NBER WORKING PAPER SERIES

MEASURING SUCCESS IN EDUCATION: THE ROLE OF EFFORT ON THE TEST ITSELF

Uri Gneezy
John A. List
Jeffrey A. Livingston
Sally Sadoff
Xiangdong Qin
Yang Xu

Working Paper 24004
http://www.nber.org/papers/w24004

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2017

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. Please direct correspondence to Sally Sadoff: ssadoff@ucsd.edu. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring Success in Education: The Role of Effort on the Test Itself
Uri Gneezy, John A. List, Jeffrey A. Livingston, Sally Sadoff, Xiangdong Qin, and Yang Xu
NBER Working Paper No. 24004
November 2017
JEL No. C93,I24

ABSTRACT

Tests measuring and comparing educational achievement are an important policy tool. We experimentally show that offering students extrinsic incentives to put forth effort on such achievement tests has differential effects across cultures. Offering incentives to U.S. students, who generally perform poorly on assessments, improved performance substantially. In contrast, Shanghai students, who are top performers on assessments, were not affected by incentives. Our findings suggest that in the absence of extrinsic incentives, ranking countries based on low-stakes assessments is problematic because test scores reflect differences in intrinsic motivation to perform well on the test itself, and not just differences in ability.

Uri Gneezy
Rady School of Management
University of California - San Diego
Otterson Hall, Room 4S136
9500 Gilman Drive #0553
La Jolla, CA 92093-0553
ugneezy@ucsd.edu

John A. List
Department of Economics
University of Chicago
1126 East 59th
Chicago, IL 60637
and NBER
jlist@uchicago.edu

Jeffrey A. Livingston
Bentley University
Department of Economics
175 Forest Street
Waltham, MA 02474
jlivingston@bentley.edu

Sally Sadoff
Rady School of Management
University of California at San Diego
Wells Fargo Hall, Room 4W121
9500 Gilman Drive #0553
La Jolla, CA 92093-0553
ssadoff@ucsd.edu

Xiangdong Qin
Antai College of Economics & Management
Department of Applied Economics
xdqin@sjtu.edu.cn

Yang Xu
HSBC Business School
Peking University
University Town, Nanshan District
Shenzhen, Guangdong 518055
China
yangxu.econ@gmail.com

1 Introduction

It is difficult to overestimate 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 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 (NCES, 2015). Similarly, on the 2012 Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Co-operation and Development (OECD), among the 65 countries and economies that participated, U.S. high school students ranked 36$^{\text{th}}$ for mathematics performance with scores declining since 2009 (OECD, 2014).[1]

In response to poor U.S. performance on such assessments, U.S. 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."[2] Student performance on international assessments has also had a

[1]U.S. rankings on international assessments varies across assessment and subject. Compared to their ranking on the PISA in mathematics, U.S. students rank relatively better on PISA in reading and science, as well as on other international assessments including the Trends in International Mathematics and Science Study and the Progress in International Reading Literacy Study (International Association for the Evaluation of Educational Achievement (IEA), `http://timssandpirls.bc.edu/isc/publications.html`).

[2]See S. Dillon, Top test scores from Shanghai stun educators. *The New York Times* (2010;

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 U.S. 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 this study, we offer an additional potential reason: students in different countries may have heterogeneous levels of intrinsic motivation to perform well on assessment tests. If so, poor U.S. 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.

To examine whether there are differences in motivation across cultures, we conduct an experiment in the U.S. 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 a surprise financial incentive to put forth effort on the test and compare their performance to students who are not given an

`http://www.nytimes.com/2010/12/07/education/07education.html`).

incentive. Importantly, students learn about the incentives 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 those who are deeply motivated to work hard at baseline, we expect incentives to have little impact on performance since students are already at or near their output frontier. In contrast, among students who lack motivation at baseline, extrinsic financial incentives have scope to increase effort and improve performance. Moving less intrinsically motivated students closer to their output frontier will result in a better measurement of relative ability across students.

Our results are consistent with this hypothesis. In response to incentives, performance among Shanghai students does not change while the scores of U.S. students increase substantially. Under incentives, U.S. students attempt more questions (particularly towards the end of the test) and are more likely to answer those questions correctly. The resulting effects on test scores are concentrated among students whose baseline performance is near the U.S. average. Finally, we simulate the impact on U.S. performance were our treatment effects to carry over to the actual PISA. We estimate that increasing student effort on the test itself would improve U.S. mathematics performance by $22 - 24$ points, equivalent to moving the U.S. from 36th to 19th in the 2012 international mathematics rankings.

The remainder of our paper proceeds as follows. Section 2 reviews the relevant literature. Section 3 describes the design of the experiment. Section 4 presents the

results. Section 5 concludes.

2 Background Literature

The finding that scores on low stakes tests do not always reflect the true ability of students has already been recognized in the literature (Wise and DeMars, 2005 and Finn, 2015 provide recent reviews). One strand of research has examined 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; Zamarro et al., 2016). Yet, important for our purposes, these studies are not able to identify the impact of effort separately from the impact of ability. For example, self-reported effort and rapid guessing may indicate that the student does not try hard because 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 et al., 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. These studies demonstrate that rewards (both financial and non-financial) as well as how the test is framed can increase effort and improve performance on the test (Duckworth et al., 2011; Braun et al., 2011; Levitt et al., 2016; Jalava et al., 2015).

Recent work in education and behavioral economics has investigated how to best structure incentives (Gneezy et al., 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 framed the incentives in our experiment as losses provided in the form of upfront cash rewards, which increases their salience and credibility.

We wish to emphasize that the goal of the current 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 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 U.S. (Zamarro et al., 2016). And there is evidence from descriptive studies that compared to the U.S., East Asian parents, teachers and students put more emphasis on diligence and effort (Stevenson and Stigler, 1992; Stevenson et al., 1990; Hess et al., 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 one's 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.

3 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 36$^{\text{th}}$ 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.[3]

In our experiment, which was conducted in the spring and fall of 2016, 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.[4] The exam consists of 13 multiple-choice questions and 12 free answer fill-in-the-blank questions (see Appendix B for screenshots of 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.

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

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

Figure 1 displays the worldwide percentage of students who answered each question correctly when the questions were administered as part of official PISA exams. We calculate the percentage correct using individual level data available from the OECD.[5] The percentage correct ranges from 25.7 to 87.27. As the figure illustrates, there is little correlation between question difficulty and question order on the test ($\rho = 0.14$). The test was administered on computers so that the results could be available immediately after students completed the test. U.S. 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 three high schools in Shanghai. While our samples are not nationally representative, we aimed to sample students throughout their respective distributions. The U.S. 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 one school with performance that is far above average. In the U.S., all students in tenth grade math classes were selected to participate.[6] In Shanghai, we randomly selected approximately 25 percent of tenth grade classes in each school to participate.

We randomly assigned students to either the Control (no incentives) group or the Treatment (incentives) group. The U.S. sample includes 447 students (227 in control

[5]Individual level responses to every question given on each iteration of the PISA by every participant are available at `http://www.oecd.org/pisa/data`.

[6]In the lower performing school, 81 percent of tenth graders were enrolled in tenth grade math. The remainder were enrolled in 9^{th} grade (18 percent) math or 11^{th} grade (1 percent) math. The tenth-grade math classes also included 89 non-tenth graders who are excluded from our primary analysis.

and 220 in treatment) and the Shanghai sample includes 280 students (141 in control and 139 in treatment). Students in the Control group received no incentive for their performance on the test. In the incentive treatment, U.S. 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 were paid in Renminbi. We used the Big Mac Index to determine currency conversion.[7] The implied exchange rate in January 2016 was 3.57. By this index $25 converts to 89.25RMB. We rounded up and gave students in the treatment group 90RMB and took away 3.6RMB for each incorrect answer.

Importantly, before the experiment began students were unaware of the purpose of the experiment or that financial incentives would be available. All they were told beforehand was that they would be participating in an experiment on decision-making, with students assigned to treatment and control receiving identical information. Immediately before they took the test, students read the instructions along with the experiment administrator and in the treatment group were informed of the incentives. Accordingly, we are assured that the incentives only influence effort on the test itself, not preparation for the exam. This design also limits the role of selective attendance on the day of testing. All students present on the day of testing took part in the experiment.[8]

[7]The index was obtained from `http://www.economist.com/content/big-mac-index`.

[8]In the higher performing U.S. school, eleven students arrived late due to a prior class and did not participate.

We randomized at the class level (except in the higher performing U.S. school where we randomized at the individual level). In the U.S., we stratified by school and re-randomized to achieve balance on the following baseline characteristics: gender, race/ethnicity and mathematics class level/track (low, regular, honors). 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 Shanghai, the randomization was stratified by school (baseline demographics were not available at the time of randomization).

Table 1 presents the results of the randomization. It displays means and standard deviations by treatment group and country for student characteristics (gender, age and, in the U.S., race/ethnicity) and baseline score on a standardized exam (baseline scores were not available at the time of randomization and are missing for 22% of the U.S. sample). As expected, there are no statistically significant differences between Treatment and Control at the 10 percent level for any observable characteristics in either the U.S. or Shanghai samples.

