AER: Insights 2019, 1(3): 291–308
https://doi.org/10.1257/aeri.20180633

Measuring Success in Education: The Role of Effort on the Test Itself[†]

By Uri Gneezy, John A. List, Jeffrey A. Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu*

US students often rank poorly on standardized tests that estimate and compare educational achievements. We investigate whether this might reflect not only differences in ability but also differences in effort on the test. We experimentally offer students incentives to put forth effort in two US high schools and four Shanghai high schools. US students improve performance substantially in response to incentives, while Shanghai students—who are top performers on assessments—do not. These results raise the possibility that ranking countries based on low-stakes assessments may not reflect only differences in ability, but also motivation to perform well on the test. (*JEL* C93, I21, I26, O15, P36)

It is difficult to overstate the value of improving education policies for both individuals and countries. A critical input to achieving improvement is accurate measurement of student learning. To that end, policymakers are increasingly interested in using student assessment tests to evaluate the quality of teachers, schools, and entire education systems. The results of these assessment tests have often raised concerns that students in the United States are falling behind their peers in other countries. For example, on the 2015 National Assessment of Educational Progress, only 40 percent of fourth graders and one-third of eighth graders performed at or above proficient levels in mathematics (National Center for Education Statistics 2015). Similarly, on the 2012 Programme for International Student Assessment (PISA), among the 65 countries and economies that participated, US high school students

*Gneezy: Rady School of Management, University of California-San Diego, 9500 Gilman Drive, La Jolla, CA 92093, University of Amsterdam, and CREED (email: ugneezy@ucsd.edu); List: The Kenneth C. Griffin Department of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, and NBER (email: jlist@uchicago.edu); Livingston: Bentley University, 175 Forest Street, Waltham, MA 02452 (email: jlivingston@bentley.edu); Qin: Antai College of Economics and Management, Shanghai Jiao Tong University, 1954 Hua Shan Road, Shanghai 200030, China (email: xdqin@sjtu.edu.cn); Sadoff: Rady School of Management, University of California-San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (email: ssadoff@ucsd.edu); Xu: The Kenneth C. Griffin Department of Economics, University of Chicago, 5757 S. University Avenue, Chicago, IL 60637 (email: yangxu@uchicago.edu). Amy Finkelstein was the coeditor for this article. We thank the University of Chicago for generous financial support. Katie Auger, Richie Aversa, Debbie Blair, Jonathan Davis, Clark Halliday, Claire Mackevicius, and Daniel Mather provided excellent research assistance. This research was conducted with approval from the University of Chicago Institutional Review Board, IRB15-0448. The study is registered in the AEA RCT Registry, AEARCTR-0003657.

[†]Go to https://doi.org/10.1257/aeri.20180633 to visit the article page for additional materials and author disclosure statement(s).

ranked thirty-sixth for mathematics performance, with scores declining since 2009 (Organisation for Economic Co-operation and Development 2014).

In response to poor US performance on such assessments, then-US Secretary of Education Arne Duncan quipped, "We have to see this as a wake-up call. I know skeptics will want to argue with the results, but we consider them to be accurate and reliable … We can quibble, or we can face the brutal truth that we're being out-educated."[1] Student performance on international assessments also has had a demonstrable impact on policy in Europe. In Finland, which performed unexpectedly well on the 2000 PISA, analysts noted that their school practices were now a model for the world, while Germany, which surprisingly underperformed, convened a conference of ministers and proposed urgent changes to improve the system (Grek 2009).

Why does the United States perform so poorly relative to other countries despite its wealth and high per pupil expenditures? Examples of answers discussed in the literature include differences in learning due to socioeconomic factors, school systems, and culture (e.g., Carnoy and Rothstein 2013, Woessmann 2016, Stevenson and Stigler 1992). In line with recent work using observational data (see, e.g., Borghans and Schils 2013; Zamarro, Hitt, and Mendez 2016; Borgonovi and Biecek 2016), we consider an additional potential reason: students in different countries may have heterogeneous levels of intrinsic motivation to perform well on assessment tests. If so, poor US performance relative to other countries may be partially explained by differential effort on the test itself. The degree to which test results actually reflect differences in ability and learning may be critically overstated if gaps in intrinsic motivation to perform well on the test are not understood in comparisons across students. Such differences are particularly important in the context of low-stakes assessments because students have no extrinsic motivation to perform well on these tests.

In this study, we present an experimental methodology for comparing test effort across student groups. We conduct an experiment in the United States and in China, between which there has historically been a large performance gap on standardized tests. In order to explore the gap in intrinsic motivation, we offer students at four schools in Shanghai and two schools in the United States a surprise financial incentive to put forth effort on a low-stakes test. We compare their performance to students who are not given an incentive. Importantly, students learn about the incentive just before taking the test, so any impact on performance can only operate through increased effort on the test itself rather than through, for example, better preparation or more studying.

If baseline effort on these tests varies across countries and cultures, then we hypothesize a differential responsiveness to extrinsic incentives. Among students who are deeply motivated to work hard at baseline, we expect incentives to have little impact on performance since they are already at or near their output frontier. In contrast, among students who lack motivation at baseline, extrinsic financial incentives have more scope to increase effort and improve performance. Moving less

[1] See, Sam Dillon, "Top Test Scores from Shanghai Stun Educators," *New York Times*, December 7, 2010, http://www.nytimes.com/2010/12/07/education/07education.html.

intrinsically motivated students closer to their output frontier will result in a better measurement of relative ability across students.[2]

Our results are consistent with this hypothesis. In response to incentives, the performance of the Chinese students does not change while the scores of US students increase substantially. Under incentives, US students attempt more questions (particularly toward the end of the test) and are more likely to answer those questions correctly. These effects are concentrated among students whose baseline performance is near the US average. It is important to note that our experimental samples are not representative (nor drawn from the same parts of their respective distributions) and therefore cannot stand in for the world distribution. We instead emphasize that our results raise the possibility that students in different countries may have different levels of intrinsic motivation to perform well on low-stakes assessment tests, which complicates the challenges of international test comparisons.

I. Background Literature

The finding that scores on low-stakes tests do not always reflect students' true ability has already been recognized in the literature (Wise and DeMars 2005 and Finn 2015 provide reviews). One strand of research uses observational data to examine correlations between performance and proxies for motivation and effort, including self-reported motivation, interest, attitudes and effort, fast response times, low item response rates, and declining performance over the course of the test (e.g., Eklöf 2010; DeMars and Wise 2010; Borghans and Schils 2013; Zamarro, Hitt, and Mendez 2016; Borgonovi and Biecek 2016; Balart, Oosterveen, and Webbink 2017; Akyol, Krishna, and Wang 2018).[3] Yet, important for our purposes, these studies are not able to identify the impact of effort separately from the impact of ability. For example, low self-reported effort and rapid guessing may indicate that the student does not try hard because he or she is unable to answer the questions; and low response rates and declining performance may partially reflect lower ability to work quickly or maintain focus rather than lower levels of motivation to do so (Sievertsen, Gino, and Piovesan 2016). It is therefore difficult to estimate from these studies whether increased motivation would translate into increased performance.

