on earlier drafts. H.H.S. acknowledges financial support from Danish Council for Independent Research Grant 09-070295.

1. US Legal I (2014) Standardized test [education] law & legal definition. Available at definitions.uslegal.com/s/standardized-test-education/. Accessed January 29, 2016.
2. Robelen EW (2002) An ESEA primer. *Educ Week* February:21.
3. Koretz D, Deibert E (1996) Setting standards and interpreting achievement: A cautionary tale from the National Assessment of Educational Progress. *Educ Assess* 3(1): 53–81.
4. Mullette-Gillman OA, Leong RLF, Kurnianingsih YA (2015) Cognitive fatigue destabilizes economic decision making preferences and strategies. *PLoS One* 10(7):e0132022.
5. Demerouti E, Bakker AB, Nachreiner F, Schaufeli WB (2001) The job demands-resources model of burnout. *J Appl Psychol* 86(3):499–512.
6. Holding D (1983) *Fatigue. Stress and Fatigue in Human Performance* (John Wiley & Sons, New York).
7. Boksem MA, Meijman TF, Lorist MM (2005) Effects of mental fatigue on attention: An ERP study. *Brain Res Cogn Brain Res* 25(1):107–116.
8. Lorist MM, Boksem MA, Ridderinkhof KR (2005) Impaired cognitive control and reduced cingulate activity during mental fatigue. *Brain Res Cogn Brain Res* 24(2):199–205.
9. Sanders AF (1998) *Elements of Human Performance: Reaction Processes and Attention in Human Skill* (Lawrence Erlbaum Associates, London).
10. van der Linden D, Frese M, Meijman TF (2003) Mental fatigue and the control of cognitive processes: Effects on perseveration and planning. *Acta Psychol (Amst)* 113(1):45–65.
11. Boksem MA, Meijman TF, Lorist MM (2006) Mental fatigue, motivation and action monitoring. *Biol Psychol* 72(2):123–132.
12. Hockey GRJ, John Maule A, Clough PJ, Bdzola L (2000) Effects of negative mood states on risk in everyday decision making. *Cogn Emotion* 14(6):823–855.
13. Danziger S, Levav J, Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. *Proc Natl Acad Sci USA* 108(17):6889–6892.
14. Vohs KD, et al. (2008) Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *J Pers Soc Psychol* 94(5):883–898.
15. Linder JA, et al. (2014) Time of day and the decision to prescribe antibiotics. *JAMA Intern Med* 174(12):2029–2031.
16. Henning RA, Sauter SL, Salvendy G, Krieg EF, Jr (1989) Microbreak length, performance, and stress in a data entry task. *Ergonomics* 32(7):855–864.
17. Gilboa S, Shirom A, Fried Y, Cooper C (2008) A meta-analysis of work demand stressors and job performance: Examining main and moderating effects. *Person Psychol* 61(2): 227–271.
18. Smith SM (1995) Getting into and out of mental ruts: A theory of fixation, incubation, and insight. *The Nature of Insight*, eds Sternberg RJ, Davidson JE (MIT Press, Cambridge, MA), pp 229–251.
19. Leonard D, Swap W (1999) *When Sparks Fly: Igniting Creativity in Groups* (Harvard Business School Press, Boston).
20. Schön DA (1983) *The Reflective Practitioner: How Professionals Think in Action* (Basic Books, New York).
21. Canay IA (2011) A simple approach to quantile regression for panel data. *Econ J* 14(3): 368–386.
22. Yoon C, May CP, Hasher L (1999) Aging, circadian arousal patterns, and cognition. *Aging, Cognition and Self Reports*, eds Schwarz N, Park D, Knauper B, Sudman S (Psychological Press, Washington, DC), pp 117–143.
23. Blake MJF (1967) Time of day effects on performance in a range of tasks. *Psychon Sci* 9(6):349–350.
24. Goldstein D, Hahn CS, Hasher L, Wiprzycka UJ, Zelazo PD (2007) Time of day, intellectual performance, and behavioral problems in morning versus evening type adolescents: Is there a synchrony effect? *Pers Individ Dif* 42(3):431–440.
25. Randler C, Frech D (2009) Young people's time-of-day preferences affect their school performance. *J Youth Stud* 12(6):653–667.
26. Hayashi F (2000) *Econometrics* (Princeton Univ, Princeton).
27. Cameron AC, Miller DL (2015) A practitioner's guide to cluster-robust inference. *J Hum Resour* 50(2):317–372.