4 Results

4.1 Effects of incentives on test performance

We begin by examining the raw data. Figure 2 shows average scores for Control and Treatment by country and school-track. The figure reveals several striking findings. First, U.S. student performance varies widely by school-track: average scores 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 U.S. students across a wide range of ability levels, with larger treatment effects among higher performing students. Third, we see only small differences with no consistent direction of effects among Shanghai students.[9]

As shown in Figure 3, the financial incentives impact the entire distribution of U.S. test scores. The figure presents the empirical cumulative distribution function (CDF) of scores by treatment group and country. In the U.S., incentives shift the CDF to the right, including in areas of common support with Shanghai. By contrast, in Shanghai, the Control and Treatment group CDFs largely overlap and cross frequently. We conduct non-parametric permutation tests of differences between the test score distributions in each country.[10] We find that treatment significantly shifts the U.S. distribution to the right ($p < 0.01$) with no significant shift in the Shanghai distribution ($p = 0.36$).

In Table 2, we estimate the effects of extrinsic incentives on test scores in the U.S. and Shanghai, by Ordinary Least Squares (OLS) using the following equation:

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

where Y_{ics} is the test score for student i in class c and school-track s; Z_c is an indi-

[9]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.

[10]We construct test statistics using permutation methods based on Schmid and Trede (1995) and run one-sided tests for stochastic dominance and separatedness of the distributions (see also Imas, 2014). The test statistics identify the degree to which one distribution lies to the right of the other, and take into account both the consistency of the differences between the distributions (i.e., how often they cross) and the size of the differences (i.e., the magnitudes). We compute p-values by Monte-Carlo methods with 100,000 repetitions.

cator 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 U.S., race/ethnicity (white non-Hispanic, black non-Hispanic, Asian, white Hispanic, non-white Hispanic, and other); μ_s is a vector of school-track fixed effects; and ϵ_{ics} is an error term.[11] The regressions reported in columns 1 and 3 include school-track fixed effects; columns 2 and 4 add controls for student characteristics.[12] In parentheses, we report p-values calculated via wild bootstrapping to adjust for clustering at the level of randomization and the small number of clusters (Cameron et al., 2008). The final column reports the p-value from a test of equality between the treatment effects in the U.S. and Shanghai, which we calculate using a randomization test (Canay et al., 2017).[13]

In response to incentives, the performance of Shanghai students does not change while the scores of U.S. students increase substantially. The estimated treatment effect in the U.S. is an increase of $1.36 - 1.59$ questions ($p < 0.01$), an effect size of approximately $0.20 - 0.23$ standard deviations (we calculate standard deviations

[11]In the higher performing U.S. school, we randomized at the individual level and so $i = c$ for those students.

[12]In the higher performing U.S. school, we pool six low track students with regular track students. For one U.S. student missing age, we impute age to be the average age in the U.S. sample. Excluding this observation does not affect the results. In the U.S., 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. Finally, the results are robust to including controls for baseline student standardized exam score rather than school-track fixed effects (Appendix Table A1).

[13]To conduct the test, we estimate Equation 1 replacing the treatment indicator with a U.S. indicator, the interaction between the U.S. indicator and a treatment group indicator, and the interaction between a Shanghai indicator and a treatment group indicator. We then save the t-statistic from a test of equality of the two interaction effects. Next, we randomly re-assign each cluster to treatment or control (using the same number of clusters that were actually assigned to each group in each school), run the same regression and estimate the t-statistic from the test of equality of the interaction effects. We repeat this procedure 100,000 times. The estimated p-value is the proportion of t-statistics from these iterations that are larger than the initial t-statistic using the actual assignments to treatment and control.

using the full sample). In contrast, the estimated effects of incentives in Shanghai, are small in magnitude, $0.22 - 0.25$ questions ($0.03 - 0.04$ standard deviations), and not statistically significant. The treatment effects in the U.S. and Shanghai are significantly different at the $p = 0.011$ level.

We next explore test taking behavior among U.S. students in order to support our interpretation that the impact of incentives on test scores is due to increased effort. First, we examine response rates. There is no penalty for wrong answers so a student who cares about performing well should attempt to answer every question. Figure 4 presents response rates for each question, by treatment group and country. In the first half of the test, response rates are high in both the U.S. and Shanghai. In the second half of the test, response rates among U.S. control group students decline dramatically. Under incentives, U.S. response rates increase, particularly towards the end of the test. There is little difference among even the final questions between Treatment and Control in Shanghai.

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

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

where Y_{qics} is the question q outcome for student i in class c and school s; Q_q is a vector of question fixed effects; ϵ_{qics} is an error term, and the other variables are as previously defined.[14] The level of observation is a student's performance on a question, so the full sample of questions and students includes $25 \times 447 =$

[14]Probit estimates yield similar results (available upon request).

11,175 observations. Column 1 reports estimates using responses to all 25 questions. Columns 2 and 3 split the sample by question number: 1-13 and 14-25. We report p-values from standard errors clustered by class and adjusted for multiple hypothesis testing within each split of the data by using the Anderson (2008) free step-down resampling method to control the family-wise error rate.[15]

We first estimate the impact of incentives on questions attempted. We use Equation 2, where the dependent variable equals 1 if the question was attempted and 0 otherwise. As shown in Panel A, incentives increase the overall probability that a student answers a question by about 4 percentage points. Consistent with the pattern seen in Figure 4, this effect is driven entirely by treatment effects on the second half of the test where response rates increase by an estimated 10 percentage points.

In Panel B, we estimate treatment effects on the proportion of attempted questions answered correctly. We use Equation 2, where the dependent variable equals 1 if the question was answered correctly and 0 otherwise and we restrict the sample to only those questions that were attempted. If incentives primarily increase guessing, then students may attempt more questions but 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 significantly increase the share of attempted questions answered correctly. The estimated effects of about 4 percentage points are similar across question order. These results suggest that the increased response

[15]The results are similar if standard errors are obtained by wild bootstrapping. Appendix Table A2 reports the estimates for Shanghai students. In the Shanghai sample, there are too few clusters to implement the resampling method, so the reported p-values are instead estimated via wild bootstrapping.

rates among U.S. students shown in Panel A are not just due to guessing but rather increased effort to answer questions correctly.

Finally, in Panel C we estimate how the effects of incentives on both response rates and share of attempts correct translate to improvement in test scores. We use the same estimating equation as in Panel B but now include all questions whether or not they were attempted. 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 U.S. students are not at their effort or output frontier at baseline, and that increasing student motivation has a significant impact on performance, particularly towards the end of the test.

We now turn to an examination of how treatment effects among U.S. students vary with baseline ability, which we proxy by predicted test score. To calculate each student's predicted score, we regress baseline standardized exam score, age, gender and race/ethnicity on test score in the control group, separately by school.[16] We then use the estimated coefficients from the relevant regression to predict each student's test score. Figure 5A plots predicted score against actual score for each U.S. student. The Treatment and Control lines are estimated using a kernel-weighted local polynomial regression. The vertical line at 14.15 is the average U.S. performance on the same test questions when administered as part of the PISA.[17]

[16]Each school uses a different baseline standardized exam. We impute missing baseline exam scores to be the school mean score and include an indicator for imputed score.

[17]We calculate the U.S. average using the proportion of U.S. students who answered each question correctly when the questions were administered as part of official PISA exams using the individual level data described in Section 3. Our estimated U.S. average score of 14.15 is equal to the sum of these proportions over the 25 questions on our exam. Appendix Figure A1 shows the same analysis

An important observation from our results is that the extrinsic incentives have the largest impact among students whose predicted scores are close to average U.S. performance. Our sample also includes students with predicted scores far below the U.S. 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 an impact for students who are able to answer the questions but do not invest effort at baseline to do so.

Figures 5B and 5C plot predicted test score against questions attempted and proportion of attempted questions correct, respectively. Compared to treatment effects on test score (Figure 5A), the effects on attempted questions are more constant across predicted score (Figure 5B); while the effects on proportion correct (Figure 5C) follow the same pattern as the effects on test score. The figures are consistent with threshold regressions, which detect a split at a predicted score of 11.04 when the dependent variable is test score, no split when the dependent variable is questions attempted and a split at a predicted score of 11.00 when the dependent variable is proportion correct.[18] These results suggest that the incentives motivate students of

for Shanghai students. We are not able to calculate an average score for Shanghai because not all 25 questions on the test have been administered in Shanghai (PISA scores are only reported for Shanghai in 2009 and 2012 because Shanghai began participation in 2009 and was grouped with three other cities in 2015). As in the full sample, there is little difference between Treatment and Control by predicted test score.

[18] Appendix Table A3 reports the threshold regression estimates of the following equation: $Y_{ic} = \sum_{j=1}^{m+1} (\alpha^j + \beta_1^j Z_c + \beta_2^j X_i) I_j(\gamma_{j-1} < PS_{ic} < \gamma_j) + \epsilon_{qics}$, where m is the number of thresholds and $j = 1, ..., m+1$ index the threshold regions; Y_{ic} is the dependent variable reported at the top of the column (test score, number of questions attempted, or proportion of attempted questions correct); PS_{ic} is the predicted test score of student i in class c (the threshold variable); $\gamma_1 < \gamma_2 < ... < \gamma_m$ are ordered thresholds with $\gamma_0 = -\infty$ and $\gamma_{m+1} = \infty$; I_j is an indicator for the jth threshold region; and other variables are as previously defined. The optimal number of thresholds is first estimated by minimizing the Bayesian Information Criterion. When this estimate indicates that at

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.

Finally, in Table 4, we examine treatment effects by gender in each country. All regressions include school fixed effects and control for student characteristics.[19] In line with the literature on gender differences in reaction to incentives (Levitt et al. 2016; Azmat et al. 2016; Attali et al. 2011; see Croson and Gneezy 2009 for a survey), the estimated impact of incentives is larger for boys than girls. In the U.S., incentives increase the score of male students by an estimated 1.76 questions, while female scores increase by an estimated 1.01 questions. Interestingly, this pattern holds in both the U.S. and Shanghai. In Shanghai, male students in treatment improve by an estimated 1.13 questions with little effect among females. Interpreted through the lens of our overall findings, these results suggest that boys in particular lack intrinsic motivation to do well on low stakes tests.