To address this concern, a second strand of the literature has used randomized interventions to exogenously vary extrinsic motivation to exert effort on the test.[4] These studies demonstrate that rewards (both financial and non-financial) as well

[2] These hypotheses can be formalized using the framework of DellaVigna and Pope (2018). Students choose optimal effort to equalize the marginal costs of effort, which are convex, with the marginal benefits of effort, which are the sum of intrinsic motivation (i.e., motivation absent extrinsic incentives) and extrinsic financial incentives. In the absence of extrinsic incentives, a student with high intrinsic motivation will exert more effort than a student with low intrinsic motivation. However, when extrinsic incentives are introduced, the *change* in effort in response to the same financial incentive will be larger for a student with low intrinsic motivation (due to the convexity of the effort cost function). See online Appendix Figure A.1 for an illustration.

[3] For example, Borghans and Schils (2013) and Zamarro, Hitt, and Mendez (2016) show that the effort students put into surveys that are given after completing the PISA correlates with declining performance over the course of the test. They argue that differential motivation and effort can explain about one-fifth to two-fifths of the variation in test scores across countries.

[4] Note that these studies are distinct from the rich literature that offers financial incentives to encourage preparation for exams and other learning activities (e.g., Fryer 2011; Levitt, List, and Sadoff 2016; Barrow and Rouse 2018, provide a review).

as how the test is framed can increase effort and improve performance (Duckworth et al. 2011; Braun, Kirsch, and Yamamoto 2011; Levitt et al. 2016; Jalava, Joensen, and Pellas 2015). Recent work in education and behavioral economics has investigated how to best structure incentives (Gneezy, Meier, and Rey-Biel 2011). Critical factors for motivating effort include: simplicity of performance criteria; credibility of actual payment; salience and stakes (incentives must be substantial enough for the students to care about); framing (e.g., framed as losses rather than gains); and, the timing of payment (immediately after the test rather than with a delay).

Building on this research, we structured our incentives to best impact behavior. We framed the incentives as losses provided in the form of upfront cash rewards, which increases their salience and credibility. We wish to emphasize that the goal of this paper is not to study how incentives work, but rather to use incentives as an experimental tool to understand the interaction of culture with motivation to do well on the test. Previous studies have noted that differential motivation can lead to biases in measures of achievement gaps. To the best of our knowledge, however, our study is novel in that we are the first to experimentally show the relevance of this underestimation of true ability for the interpretation of ability gaps across cultures on low-stakes tests.

In this spirit, with respect to the students in our sample, observational studies find that proxies for effort, such as survey response rates and consistent performance over the course of the test, are higher on average in East Asian countries than in the United States (Zamarro, Hitt, and Mendez 2016). There is also evidence from descriptive studies showing that, compared to the United States, East Asian parents, teachers, and students put more emphasis on diligence and effort (Stevenson and Stigler 1992; Stevenson et al. 1990; Hess, Chang, and McDevitt 1987). Traditional East Asian values also emphasize the importance of fulfilling obligations and duties (Aoki 2008). These include high academic achievement, which is regarded as an obligation to oneself as well as to the family and society (Tao 2016, Hau and Ho 2010). Hence, East Asian students may put forth higher effort on standardized tests if doing well on those tests is considered an obligation.

II. Experimental Design

We conducted the experiment in high schools in Shanghai, which was ranked first in mathematics on the 2012 PISA test, and in the United States, which was ranked thirty-sixth on the same test. The PISA is conducted by the Organisation for Economic Co-operation and Development (OECD) in member and non-member nations. Administered every three years since 2000, the test assesses 15-year-olds in mathematics, science, and reading with the goal of allowing educators and policymakers to learn what works better in advancing the success of students.[5]

Our experiment was conducted in the spring and fall of 2016 in the United States and Shanghai and in the spring of 2018 in Shanghai only. In all experimental sessions students took a 25-minute, 25-question mathematics test that we constructed from questions that have been used on the mathematics PISA in the past.[6] The exam

[5] See http://www.oecd.org/pisa/aboutpisa/.

[6] The questions are drawn from PISA tests given in 2000, 2003, and 2012 (https://www.oecd.org/pisa/pisaproducts/Take%20the%20test%20e%20book.pdf and https://nces.ed.gov/surveys/pisa/pdf/items2_math2012.pdf).

consists of 13 multiple-choice questions and 12 free answer fill-in-the-blank questions (see online Appendix B for the test questions). To determine the question order, we first grouped related questions together and then assigned a random number to each group. For example, questions 14 through 16 all reference the same bar chart, so they were kept together. The question order was the same for all students. As shown in online Appendix Figure A.2, the worldwide percentage of students who answered each question correctly when the questions were administered as part of official PISA exams ranges from 25.7 to 87.3, with little correlation between question difficulty and question order on the test ($p = 0.14$). US students took the test in English and Shanghai students took the test in Mandarin.

The experiment was conducted in two high schools in the United States and four high schools in Shanghai. While our samples are not nationally representative, we aimed to sample students throughout their respective distributions. The US sample includes a high-performing private boarding school and a large public school with both low- and average-performing students. The Shanghai sample includes one below-average-performing school, one school with performance that is just above average, and two schools with performance that is well above average.[7] In the United States, all students in tenth grade math classes were selected to participate.[8] In Shanghai, we randomly selected approximately 25 percent of tenth grade classes in each school to participate. All students present on the day of testing took part in the experiment.[9]

We randomly assigned students to either the Control (no incentive) group or the Treatment (incentive) group. The US sample includes 447 students (227 in Control and 220 in Treatment) and the Shanghai sample includes 656 students (333 in Control and 323 in Treatment).[10] Students in the Control group received no incentive for their performance on the test. In the incentive treatment, US students were given an envelope with $25 in one dollar bills and were told that the money was theirs, but that we would take away one dollar for each question that was answered incorrectly (unanswered questions counted as incorrect). Immediately after students completed the test, we took away any money owed based on their performance. In Shanghai, students received the equivalent in renminbi (RMB).[11]

Importantly students had no advance notice of the incentives. Immediately before they took the test, students read the instructions along with the experiment administrator (see online Appendix C for instructions). Accordingly, we are assured that the incentives only influence effort on the test itself, not preparation for the exam.