4.2 Effects of incentives on measurement of student learning

In this section, we explore how U.S. performance would improve if our treatment effects applied to the PISA. We do so with the caveats that the PISA is administered to nationally representative samples using a carefully developed framework

least one threshold is optimal, the threshold values of PS_{ic} are then estimated by minimizing the sum of squared residuals of the equation stated above. Finally, the remaining parameters are then estimated by OLS.

[19] We cluster standard errors at the level of randomization and, as in Table 3, adjust the p-values from the U.S. estimates for multiple hypothesis testing using the resampling procedure of Anderson (2008). Also as above, Shanghai corrections are not possible because of the small number of clusters; those p-values are calculated using standard errors obtained via wild bootstrapping.

that includes different knowledge dimensions within each content domain (OECD, 2013), and that the effects of our incentives may not fully carry over to the PISA.

We first simulate the effects of incentives on PISA mathematics performance in the national U.S. sample. On the PISA, students receive different testing booklets, which include different blocks of mathematics, science, and English questions interspersed in different orders. We use the individual level PISA data from 2009, which reports the questions each student received, the order of the questions and question level responses. The solid line in Figure 6 estimates the proportion correct for U.S. students on PISA mathematics questions by question position (students received an average of 16 mathematics questions that appeared in positions 1-54).[20]

Similar to control group students in our experimental sample, the performance of U.S. students declines substantially over the course of the test. The probability an American student answers a question correctly falls by about 6 percentage points if it appears at the end of the test rather than the beginning (using the same data, Zamarro et al. (2016) find similar results). The dashed line in Figure 6 estimates counterfactual performance under incentives. To do this, we first estimate treatment effects by question position in our experimental sample and then map these effects to performance by question position on PISA.[21] As shown in Figure 6 (and discussed

[20]The proportion is the question position fixed effect plus the constant from a probit regression where the dependent variable equals one if the student answered the question correctly, controlling for question and question position fixed effects. There are 35 possible mathematics questions a student could receive. In the U.S., 3,640 (out of 5,233) students received at least one mathematics question (mean = 15.56, median=12, min=7, max =24).

[21]We estimate question level treatment effects using a probit regression of Equation 2 that includes an indicator for treatment interacted with question number. The treatment effect for question number q is is the average marginal effect implied by the overall treatment coefficient and treatment question number interaction. We map the 25 question specific effects from the experimental sample to the 54 questions on the PISA by rounding (25/54) times the question number to the nearest

in Section 4.1), treatment effects increase over the course of the test. Our estimates suggest that these effects could largely offset the decline in performance over the course of the exam that occurs in the absence of incentives.

We next simulate the effects of incentives on PISA mathematics scores by estimating PISA scores for the treatment and control groups in our experimental sample. We estimate PISA scores for each student in our sample using the individual level PISA data from the years in which our questions appeared: 2000 (five questions), 2003 (eleven questions) and 2012 (nine questions). We use the following fully flexible non-parametric approach. For each year (2000, 2003 and 2012), we calculate the average PISA score for every possible combination of correct and incorrect answers to the questions from our test that appeared on the actual PISA in that year. For example, for the five test questions that were taken on the 2000 PISA, there are $2^5 = 32$ cells, which represent possible combinations of answering each question correctly or incorrectly (non-response counts as incorrect).[22] PISA reports five plausible values (PVs) for each student's score, which enables researchers to estimate the distribution of each student's score. We average the five PVs to generate an average PISA score for each cell.

Each student in our sample receives predicted PISA scores from each year, which are the average scores from the cells that match his/her performance on the questions

integer, except for question 1 which we set to map to the question 1 treatment effect.

[22]Because different questions appear in different testing booklets in the same year, not every student in the individual level data received every question on our test in a given year. In 2000, 38,615 students in the individual level data received all five questions that appear on our test. In 2003, no student received all 11 questions from our test. However, 20,940 students received 8 of the 11 questions, and 21,940 receive the remaining 3 of the 11 questions, so we estimate two scores for 2003. In 2012, 106,381 students in the individual level data received the nine questions from our test.

in the 2000, 2003 and 2012 PISAs respectively. We then construct a weighted average of the student's predicted PISA scores in each year, where each score is weighted by the proportion of the total questions contributed from a given PISA. Non-parametric regressions in which we regress each possible cell on each plausible PISA score yields a weighted R^2 of 0.64, where we average over the five plausible scores and then weight the average by the proportion of total questions contributed from a given PISA.

Figure 7 presents the distributions of estimated PISA scores for Control and Treatment, with vertical lines indicating the mean for each group. Panel A presents the unweighted distributions, and Panel B presents the distributions weighted to match the U.S. national distribution.[23] We estimate an average treatment effect of 22 points in the unweighted distributions and an effect of 24 points in the weighted distributions. On the 2012 mathematics PISA, this is equivalent to moving from the U.S. ranking of 36^{th} to approximately the Australian ranking of 19^{th}.

5 Conclusion

The conjecture we raise and test in this article is that success on low stakes assessments does not solely reflect differences in ability between students across countries.

[23]We generate the weighted U.S. national distribution by pooling the 2000, 2003 and 2012 U.S. PISA data, weighted by the proportion of questions contributed by a given PISA. We weight estimated scores to make the Control distribution match the estimated U.S. national distribution as closely as possible. To generate these weights, we separate the estimated Control scores into 5 point bins and calculate the share of the sample in each bin, s_b^{Exp}. We then calculate the share of the national sample across the set of 5 point bins in the support of the experimental distribution, s_b^{USA}. Finally, we calculate a weight for each bin as the U.S. sample share divided by the experimental sample share. We assign treatment group weights from equivalent control group weights, under the assumption that the treatment effects are rank preserving, $w_b = \frac{s_b^{USA}}{s_b^{Exp}}$.

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. Rather we are focusing on between-country differences in effort on the test itself.[24] In this manner, we show that policy reforms that ignore the role of intrinsic motivation to perform well on the test may be misguided and have unintended consequences.

Our study may also shed light on two puzzles in the literature regarding the correlation between performance on low stakes assessments and economic outcomes. Test performance is highly correlated with both individual income and economic growth, but explains little of the variation in income across individuals in the U.S., and particularly under-predicts U.S. economic growth in cross-country comparisons (Murnane et al., 2000; 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 between ability and long-term outcomes (e.g., Borghans and Schils, 2013; Balart et al., 2015; Segal, 2012).

Our goal in this article is to highlight that low-stakes tests do not measure and compare ability in isolation, and as such the conclusions drawn from them should be more modest than current practice. Policymakers can allocate resources in a more efficient and productive manner by understanding the underlying reasons for test score differences. In addition, we hope that our findings serve as a catalyst to exploring the relevance of our conjecture in different domains, such as black-white or

[24]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.

male-female performance gaps. This can serve to not only deepen our understanding of test score differences across all groups in society, but also lead to a new discussion revolving around why such differences persist.

References

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: a reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Aoki, K. (2008). Confucious vs socrates: The impact of educational traditions of east and west in a global age. *International Journal of Learning*, 14:11.

Attali, Y., Neeman, Z., and Schlosser, A. (2011). Rise to the challenge or not give a damn: Differential performance in high vs low stakes tests. IZA discussion paper 5693.

Azmat, G., Calsamiglia, C., and Iriberri, N. (2016). Gender differences in response to big stakes. *Journal of the European Economic Association*, 14(6):1372–1400.

Balart, P., Oosterveen, M., and Webbink, H. D. (2015). Test scores, noncognitive skills and economic growth. Discussion paper, IZA.

Borghans, L. and Schils, T. (2013). The leaning tower of pisa: decomposing achievement test scores into cognitive and noncognitive components. Technical report, Unpublished manuscript. Draft version: July 22, 2013.

Braun, H., Kirsch, I., and Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade naep reading assessment. *Teachers College Record*, 113:2309–2344.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90:414–427.

Canay, I. A., Romano, J. P., and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*. forthcoming.

Carnoy, M. and Rothstein, R. (2013). What do international tests really show about us student performance. *Economic Policy Institute*, 28:32–33.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47:448–474.

DeMars, C. E. and Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *Intl. J. Test*, 10:207–229.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., and Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. In *Proceedings of the national Academy of Sciences 108*, pages 7716–7720.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy, and Practice*, 17:345–356.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, pages 1–17.

Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *The Journal of Economic Perspectives*, 25:191–210.

Grek, S. (2009). Governing by numbers: the pisa effect in europe. *Journal of Education Policy*, 24:23–37.

Hanushek, E. A. and Woessmann, L. (2011). How much do educational outcomes matter in oecd countries? *Economic Policy*, 26(67):427–491.

Hau, K. T. and Ho, I. T. (2010). Chinese students' motivation and achievement in the oxford handbook of chinese psychology, 187-204, m.h. In *Bond*. Oxford University Press.

Hess, R. D., Chang, C. M., and McDevitt, T. M. (1987). Cultural variations in family beliefs about children's performance in mathematics: Comparisons of peoples' republic of china, chinese american, and caucasian-american families. *Journal of Educational Psychology*, 79:179–188.

Imas, A. (2014). Working for the warm glow: On the benefits and limits of prosocial incentives. *Journal of Public Economics*, 114:14–18.

Jacob, B. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of Public Economics*, 89:761–796.

Jalava, N., Joensen, J. S., and Pellas, E. (2015). Grades and rank: impacts of non-financial incentives on test performance. *Journal of Economic Behavior & Organization*, 115:161–196.

Levitt, S., List, J. A., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8:183–219.

Murnane, R. J. et al. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, pages 547–568.

National Center for Education Statistics (NCES) (2015). The Nation's Report Card. U.S. Department of Education, Institute of Education Sciences, National Assessment of Educational Progress (NAEP).

Organisation for Economic Cooperation and Development, (OECD) (2013). Pisa 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy.

Organisation for Economic Cooperation and Development (OECD) (2014). Pisa 2012 results: What students know and can do, student performance in mathematics, reading and science. Volume I, Available at https://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-I.pdf.

Schmid, F. and Trede, M. (1995). A distribution free test for the two sample problem for general alternatives. *Computational Statistics & Data Analysis*, 20:409–419.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science*, 58:8.

Sievertsen, H. H., Gino, F., and Piovesan, M. (2016). Cognitive fatigue influences students' performance on standardized tests. *Proceedings of the National Academy of Sciences USA*, 113:2621–2624.

Stevenson, H. W., Lee, S., Chen, C., Stigler, J. W., Hsu, C., and Kitamura, S. (1990). Context of achievement: A study of american, chinese, and japanese children. *Monograph of the Society for Research in Child Development*, 55:221.

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

Tao, Y. K. V. (2016). Understanding chinese students' achievement patterns: Perspectives from social-oriented achievement motivation. *The Psychology of Asian Learners*, pages 621–634.

Wise, S. L. and DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment*, 10:1–17.

Woessmann, L. (2016). The importance of school systems: evidence from international differences in student achievement. *Journal of Economic Perspectives*, 30:3–31.

Zamarro, G., Hitt, C., and Mendez, I. (2016). reexamining international differences in achievement and non-cognitive skills, When students don't care.

Figure 1: PISA worldwide percentage correct, by question

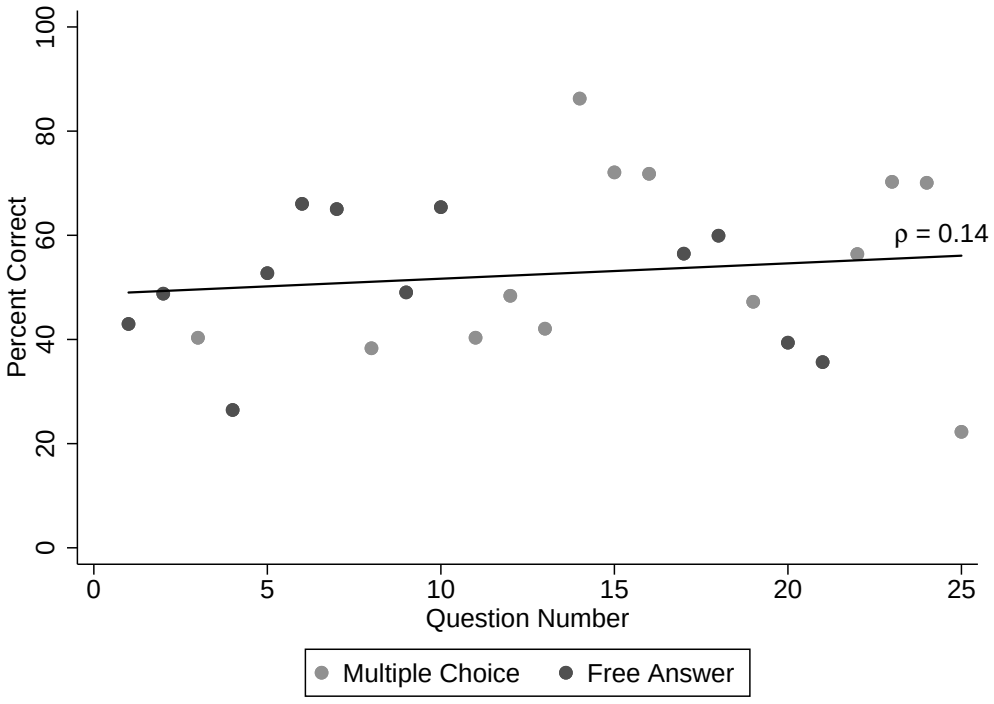
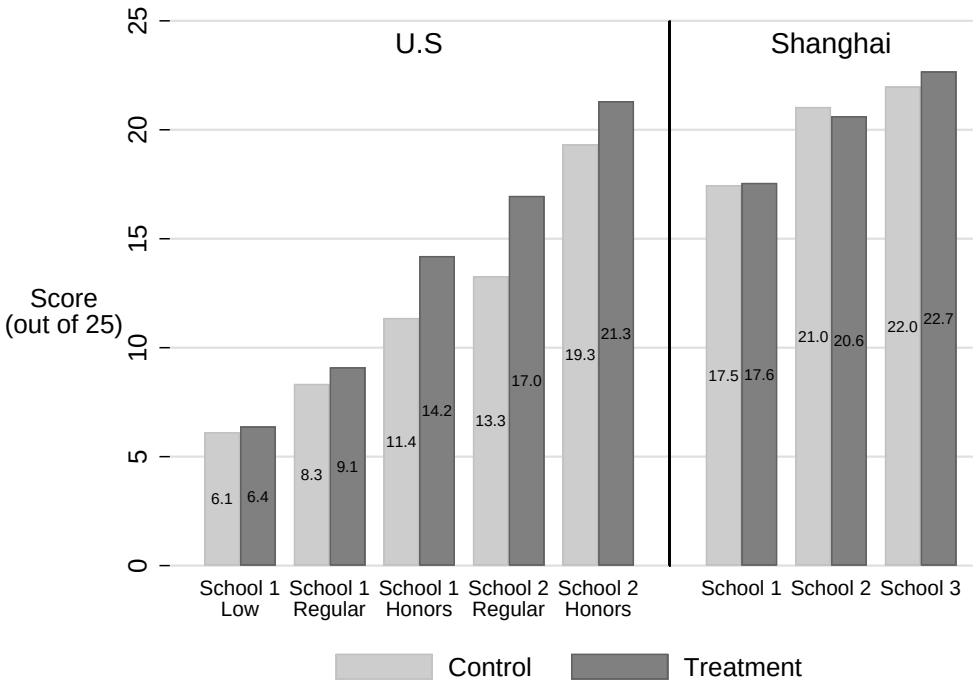
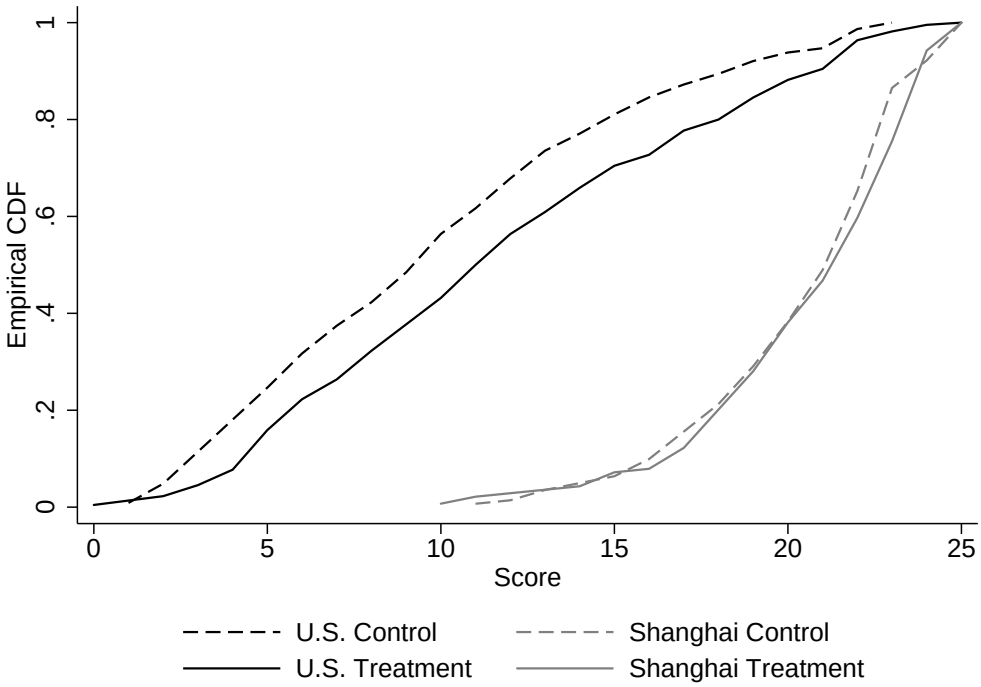


Figure 2: Average test score by group and treatment: U.S. vs. Shanghai



Notes: Average score for students who received no incentives (Control) and for students who received incentives (Treatment) by school and track.

Figure 3: Distribution of test scores by treatment group



Notes: Test of equality of distributions p-values: U.S. $p = 0.0031$, Shanghai $p = 0.36$.