[7] School performance is rated compared to the average Shanghai 2015 Senior High School Entrance Examination score of 473.5. The average 2015 scores for the four schools (from lowest to highest) were: 464, 516.5, 552, and 573.5. The 2016 sessions in Shanghai included all but the second highest performing school. The 2018 sessions included all but the lowest performing school.

[8] In the lower performing school, 81 percent of tenth graders were enrolled in tenth grade math. The remainder were enrolled in ninth (18 percent) or eleventh (1 percent) grade math. The tenth grade math classes also included 89 non-tenth graders who are excluded from our primary analysis.

[9] In the higher performing US school, 11 students arrived late due to a prior class and did not participate.

[10] The sample sizes in order of school performance (lowest to highest) in the United States are: $n = 341$ and $n = 106$; and in Shanghai are: $n = 60$, $n = 208$, $n = 126$, and $n = 262$.

[11] We used the Big Mac Index to determine currency conversion (http://www.economist.com/content/big-mac-index). The implied exchange rate in January 2016 was 3.57. By this index $25 converts to RMB 89.25. We rounded up and gave students in the Treatment group RMB 90 and took away RMB 3.6 for each incorrect answer.

Table 1—Sample Characteristics by Treatment Group

	United States			Shanghai		
	Control	Treatment	National average	Control	Treatment	National average
Female	0.50	0.49	0.49	0.54	0.52	0.53
White	0.39	0.45	0.50			
Black	0.18	0.18	0.16			
Asian	0.07	0.06	0.05			
Hispanic white	0.30	0.27	0.25			
Hispanic non-white	0.05	0.03	0.03			
Other	0.00	0.01	0.01			
Age	16.19	16.06		16.23	16.17	
	(0.76)	(0.65)		(0.42)	(0.38)	
Standardized baseline exam score	−0.09	0.09		0.01	−0.01	
	(0.94)	(1.05)		(0.93)	(1.07)	
Missing baseline exam score	0.24	0.20		0	0	
	(0.43)	(0.40)		(0)	(0)	
Observations (students)	227	220		333	323	

Notes: The table reports group means. Standard deviations in parentheses. US national 10th grade averages for gender and ethnicity categories are computed from enrollment numbers from the US Department of Education, National Center for Education Statistics, Common Core of Data (CCD), Local Education Agency (School District) Universe Survey Membership Data, 2015–16 v.1a. The US national average for Hispanic Non-white includes all multi-racial 10th graders. The Shanghai-wide average percentage of female students is reported by the Shanghai Municipal Education Bureau. The baseline exam is the 7th grade Massachusetts Comprehensive Assessment System test in mathematics for US school 1, the Quantitative Secondary School Admissions Test (SSAT) for US school 2, and the Senior High School Entrance Examination for the Shanghai schools. These baseline test scores are standardized within sample separately for each test. No within-country differences between Treatment and Control are significant at the 10 percent level for any characteristic.

We randomized at the class level in the lower performing school in the United States and in the 2016 sessions in Shanghai. We randomized at the individual level in the higher performing school in the United States and in the 2018 sessions in Shanghai.[12] In the United States, we stratified by school and re-randomized to achieve balance on the following baseline characteristics: gender, ethnicity, and mathematics class level/track: low, regular, and honors.[13] For each school's randomization, we re-randomized until the *p*-values of all tests of differences between Treatment and Control were above 0.4. In the 2016 Shanghai sessions, we stratified the randomization by school (baseline demographics were not available at the time of randomization). In the 2018 Shanghai sessions, we stratified the randomization by class, gender, and senior entrance exam score quartile.

III. Results

Table 1 presents the results of the randomization and average test scores by treatment group and country. We also present national averages where applicable and available. The table displays means of student characteristics (gender, age, and race/ethnicity) and a baseline exam score. The exam scores are standardized within

[12] Differences in the randomization across schools and waves of the experiment were driven by logistical constraints. We randomized at the individual level when possible.

[13] We did not balance the randomization on baseline test scores because they were not available at the time of the randomization and are missing for 22 percent of the sample.

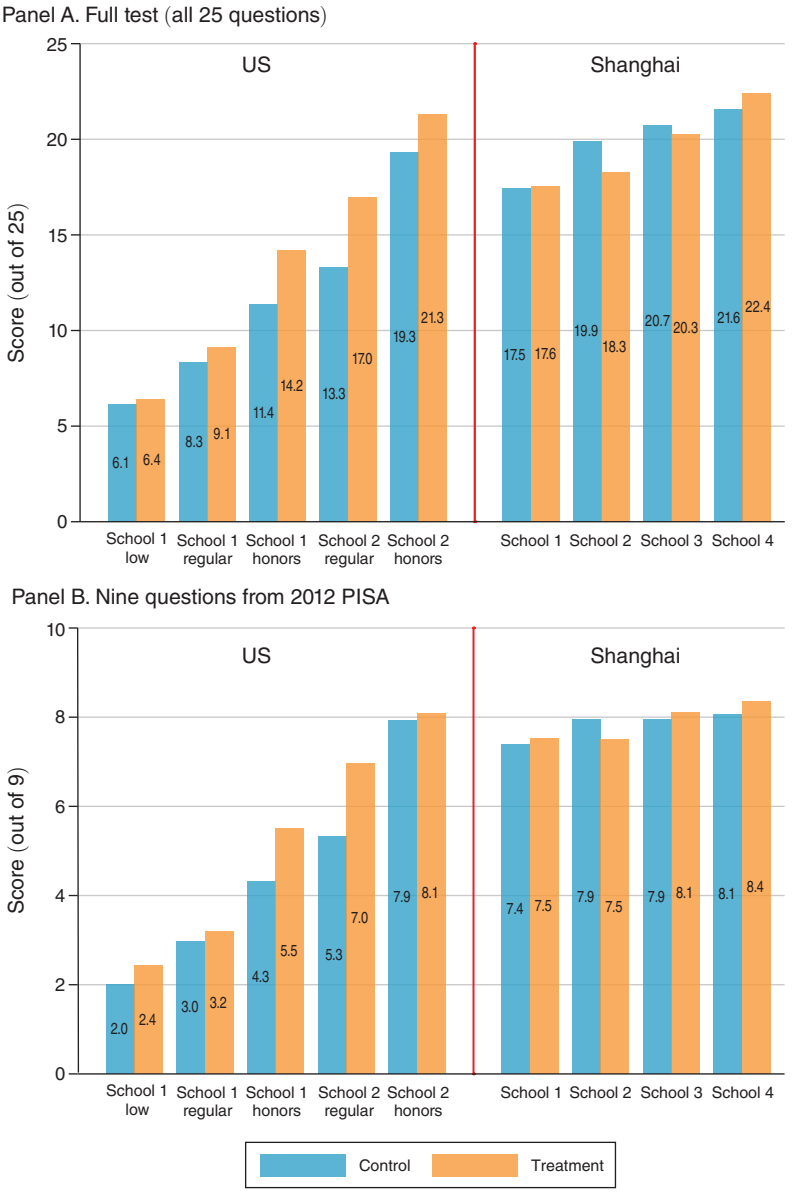

FIGURE 1. AVERAGE TEST SCORE BY GROUP AND TREATMENT: UNITED STATES VERSUS SHANGHAI

Notes: Panel A shows the average score on the full 25 question test for students who received no incentives (Control) and for students who received incentives (Treatment) by school and track. The national average score among US students when these questions were administered as part of official PISA tests is 14.15. We calculate this average using the proportion of US students who answered each question correctly when the questions were administered as part of official PISA exams. The estimated average score of 14.15 is equal to the sum of these proportions over the 25 questions on our exam. This average cannot be calculated for Shanghai because only nine of the questions have been administered there. Panel B shows the average score on the nine questions that have been administered in both the United States and Shanghai. The average score on these nine questions among US students on the official PISA is 5.09. The average score in Shanghai is 7.37.

sample by exam.[14] Standard deviations of the continuous variables (age and standardized baseline exam score) are also displayed. There are no statistically significant differences of means between Treatment and Control at the 10 percent level for any observable characteristics in either the United States or Shanghai samples (standard errors are clustered at the level of randomization). We note that our US sample includes a slightly lower proportion of white students and slightly higher proportion of minority students (Asian, black, and Hispanic) than the national average.

A. *Effects of Incentives on Test Scores*

Figure 1 shows average scores for Control and Treatment by country and school-track. Panel A displays results for the full 25 question test. Panel B shows results for the subset of nine test questions administered on the 2012 PISA, which is when Shanghai first participated in the PISA. We report these scores separately in order to compare performance in our Shanghai sample to national averages on the PISA.[15]

Panel A reveals several striking findings. First, US student performance varies widely by school-track: average scores on the full test without incentives range from 6.1 in the lowest performing group to 19.3 in the highest performing group. Second, the effect of incentives is positive for every group of US students across a wide range of ability levels. The effects are largest for school-tracks in the middle of the ability distribution, which score near the US national average of 14.15. Third, among Shanghai students, we see only small differences between Treatment and Control with no consistent direction of effects.[16] In contrast to the results from the US sample, we find no evidence of treatment effects among students who score near the Shanghai national average of 7.37 (see Shanghai School 1 in panel B). As shown in online Appendix Figure A.3, the financial incentives shift the entire distribution of US test scores to the right, including in areas of common support with Shanghai. By contrast, in Shanghai, the Control and Treatment group distributions largely overlap.

In Table 2, we estimate the effects of extrinsic incentives on test scores in the United States and Shanghai by OLS, estimating the following equation separately in each country:

$$Y_{icsw} = \alpha + \beta_1 Z_c + \beta_2 X_i + \mu_s + \gamma_w + \epsilon_{icsw}, \tag{1}$$

where Y_{icsw} is the score (out of 25) achieved on the exam by student i in class c, school-track s, and wave w (Shanghai only); Z_c is an indicator variable for treatment in class c (the level of randomization); X_i is a vector of individual-level student characteristics: age, gender, and in the United States, race/ethnicity (Asian, black,

[14] The standardized tests for which we have data are the seventh grade Massachusetts Comprehensive Assessment System mathematics assessment for US school 1, the mathematics Secondary School Admissions Test for US school 2, and the 2015 Shanghai Senior High School Entrance Examination for all Shanghai schools.

[15] We calculate national averages using PISA data from the OECD, which provides individual-level responses. We calculate the percentage of test takers who answered each question correctly (weighting each response using the weight variable provided by the OECD to generate nationally representative results) and sum these percentages over the full 25 (or 9) questions. Sixteen of the questions on our test were administered on the PISA prior to 2012 and so we cannot calculate the Shanghai national average for the full test.

[16] We note that the largest positive effect in Shanghai is in the highest performing school, which suggests the results in Shanghai are not due to ceiling effects.

TABLE 2—EFFECTS OF INCENTIVES ON TEST SCORES, BY COUNTRY

	United States		Shanghai		US = Shanghai
	(1)	(2)	(3)	(4)	*p*-value
Treatment	1.59	1.34	−0.26	−0.28	0.0008
SE	(0.40)	(0.34)	(0.27)	(0.26)	
[*p*-value]	[0.002]	[0.001]	[0.379]	[0.353]	
Control mean	10.22		20.50		
SD	(5.64)		(2.95)		
Baseline characteristics	No	Yes	No	Yes	
Standardized effect size	0.28	0.24	−0.09	−0.09	
Students	447	447	656	656	
Clusters	133	133	384	384	

Notes: OLS estimates of equation (1). Robust standard errors clustered by class (except US school 2 and Shanghai schools visited in 2018, which were randomized at the individual level) in parentheses. *p*-values in brackets. Inference in each column is based on a randomization test using the procedure of Young (2019). The dependent variable is the student's score on the full 25 question test. All regressions control for school-track (United States) or school (Shanghai) fixed effects and a wave fixed effect (Shanghai only). Columns 2 and 4 add controls for race/ethnicity (US only), gender, and age. One observation from column 2 imputes age to be the average age in the US sample because age is not recorded for that student. The final column tests whether the treatment effect is equal in the United States and Shanghai. To conduct this test, we pool the US and Shanghai samples and estimate an OLS regression on test score that controls for a treatment assignment indicator, a US indicator, and their interaction. School-track fixed effects, a wave fixed effect, and all student characteristics are also controlled for, with standard errors clustered by the level of randomization. We then conduct a randomization test of the null hypothesis that the effect of the interaction term is zero using the procedure of Young (2019). Effect sizes are standardized using the full sample.

Hispanic white, Hispanic non-white, white, and other); μ_s is a vector of school-track fixed effects; γ_w is a fixed effect in Shanghai for the wave of the experiment (2016, 2018); and ϵ_{icsw} is an error term.[17]

All regressions in Table 2 control for school-track (United States) or school (Shanghai) fixed effects and a wave fixed effect (Shanghai only). Columns 2 and 4 add controls for student characteristics.[18] We report standard errors clustered by the level of randomization in parentheses. All statistical inference is based on randomization tests. The *p*-values from these tests are reported in brackets.[19] The final column reports the *p*-value from a test of equality between the treatment effects in the United States and Shanghai, which we also calculate using a randomization test.[20]

[17] In the higher performing US school and in the 2018 Shanghai sessions, we randomized at the individual level and so $i = c$ for those students.