Figure 4: Proportion of questions answered by question and treatment group

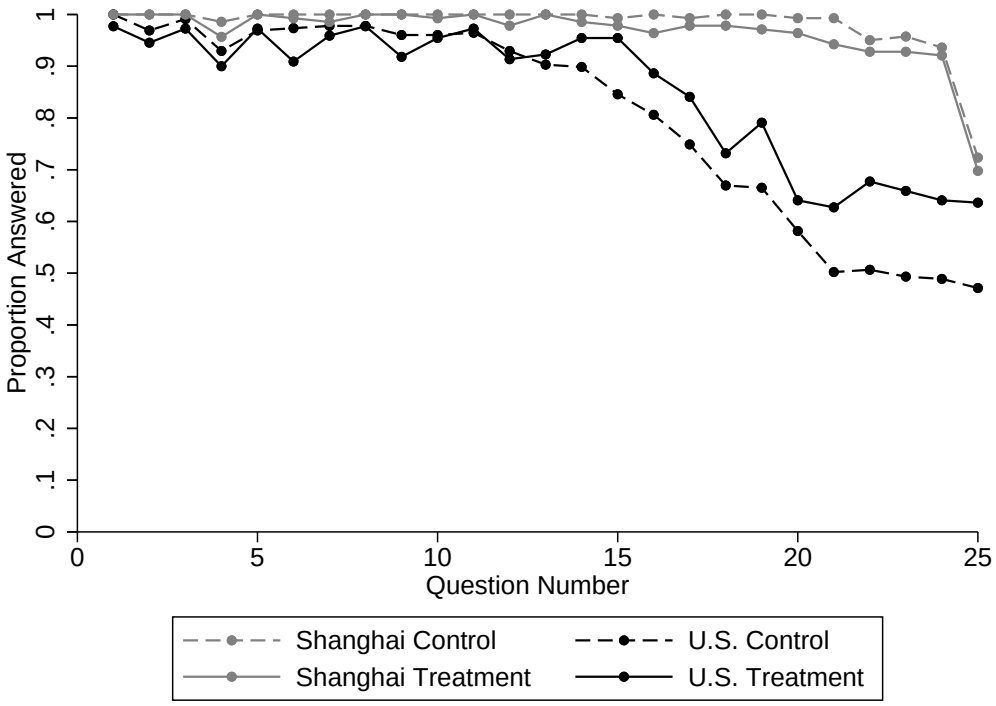
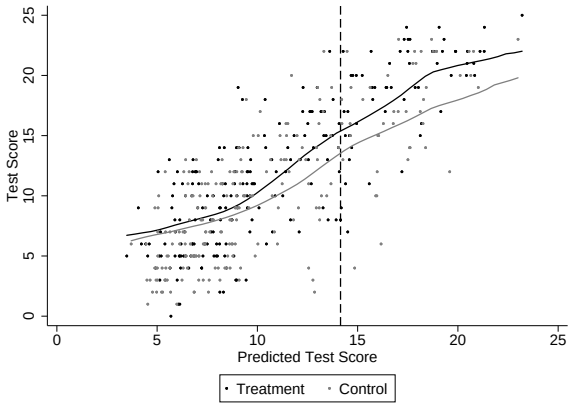
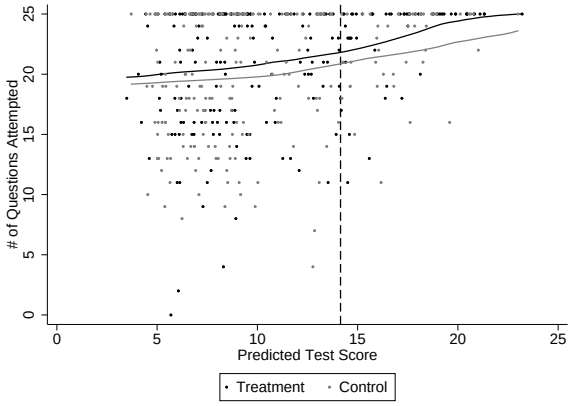


Figure 5: Treatment effects by predicted score, U.S.

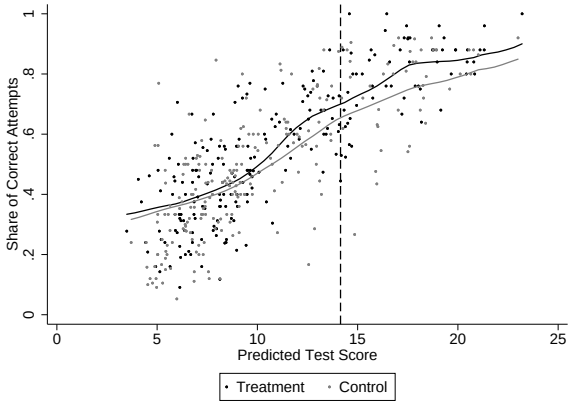
(a) Test score



(b) Questions attempted

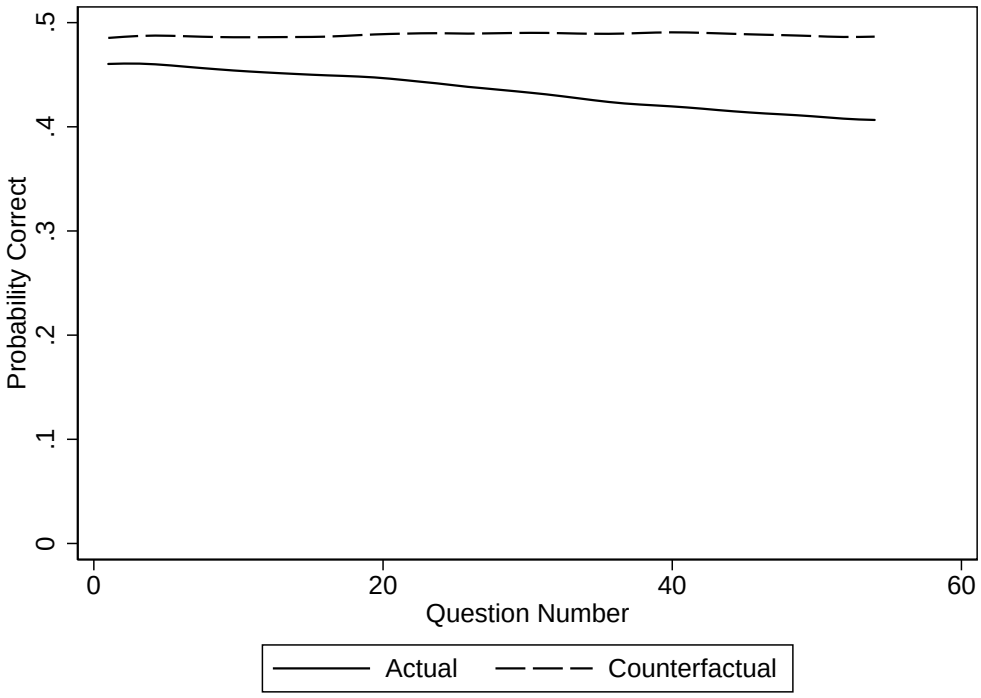


(c) Share of attempts correct



Notes: We predict score using age, gender, race/ethnicity and baseline exam score in the U.S. control group. We estimate the control and treatment lines using kernel weighting. The vertical line at 14.15 is the U.S. national average.

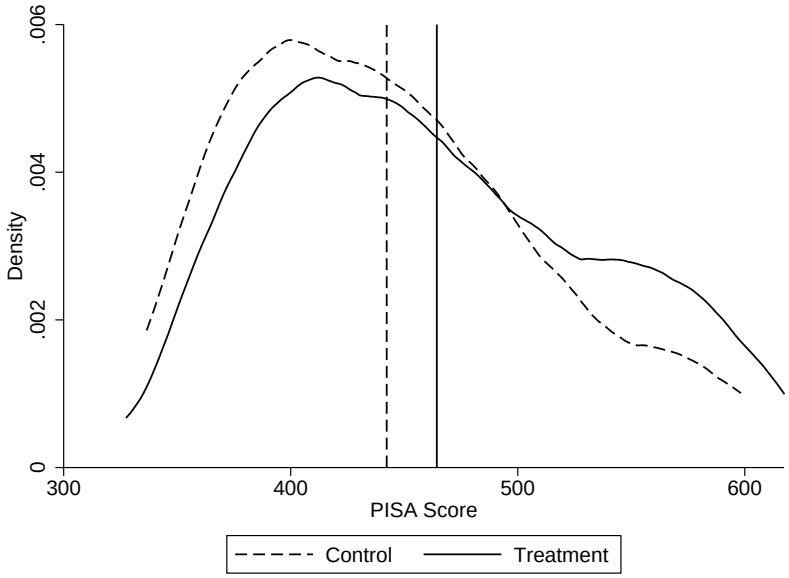
Figure 6: Estimated treatment effects on PISA by question order, U.S.



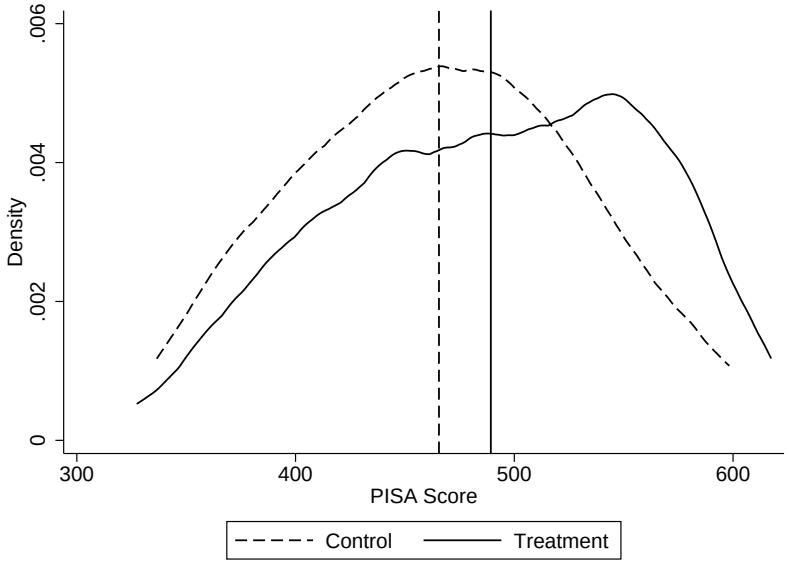
Notes: Lines are smoothed using local polynomial regression.