[18] In the United States, we exclude students who are not in tenth grade and students who are English Language Learners (ELL). Including these students does not affect the results. For one US student missing age, we impute age to be the average age in the US sample. Excluding this observation does not affect the results. Finally, the results are robust to including controls for baseline student standardized exam score rather than school-track fixed effects. See online Appendix Table A.1 for results.

[19] See Young (2019) for an explanation of how these tests are conducted. Each randomization test re-randomizes the allocation of treatment 10,000 times.

[20] To conduct this test, we pool the US and Shanghai samples and estimate an OLS regression on test score that controls for a treatment assignment indicator, a US indicator, and their interaction. School-track fixed effects, a wave fixed effect and all student characteristics are also controlled for, with standard errors clustered by the level of randomization. We then conduct a randomization test of the null hypothesis that the effect of the interaction term is zero.

In response to incentives, the performance of Shanghai students does not change while the scores of US students increase substantially. The estimated treatment effect in the United States is an increase of 1.34 to 1.59 questions ($p < 0.01$), which is equivalent to an effect size of approximately 0.24 to 0.28 standard deviations (we calculate standard deviations using the full sample). In contrast, the estimated effects of incentives in Shanghai are small in magnitude (-0.26 to -0.28 questions, or -0.09 standard deviations) and not statistically significant. The treatment effects in the United States and Shanghai are significantly different at the 1 percent level. These results are consistent with our hypothesis that US students are more responsive than Shanghai students to incentives for effort because they are less motivated at baseline.[21]

B. *Effects of Incentives on Proxies for Effort*

We next study test-taking behavior to support our interpretation that the improvement in test scores is due to increased effort. We examine three proxies for effort, which we discuss in more detail below: questions attempted, proportion of attempted questions answered correctly, and proportion of questions correct. We estimate the effect of incentives for the full test, as well as separately for the first half of the test (questions 1 to 13) and the second half of the test (questions 14 to 25). This analysis builds on prior work, which argues that declining performance over the course of the test is indicative of declining effort (Borghans and Schils 2013, Zamarro, Hitt, and Mendez 2016).

In Table 3, we report regression results for effort proxies, using the following equation estimated by OLS:

$$Y_{qicsw} = \alpha + \beta_1 Z_c + \beta_2 X_i + Q_q + \mu_s + \gamma_w + \epsilon_{qicsw}, \tag{2}$$

where Y_{qicsw} is the question q outcome for student i in class c, school s, and wave w (Shanghai only); Q_q is a vector of question fixed effects; ϵ_{qicsw} is an error term, and the other variables are as previously defined. For each country, the first column (column 1 for the United States, column 4 for Shanghai) reports the results using responses to all 25 questions. The next two columns (columns 2 and 3 for the United States, columns 5 and 6 for Shanghai) split the sample by question number: 1 to 13 and 14 to 25.

For the pooled samples in each country (columns 1 and 4), the reported p-values in brackets are calculated using randomization tests. For the subsamples split by question order, we report p-values adjusted for multiple hypothesis testing within

[21] One potential concern with the null result in Shanghai is that financial incentives might not increase Shanghai students' motivation to put forth effort. To investigate this, we tested the impact of incentives on an effort task in which subjects alternately press the "a" and "b" buttons on their keyboards (see, e.g., Ariely et al. 2009, DellaVigna and Pope 2018). The sample included 194 students at the three high schools in the 2018 Shanghai wave (these students did not participate in the main experiment). Students performed the task for ten minutes, scoring one point for each alternate press. After completing a practice round, the Treatment group ($n = 98$) received RMB 1.8 for every 100 points scored; the Control group ($n = 96$) did not receive incentives. Financial incentives increased performance by an estimated 724 points ($p < 0.01$), a 32 percent increase compared to average performance in the practice round. These results suggest that Shanghai students are responsive to financial incentives.

Table 3—Treatment Effects on Questions Attempted and Questions Correct

	United States			Shanghai		
	All questions (1)	Q 1–13 (13 questions) (2)	Q 14–25 (12 questions) (3)	All questions (4)	Q 1–13 (13 questions) (5)	Q 14–25 (12 questions) (6)
Panel A. Questions attempted						
Treatment	0.037	−0.022	0.102	−0.030	−0.005	−0.057
SE	(0.017)	(0.016)	(0.028)	(0.008)	(0.002)	(0.017)
[*p*-value]	[0.060]	[0.324]	[0.023]	[0.002]	[0.022]	[0.017]
Control mean	0.807	0.962	0.640	0.970	0.998	0.940
SD	(0.394)	(0.191)	(0.480)	(0.170)	(0.046)	(0.238)
Observations	11,175	5,811	5,364	16,400	8,528	7,872
Clusters	133	133	133	384	384	384
Panel B. Proportion of attempted questions correct						
Treatment	0.038	0.041	0.035	0.012	−0.002	0.029
SE	(0.012)	(0.013)	(0.019)	(0.007)	(0.008)	(0.009)
[*p*-value]	[0.004]	[0.017]	[0.119]	[0.119]	[0.801]	[0.012]
Control mean	0.515	0.494	0.549	0.852	0.856	0.848
SD	(0.500)	(0.500)	(0.498)	(0.355)	(0.351)	(0.359)
Observations	9,276	5,544	3,732	15,667	8,490	7,177
Clusters	133	133	130	384	384	380
Panel C. Proportion of questions correct						
Treatment	0.053	0.030	0.079	−0.013	−0.007	−0.020
SE	(0.013)	(0.015)	(0.019)	(0.011)	(0.008)	(0.017)
[*p*-value]	[0.001]	[0.086]	[0.002]	[0.269]	[0.475]	[0.475]
Control mean	0.416	0.475	0.351	0.827	0.854	0.797
SD	(0.493)	(0.499)	(0.477)	(0.379)	(0.353)	(0.402)
Observations	11,175	5,811	5,364	16,400	8,528	7,872
Clusters	133	133	133	384	384	384

Notes: OLS estimates of equation (2). Robust standard errors clustered by class (except US school 2 and 2018 Shanghai wave, which were randomized at the individual level) in parentheses. *p*-values in brackets. Inference is based on a randomization test using the procedure of Young (2019) in columns 1 and 4 and is adjusted for multiple hypothesis testing of estimates from two subsamples by controlling the family-wise error rate using the free step-down resampling methodology of Westfall and Young (1993) in columns 2–3 and columns 4–5. This adjustment is done within each panel over the two columns. All columns include school-track fixed effects, question fixed effects, and the following covariates: age, gender, race/ethnicity in the United States, and a wave fixed effect in Shanghai.

each country using the Westfall and Young (1993) free step-down resampling method to control the family-wise error rate. This adjustment is done within each panel over the two columns (columns 2 and 3 for the United States; columns 5 and 6 for Shanghai).