Figure 7: Distribution of estimated PISA scores

(a) Unweighted



(b) Nationally weighted



Notes: The vertical lines in Panel A indicate the mean for Control (442.29) and Treatment (464.37). The vertical lines in Panel B indicate the mean for Control (465.56) and Treatment (489.26). In Panel B, we weight estimated scores in Control to match the estimated U.S. national distribution. We weight estimated scores in Treatment using Control weights for equivalent ranks.

Table 1: Baseline characteristics by treatment group

	U.S.		Shanghai	
	Control	Treatment	Control	Treatment
N	227	220	141	139
Female	0.50	0.49	0.55	0.47
	(0.50)	(0.50)	(0.73)	(050)
Age	16.19	16.06	16.23	16.19
	(0.76)	(0.65)	(0.40)	(0.38)
Baseline exame score	-0.09	0.09	0.01	-0.01
	(0.94)	(1.05)	(1.03)	(0.98)
Asian	0.07	0.06		
	(0.26)	(0.24)		
Black	0.18	0.18		
	(0.39)	(0.39)		
Hispanic white	0.30	0.27		
	(0.46)	(0.45)		
Hispanic non-white	0.05	0.03		
	(0.22)	(0.18)		
White	0.39	0.45		
	(0.49)	(0.50)		
Other	0.00	0.01		
	(0.00)	(0.10)		

Notes: The table reports group means. Standard deviations in parentheses. Asterisks indicate within-country difference of group means with standard errors clustered by class (except U.S. school 2, which was randomized at the individual level) at the 10*/5**/1*** percent level.

Table 2: Effects of incentives on test scores, by country

	U.S.		Shanghai		U.S. = Shanghai
	(1)	(2)	(3)	(4)	p-value
Treatment	1.59	1.36	0.25	0.22	0.011
(p-value)	(0.001)	(0.001)	(0.524)	(0.573)	
Control mean	10.22		20.76		
(Std. devatiation)	(5.64)		(3.06)		
School-track FE	Yes	Yes	Yes	Yes	
Covariates	No	Yes	No	Yes	
Standardized effect size	0.23	0.20	0.04	0.03	
Students	447	447	280	280	
Clusters	131	131	8	8	

Notes: OLS estimates. p-values in parentheses calculated via wild bootstrapping with clustering by class (except U.S. school 2, which was randomized at the individual level). All regressions control for school-track (U.S.) or school (Shanghai) fixed effects. Columns 2 and 4 add controls for gender, age and race/ethnicity (U.S. only). One observation from column 2 imputes age to be the average age in the U.S. sample because age is not recorded for that student. The final column tests whether the treatment effect is equal in the U.S. and Shanghai using a randomization test. Effect sizes are standardized using the full sample.

Table 3: Effects of incentives on proxies for effort, U.S.

	All questions (1)	Q 1-13 (13 questions) (2)	Q 14-25 (12 questions) (3)
Panel A: Questions Attempted			
Treatment	0.038	-0.022	0.102
(p-value)	(0.029)	(0.245)	(0.010)
Control mean	0.807	0.962	0.640
(Std. deviation)	(0.394)	(0.191)	(0.480)
Observations	11,175	5,811	5,364
Clusters	447	447	447
Panel B: Proportion of Attempted Questions Correct			
Treatment	0.039	0.041	0.035
(p-value)	(0.003)	(0.010)	(0.099)
Control mean	0.516	0.494	0.549
(Std. deviation)	(0.500)	(0.500)	(0.498)
Observations	9,276	5,544	3,732
Clusters	446	446	417
Panel C: Proportion of Questions Correct			
Treatment	0.054	0.030	0.079
(p-value)	(0.0001)	(0.072)	(0.001)
Control mean	0.416	0.475	0.351
(Std. deviation)	(0.493)	(0.499)	(0.477)
Observations	11,175	5,811	5,364
Clusters	447	447	447
School-track FE	Yes	Yes	Yes
Question FE	Yes	Yes	Yes
Covariates	Yes	Yes	Yes

Notes: OLS estimates. p-values in parentheses calculated via wild bootstrapping with clustering by class (except U.S. school 2, which was randomized at the individual level). p-values in columns (2) and (3) are adjusted for multiple hypothesis testing following the procedure of Anderson (2008). All columns control for school-track fixed effects, age, gender and race/ethnicity.

Table 4: Effects of incentives on test scores, by gender

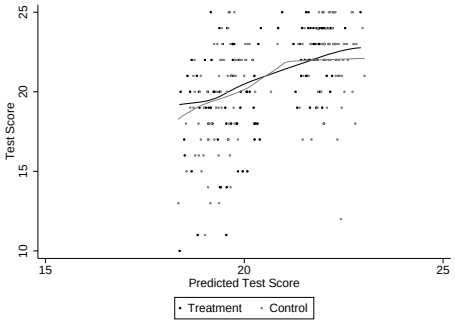
	U.S.		Shanghai	
	Male	Female	Male	Female
	(1)	(2)	(3)	(4)
Treatment	1.67	0.97	1.13	-0.39
(p-value)	(0.014)	(0.018)	(0.057)	(0.586)
Control mean	10.36	10.08	20.23	21.19
(Std. deviation)	(5.97)	(5.31)	(3.08)	(2.99)
Observations	226	221	137	143
Clusters	85	70	8	8

Notes: OLS estimates. U.S. p-values are adjusted for multiple hypothesis testing following the procedure of Anderson (2008) Shanghai p-values are calculated using standard errors obtained via wild bootstrapping. All columns control for school-track fixed effects, age, gender and race/ethnicity.(U.S. only).

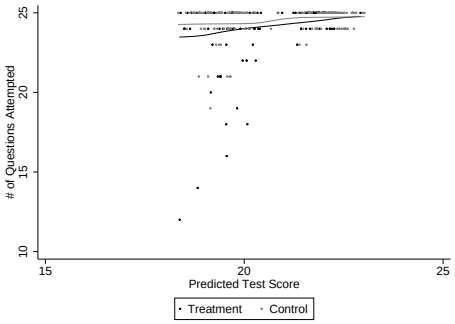
Appendix A. Figures and Tables

Figure A1: Treatment effects by predicted score, Shanghai

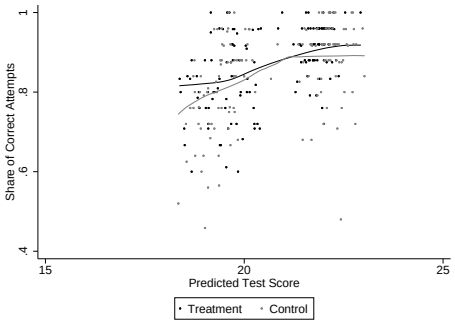
(a) Test score



(b) Questions attempted



(c) Questions attempted



Notes: We predict score using age, gender, and baseline exam score in the Shanghai control group. We estimate the control and treatment lines using kernel weighting.

Table A1: Effects of incentives on U.S. test scores, sensitivity checks

	Main sample	Drop if missing age	Keep non-10th	Keep ELL	Control for baseline exam score
	(1)	(2)	(3)	(4)	(5)
Treatment	1.36	1.34	1.37	1.36	1.38
(p-value)	(0.001)	(0.001)	(0.001)	(0.000)	(0.036)
Control mean	10.22	10.22	9.59	9.91	12.14
(Std. deviation)	(5.64)	(5.64)	(5.58)	(5.69)	(5.84)
School-track FE	Yes	Yes	Yes	Yes	No
Covariates	Yes	Yes	Yes	Yes	Yes
Baseline exam score	No	No	No	No	Yes
Students	447	446	535	469	348
Clusters	131	130	132	134	121

Notes: OLS estimates. p-values in parentheses calculated via wild bootstrapping with clustering by class (except U.S. school 2, which was randomized at the individual level). Columns (1)-(4) include school fixed effects. Column (5) controls for baseline exam score (students with missing baseline scores are dropped). All columns control for age, gender and race/ethnicity.

Table A2: Effects of incentives on proxies for effort, Shanghai

	All questions (1)	Q 1-13 (13 questions) (2)	Q 14-25 (12 questions) (3)
Panel A: Questions Attempted			
Treatment	-0.014	-0.006	-0.024
(p-value)	(0.223)	(0.034)	(0.276)
Control mean	0.981	0.998	0.962
(Std. deviation)	(0.137)	(0.033)	(0.192)
Observations	7,000	3,640	3,360
Clusters	280	280	280
Panel B: Proportion of Attempted Questions Correct			
Treatment	0.019	0.021	0.016
(p-value)	(0.177)	(0.081)	(0.353)
Control mean	0.861	0.861	0.862
(Std. deviation)	(0.345)	(0.346)	(0.345)
Observations	6,814	3,625	3,189
Clusters	280	280	280
Panel C: Proportion of Questions Correct			
Treatment	0.006	0.016	-0.004
(p-value)	(0.669)	(0.210)	(0.866)
Control mean	0.845	0.860	0.829
(Std. deviation)	(0.362)	(0.347)	(0.376)
Observations	7,000	3,640	3,360
Clusters	280	280	280
School-track FE	Yes	Yes	Yes
Question FE	Yes	Yes	Yes
Covariates	Yes	Yes	Yes

Notes: OLS estimates. p-values in parentheses calculated via wild bootstrapping with clustering by class. All columns include controls for school fixed effects, age, and gender.

Table A3: Effects of incentives by predicted score threshold, U.S.

	Score		Attempted	Proportion Correct	
Predicted score threshold:	< 11.04	≥ 11.04	n/a	< 11.002	≥ 11.002
	(1)	(2)	(3)	(4)	(5)
Treatment	0.79	2.24	1.01	0.028	0.054
(p-value)	(0.221	(0.012)	(0.060)	(0.306)	(0.048)
Control mean	7.37	15.27	20.19	0.388	0.711
(Std. deviation)	(3.63)	(4.99)	(5.00)	(0.160)	(0.162)
School-track FE	No	No	No	No	No
Covariates	Yes	Yes	Yes	Yes	Yes
Std. effect size	0.11	0.33	0.23	0.11	0.22
Students	270	177	447	269	178
Clusters	29	120	131	29	121

Notes: The table reports estimates of threshold regression models where the optimal number of predicted score thresholds for each outcome are estimated by minimizing the Bayesian Information Criterion, the threshold values are estimated by minimizing the sum of squared residuals of the equation defined in footnote 18, and the remaining parameters are estimated using least squares. p-values in parentheses adjusted for multiple hypothesis testing following the procedure of Anderson (2008) in columns (1)-(2) and columns (4)-(5). p-values in column (3) calculated via wild bootstrapping with clustering by class (except U.S. school 2, which was randomized at the individual level). All columns include the following covariates: age, gender, race/ethnicity. School-track fixed effects are not included because they are collinear with predicted score.

Appendix B. Test Questions (for online publication)

Question 1

Mark (from Sydney, Australia) and Hans (from Berlin, Germany) often communicate with each other using "chat" on the internet. They have to log on to the internet at the same time to be able to chat.

To find a suitable time to chat, Mark looked up a chart of world times and found the following:

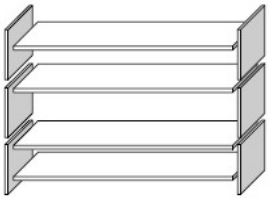
At 7:00 PM in Sydney, what time is it in Berlin?

NOTE: In your answer, please specify the hour, minutes, and whether it is AM or PM. For example, if your answer is 3 PM, write your answer as 3:00 PM.

Question 2

To complete one set of bookshelves a carpenter needs the following components:

4 long wooden panels,
6 short wooden panels,
12 small clips,
2 large clips and
14 screws.



The carpenter has in stock 26 long wooden panels, 33 short wooden panels, 200 small clips, 20 large clips and 510 screws.

How many sets of bookshelves can the carpenter make? (units not required)

Question 3

A documentary was broadcast about earthquakes and how often earthquakes occur. It included a discussion about the predictability of earthquakes.

A geologist stated: "In the next twenty years, the chance that an earthquake will occur in Zed City is two out of three".

Which of the following best reflects the meaning *of the geologist's statement*?

- ○ 2/3 x 20 = 13.3, so between 13 and 14 years from now there will be an earthquake in Zed City.
- ○ 2/3 is more than 1/2, so you can be sure there will be an earthquake in Zed City at some time during the next 20 years.
- ○ The likelihood that there will be an earthquake in Zed City at some time during the next 20 years is higher than the likelihood of no earthquake.
- ○ You cannot tell what will happen, because nobody can be sure when an earthquake will occur.

<< >>

Question 4

Infusions (or intravenous drips) are used to deliver fluids and drugs to patients.

Nurses need to calculate the drip rate, D, in drops per minute for infusions.

They use the formula:

$$D = \frac{dv}{60n}$$, where

d is the drop factor measured in drops per millilitre (mL)

v is the volume in mL of the infusion

n is the number of hours the infusion is required to run

Nurses need to calculate the volume of the infusion, v, from the drip rate, D.

An infusion with a drip rate of 50 drops per minute has to be given to a patient for 3 hours. For this infusion, the drop factor is 25 drops per milliliter.

What is the volume in mL of the infusion? (units not required)

Question 5

You are making your own dressing for a salad.

Here is a recipe for 100 milliliters (mL) of dressing.

Salad Oil:	60 mL
Vinegar:	30 mL
Soy sauce:	10 mL

How many milliliters (mL) of salad oil do you need to make 150 mL of this dressing? (units not required)

Question 6

A car magazine uses a rating system to evaluate new cars, and gives the award of "The Car of the Year" to the car with the highest total score. Five new cars are being evaluated, and their ratings are shown in the table.

Car	Safety Features (S)	Fuel Efficiency (F)	External Appearance (E)	Internal Fittings (T)
Ca	3	1	2	3
M2	2	2	2	2
Sp	3	1	3	2
N1	1	3	3	3
KK	3	2	3	2

The ratings are interpreted as follows:

3 points = Excellent
2 points = Good
1 point = Fair

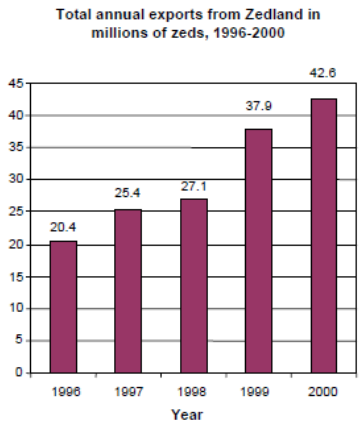
To calculate the total score for a car, the car magazine uses the following rule, which is a weighted sum of the individual score points:

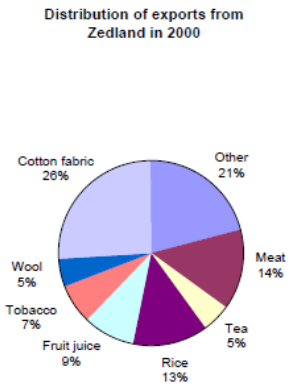
$$\text{Total Score} = (3 \times S) + F + E + T$$

Calculate the total score for Car "Ca". Write your answer in the space below. (units not required)

Question 7

The graphics below show information about exports from Zedland, a country that uses zeds as its currency.

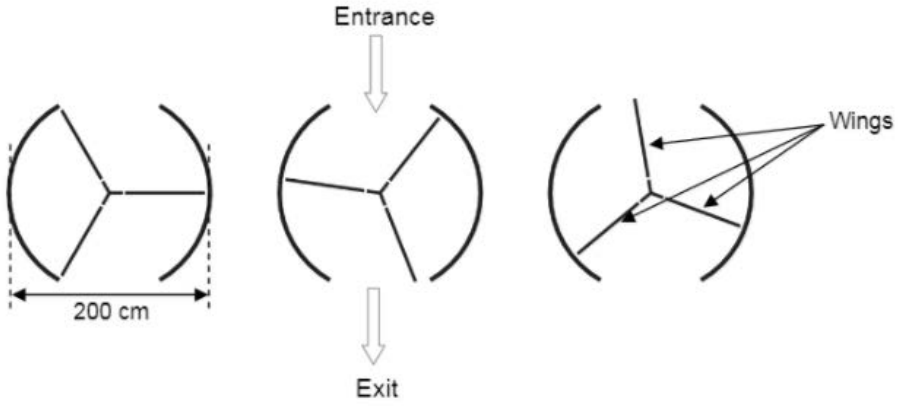




What was the total value (in millions of zeds) of exports from Zedland in 1998? (units not required)

Question 8

A revolving door includes three wings which rotate within a circular-shaped space. The inside diameter of this space is 2 meters (200 centimeters). The three door wings divide the space into three equal sectors. The plan below shows the door wings in three different positions viewed from the top.



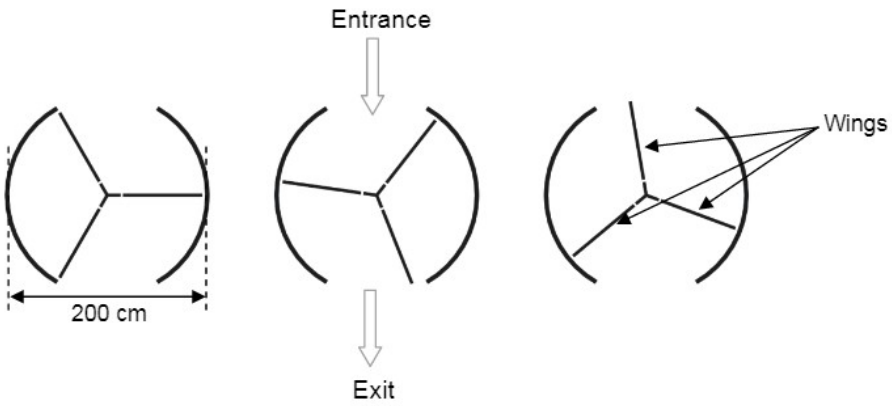
The door makes 4 complete rotations in a minute. There is room for a maximum of two people in each of the three door sectors.

What is the maximum number of people that can enter the building through the door in 30 minutes?