We first estimate the effect of incentives in the United States on questions attempted. There is no penalty for wrong answers so a student who cares about performing well should attempt to answer every question. As shown in column 1 of panel A, incentives increase the overall probability that a US student answers a question by about 4 percentage points. The average impact is driven entirely by treatment effects on the second half of the test where response rates increase by an estimated 10 percentage points (column 3). The impact of incentives helps offset the

dramatic decline in response rates among the Control group, which drop from 96 percent in the first half of the test to 64 percent in the second half.[22]

In Table 3, panel B (columns 1 through 3), we estimate the effects of incentives in the United States on the percentage of attempted questions answered correctly. If incentives primarily increase guessing, then students may attempt more questions but may be less likely to answer those questions correctly; whereas, if students are truly thinking harder about each question, we would expect that they answer a higher share correctly (Jacob 2005 provides discussion). We find that incentives increase the share of attempted questions answered correctly by US students. The estimated effects of about 4 percentage points are similar across question order. These results suggest that the increased response rates shown in panel A are not just due to guessing but rather increased effort to answer questions correctly.

Finally, in panel C (columns 1 through 3), we estimate how the effects of incentives in the United States on both response rates and share correct translate to improvement in test scores. Incentives improve correct answer rates by about 5 percentage points, with estimated effects increasing from 3 percentage points in the first half of the test to 8 percentage points in the second half. Together our results suggest that US students are not at their effort or output frontier at baseline, and that increasing student motivation has a significant impact on performance, particularly toward the end of the test.

In Shanghai, there is little impact of treatment on the first half of the test (column 5 of each panel). On the second half (column 6 of each panel), Shanghai students attempt fewer questions (panel A) but are more likely to answer correctly those that they do attempt (panel B). The net effect on correct answers is small and not statistically significant (panel C). One possible explanation for these results is that in response to treatment, Shanghai students reallocate effort by answering fewer questions but putting more effort into the ones they do answer, such that average performance remains unchanged. Taken together, the findings are consistent with students in Shanghai having little scope to meaningfully increase their overall effort.

C. *Heterogeneity*

We now turn to an examination of treatment effects by ability, as measured by predicted test score. To calculate each student's predicted score, we regress baseline standardized exam score, age, gender, and (in the United States) race/ethnicity on test score in the Control group, separately by school.[23] We then use the estimated coefficients from the relevant regression to predict each student's test score. Panel A of Figure 2 plots predicted score against actual score for each US student. The Treatment and Control lines are estimated by performing a kernel-weighted local

[22] Online Appendix Figure A.4 plots response rates by question, treatment group, and country. The declines in US performance over the course of the test are similar to those found among US students on the PISA (Borghans and Schils 2013, Zamarro, Hitt, and Mendez 2016).

[23] In the United States, each school uses a different baseline standardized exam. We impute missing baseline exam scores to be the school mean and include an indicator for imputed score. Baseline exam scores are available for all students in the Shanghai sample.

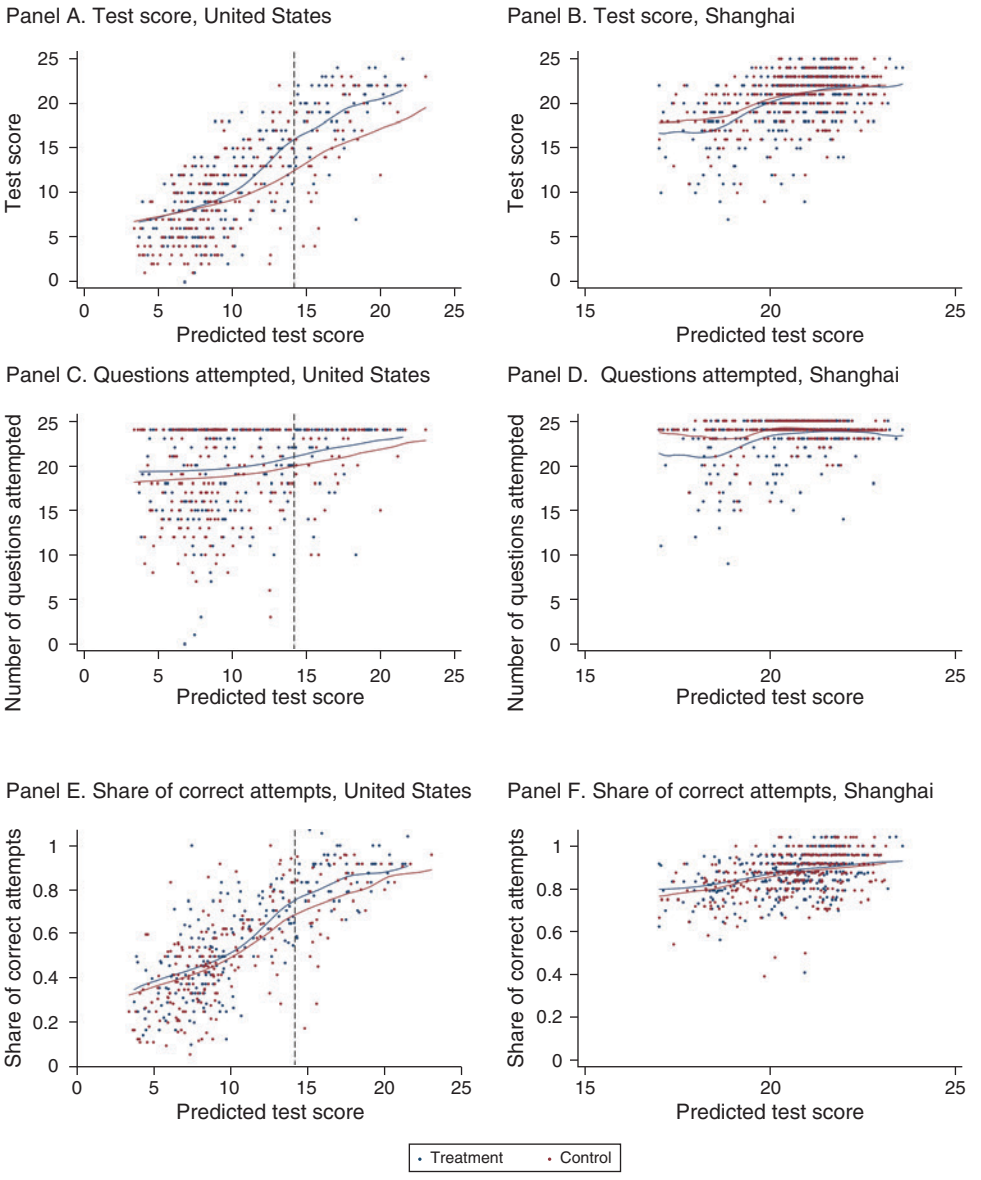

FIGURE 2. TREATMENT EFFECTS BY PREDICTED SCORE

Notes: For the United States, we predict score using age, gender, race/ethnicity, and baseline exam score in the US Control group. The vertical line at 14.15 in the US panels is the US national average. We calculate this average using the proportion of US students who answered each question correctly when the questions were administered as part of official PISA exams. The estimated average score of 14.15 is equal to the sum of these proportions over the 25 questions on our exam. For Shanghai, we predict score using age, gender, baseline exam score, and a wave fixed effect in the Shanghai control group. This average cannot be calculated for Shanghai because only nine of the questions have been administered there. In both cases, we estimate the Control and Treatment lines using kernel weighting.

polynomial regression. The vertical line at 14.15 is the average US performance on the same test questions when administered as part of the PISA.[24]

As shown in Figure 2, panel A, extrinsic incentives have the largest impact among students whose predicted scores are close to average US performance. Our sample also includes students with predicted scores far below the US average. For these students, the incentives have little impact on performance, possibly because they simply do not understand the material, and incentives cannot change that fact. In contrast, incentives do have a large impact on students who are able to answer the questions but do not invest effort at baseline to do so.

Panels C and E of Figure 2 plot predicted baseline score against questions attempted and proportion of attempted questions correct, respectively. Compared to the impact on test scores (panel A), the treatment effect on attempted questions is more constant across predicted score (panel C); while the treatment effect on proportion correct (panel E) follows the same pattern as test scores. The figures are consistent with threshold regressions reported in online Appendix Table A.2, which detect a split at a predicted test score of 10.15 or 10.66 when the dependent variable is test score or proportion correct, respectively, but no split when the dependent variable is questions attempted. These results suggest that the incentives motivate students of all ability levels to try harder on the test (i.e., attempt more questions), but that increased effort only translates into higher scores for students who are able to answer the questions correctly.

Panels B, D, and E of Figure 2 show the same analysis for Shanghai students. Throughout most of the Shanghai ability distribution, there is little difference between Treatment and Control on any measure. We find suggestive evidence that lower ability students attempt fewer questions in response to treatment, which may reduce their scores, but note that this is based on a small number of data points in the left tail of our ability distribution.[25]

IV. Conclusion

Our goal in this article is to highlight that low-stakes assessments may not measure and compare ability in isolation, and as such differences across countries may not solely reflect differences in ability across students. If correct, the conclusions drawn from such assessments should be more modest than current practice. Note that this paper is not about the importance of intrinsic motivation in learning, or the impact of incentives to invest more effort in preparing for the test or studying in general.[26] Rather we are focusing on between-country differences in effort on the test itself. In this manner, we show that policy reforms that ignore the role of

[24] There is no vertical line indicating the Shanghai national average because, as noted above, we cannot calculate the Shanghai average for the full test. As shown in Table 1, panel B, and Figure 1, panel B, our sample mainly consists of students who score above the national average on the subset of questions administered on official PISA exams.

[25] We also examine treatment effects by gender (online Appendix Table A.3). Incentives significantly increase both male and female scores in the United States, with larger point estimates among boys (the differences by gender are not statistically significant). In Shanghai, the treatment effects for both male and female students are small and not statistically significant.

[26] Similarly, our results may not generalize to high-stakes tests, such as end of the year final exams, high school exit exams or college entrance exams, on which students have large extrinsic incentives to work hard and perform well.

intrinsic motivation to perform well on the test may be misguided and have unintended consequences.

We regard this study as a starting point. Our field experiment provides a methodology for estimating the causal effect of differential effort levels on test performance, but we implement this approach with samples that are not nationally representative, and include students from only two US high schools and four Shanghai high schools. The results from our experimental samples suggest that motivation may be an important confound in international comparisons, and we hope future work will employ this methodology using nationally representative samples in many countries. This would make it possible to better quantify how international rankings might change if differences in motivation and test taking effort across countries are taken into account.[27]

Should our results replicate when our method is applied more broadly, the findings may also shed light on two puzzles in the literature regarding the correlation between performance on low-stakes assessments and economic outcomes. In the United States, low-stakes test performance is highly correlated with individual income, but explains little of the variation across individuals (Murnane et al. 2000). Relatedly, while low-stakes test performance is highly correlated with economic growth across countries, the United States is an outlier, with higher economic growth than its test scores predict (Hanushek and Woessmann 2011). Differences in test-taking effort across students and across cultures may add explanatory power to these analyses and better inform our understanding of the relationship among ability, intrinsic motivation, and long-term outcomes (e.g., Borghans and Schils 2013; Balart, Oosterveen, and Webbink 2017; Segal 2012). Future work could also explore how best to interpret or perhaps even redesign low-stakes assessment tests so that policymakers can use the results to allocate resources in a more efficient and productive manner. Finally, we hope that our findings serve as a catalyst to explore their relevance in different domains, such as black-white or male-female performance gaps. This can not only deepen our understanding of test score differences across groups in society, but also lead to a new discussion revolving around why such differences persist.

REFERENCES

Akyol, Ş. Pelin, Kala Krishna, and Jinwen Wang. 2018. "Taking PISA Seriously: How Accurate are Low Stakes Exams?" National Bureau of Economic Research Working Paper 24930.

Aoki, Kumiko. 2008. "Confucius vs. Socrates: The Impact of Educational Traditions of East and West in a Global Age." *International Journal of Learning* 14 (11): 35–40.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar. 2009. "Large Stakes and Big Mistakes." *Review of Economic Studies* 76 (2): 451–69.

Balart, Pau, Matthijs Oosterveen, and Dinand Webbink. 2017. "Test Scores, Noncognitive Skills and Economic Growth." *Economics of Education Review* 63: 134–53.

Barrow, Lisa, and Cecilia Elena Rouse. 2018. "Financial Incentives and Educational Investment: The Impact of Performance-Based Scholarships on Student Time Use." *Education Finance and Policy* 13 (4): 419–48.

[27] In Gneezy et al. (2017), we provide a back-of-the-envelope calculation that suggests that if our treatment effects carried over to the PISA, increasing student effort on the test itself would improve US mathematics performance by 22 to 24 points, equivalent to moving the United States from thirty-sixth to nineteenth in the 2012 international mathematics rankings. While this gives a sense of magnitudes, we note that it is based on sample calculations for a non-representative sample, and holds constant the effort level exerted in all other countries.

Borghans, Lex, and Trudie Schils. 2013. "The Leaning Tower of Pisa: Decomposing Achievement Test Scores into Cognitive and Noncognitive Components." https://www.sole-jole.org/13260.pdf (accessed May 1, 2018).

Borgonovi, Francesca, and Przemyslaw Biecek. 2016. "An International Comparison of Students' Ability to Endure Fatigue and Maintain Motivation during a Low-Stakes Test." *Learning and Individual Differences* 49: 128–37.

Braun, Henry, Irwin Kirsch, and Kentaro Yamamoto. 2011. "An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment." *Teachers College Record* 113 (11): 2309–44.

Carnoy, Martin, and Richard Rothstein. 2013. *What Do International Tests Really Show about U.S. Student Performance?* Washington, DC: Economic Policy Institute.

DellaVigna, Stefano, and Devin Pope. 2018. "What Motivates Effort? Evidence and Expert Forecasts." *Review of Economic Studies* 85 (2): 1029–69.

DeMars, Christine E., and Steven L. Wise. 2010. "Can Differential Rapid-Guessing Behavior Lead to Differential Item Functioning?" *International Journal of Testing* 10 (3): 207–29.

Duckworth, Angela Lee, Patrick D. Quinn, Donald R. Lynam, Rolf Loeber, and Magda Stouthamer-Loeber. 2011. "Role of Test Motivation in Intelligence Testing." *Proceedings of the National Academy of Sciences* 108 (19): 7716–20.

Eklöf, Hanna. 2010. "Skill and Will: Test-Taking Motivation and Assessment Quality." *Assessment in Education: Principles, Policy, & Practice* 17 (4): 345–56.

Finn, Bridgid. 2015. "Measuring Motivation in Low-Stakes Assessments." Educational Testing Service Research Report ETS RR-15-19.

Fryer, Roland G., Jr. 2011. "Financial Incentives and Student Achievement: Evidence from Randomized Trials." *Quarterly Journal of Economics* 126 (4): 1755–98.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu. 2019. "Measuring Success in Education: The Role of Effort on the Test Itself: Dataset." *AER: Insights*. https://doi.org/10.1257/aeri.20180633.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Sally Sadoff, Xiangdong Qin, and Yang Xu. 2017. "Measuring Success in Education: The Role of Effort on the Test Itself." National Bureau of Economic Research Working Paper 24004.

Gneezy, Uri, Stephan Meier, and Pedro Rey-Biel. 2011. "When and Why Incentives (Don't) Work to Modify Behavior." *Journal of Economic Perspectives* 25 (4): 191–210.

Grek, Sotiria. 2009. "Governing by Numbers: The PISA 'Effect' in Europe." *Journal of Education Policy* 24 (1): 23–37.

Hanushek, Eric A., and Ludger Woessmann. 2011. "How Much Do Educational Outcomes Matter in OECD Countries?" *Economic Policy* 26 (67): 427–91.

Hau, Kit-Tai, and Irene T. Ho. 2010 "Chinese Students' Motivation and Achievement." In *Oxford Handbook of Chinese Psychology*, edited by Michael Harris Bond, 187–204. Oxford, UK: Oxford University Press.

Hess, Robert D., Chih-mei Chang, and Teresa M. McDevitt. 1987. "Cultural Variations in Family Beliefs about Children's Performance in Mathematics: Comparisons among Peoples' Republic of China, Chinese American, and Caucasian-American Families." *Journal of Educational Psychology* 79 (2): 179–88.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89 (5-6): 761–96.

Jalava, Nina, Juanna Schrøter Joensen, and Elin Pellas. 2015. "Grades and Rank: Impacts of Non-Financial Incentives on Test Performance." *Journal of Economic Behavior & Organization* 115: 161–96.

Levitt, Steven D., John A. List, and Sally Sadoff. 2016. "The Effect of Performance-Based Incentives on Educational Achievement: Evidence from a Randomized Experiment." National Bureau of Economic Research Working Paper 22107.

Levitt, Steven D., John A. List, Susanne Neckermann, and Sally Sadoff. 2016. "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance." *American Economic Journal: Economic Policy* 8 (4): 183–219.

Murnane, Richard J., John B. Willett, Yves Duhaldeborde, and John H. Tyler. 2000. "How Important are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?" *Journal of Policy Analysis and Management* 19 (4): 547–68.

National Center for Education Statistics (NCES). 2015. "The Nation's Report Card." https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2015136.

Organisation for Economic Co-operation and Development (OECD). 2014. *PISA 2012 Results: What Students Know and Can Do: Student Performance in Mathematics, Reading and Science*, Volume I, Revised Edition. Paris: OECD Publishing.

Segal, Carmit. 2012. "Working When No One is Watching: Motivation, Test Scores, and Economic Success." *Management Science* 58 (8): 1438–57.

Sievertsen, Hans Henrik, Francesca Gino, and Marco Piovesan. 2016. "Cognitive Fatigue Influences Students' Performance on Standardized Tests." *Proceedings of the National Academy of Sciences* 113 (10): 2621–24.

Stevenson, Harold W., Shin-Ying Lee, Chuansheng Chen, James W. Stigler, Chen-Chin Hsu, Seiro Kitamura, and Giyoo Hatano. 1990. "Contexts of Achievement: A Study of American, Chinese, and Japanese Children." *Monograph of the Society for Research in Child Development* 55 (1-2): 221.

Stevenson, Harold W., and James W. Stigler. 1992. *The Learning Gap: Why Our Schools Are Failing and What We Can Learn from Japanese and Chinese Education*. New York: Summit Books.

Tao, Vivienne Y. K. 2016. "Understanding Chinese Students' Achievement Patterns: Perspectives from Social-Oriented Achievement Motivation." In *The Psychology of Asian Learners*, edited by Ronnel B. King and Allan B. I. Bernardo, 621–34. Singapore: Springer.

Westfall, Peter H., and S. Stanley Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: John Wiley and Sons.

Wise, Steven L., and Christine E. DeMars. 2005. "Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions." *Educational Assessment* 10 (1): 1–17.

Woessmann, Ludger. 2016. "The Importance of School Systems: Evidence from International Differences in Student Achievement." *Journal of Economic Perspectives* 30 (3): 3–32.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *Quarterly Journal of Economics* 134 (2): 557–98.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez. 2016. "When Students Don't Care: Reexamining International Differences in Achievement and Non-Cognitive Skills." EDRE Working Paper 2016-18.