- ○ 60
- ○ 180
- ○ 240
- ○ 720

<< >>

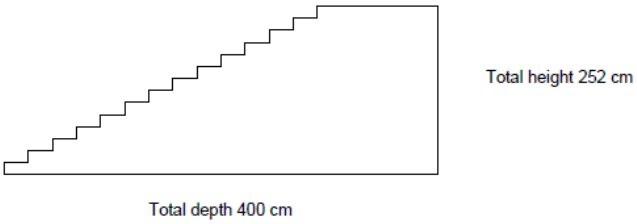
Question 9



What is the size in degrees of the angle formed by two door wings? (units not required)

Question 10

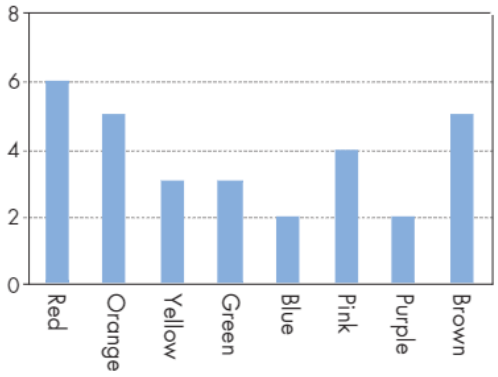
The diagram below illustrates a staircase with 14 steps and a total height of 252 cm:



What is the height of each of the 14 steps (in cm)? (units not required)

Question 11

Robert's mother lets him pick one candy from a bag. He can't see the candies. The number of candies of each color in the bag is shown in the following graph.



What is the probability that Robert will pick a red candy?

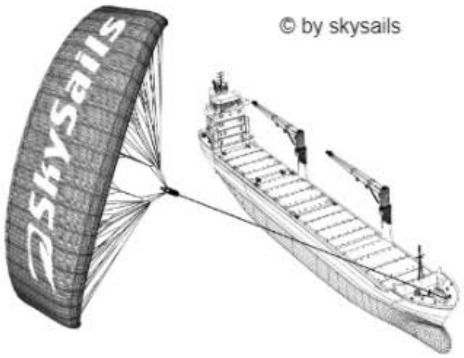
- ○ 10%
- ○ 20%
- ○ 25%
- ○ 50%

<< >>

Question 12

Ninety-five percent of world trade is moved by sea, by roughly 50,000 tankers, bulk carriers and container ships. Most of these ships use diesel fuel.

Engineers are planning to develop wind power support for ships. Their proposal is to attach kite sails to ships and use the wind's power to help reduce diesel consumption and the fuel's impact on the environment.



© by skysails

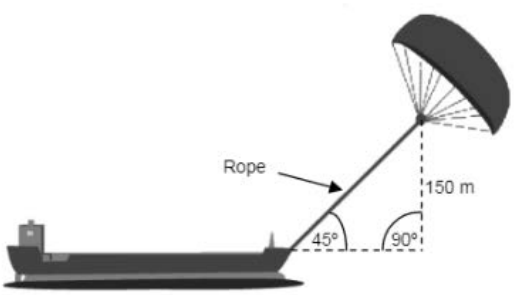
One advantage of using a kite sail is that it flies at a height of 150 m. There, the wind speed is approximately 25% higher than down on the deck of the ship.

At what approximate speed does the wind blow into a kite sail when a wind speed of 24 km/h is measured on the deck of the ship?

- ○ 6 km/h
- ○ 18 km/h
- ○ 25 km/h
- ○ 30 km/h
- ○ 49 km/h

<< >>

Question 13



Note: Drawing not to scale.
© by skysails

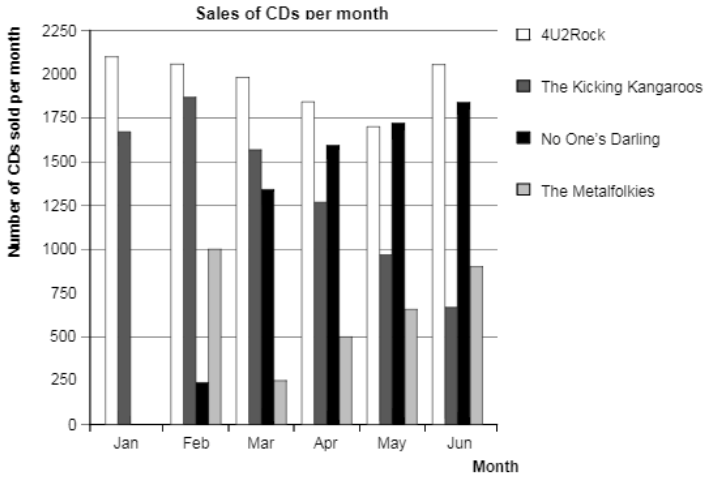
Approximately what is the length of the rope for the kite sail, in order to pull the ship at an angle of 45 degrees and be at a vertical height of 150 m, as shown in the diagram above?

- O 173 m
- O 212 m
- O 285 m
- O 300 m

<< >>

Question 14

In January, the new CDs of the bands 4U2Rock and The Kicking Kangaroos were released. In February, the CDs of the bands No One's Darling and The Metalfolkies followed. The following graph shows the sales of the bands' CDs from January to June.

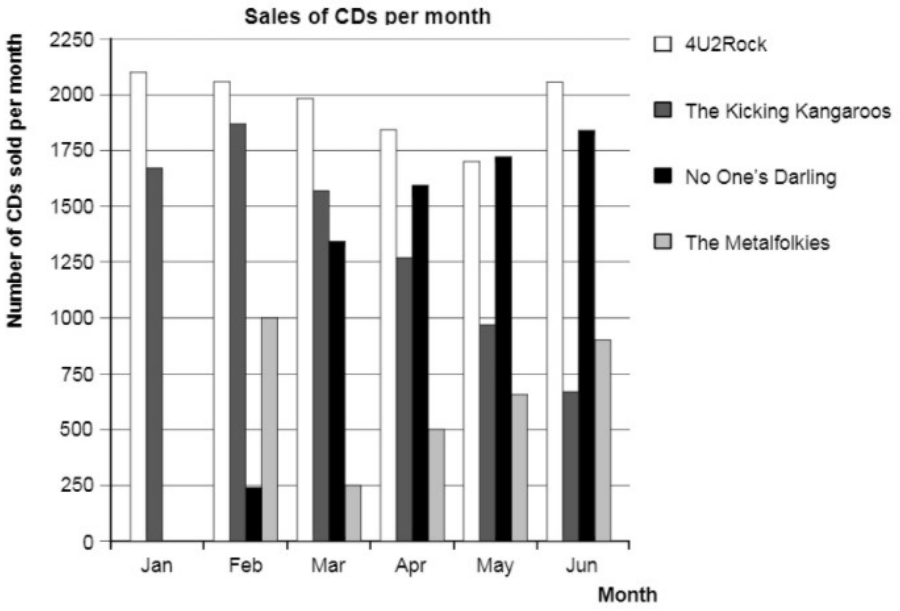


How many CDs did the band The Metalfolkies sell in April?

- ○ 250
- ○ 500
- ○ 1000
- ○ 1270

<< >>

Question 15

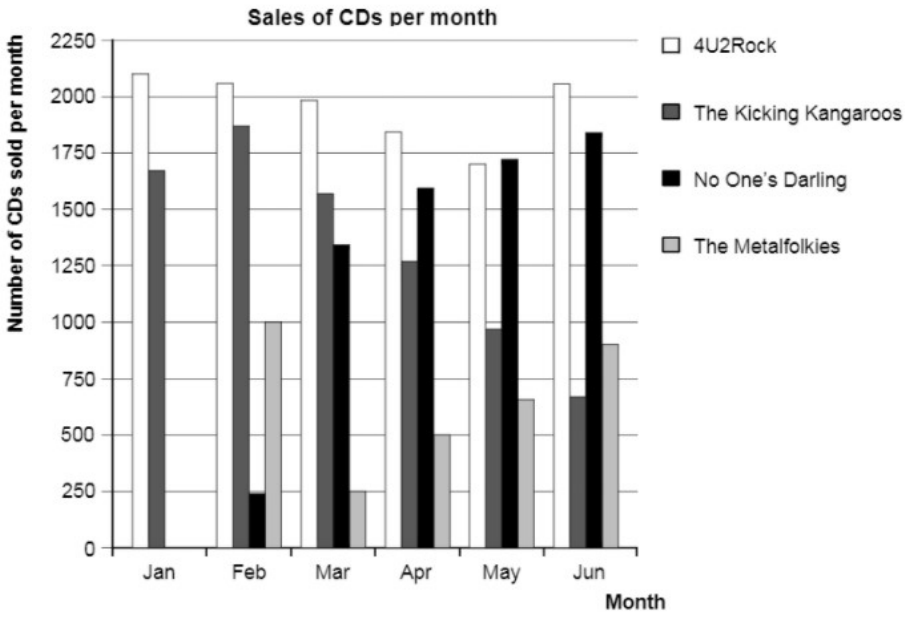


In which month did the band No One's Darling sell more CDs than the band The Kicking Kangaroos for the first time?

- ○ No Month
- ○ March
- ○ April
- ○ May

<< >>

Question 16



The manager of *The Kicking Kangaroos* is worried because the number of their CDs that sold decreased from February to June.

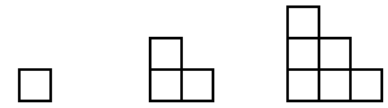
What is the estimate of their sales volume for July if the same negative trend continues?

- ○ 70 CDs
- ○ 370 CDs
- ○ 670 CDs
- ○ 1340 CDs

<< >>

Question 17

Robert builds a step pattern using squares. Here are the stages he follows.



As you can see, he uses one square for Stage 1, three squares for Stage 2 and six for Stage 3.

How many squares should he use for the fourth stage? (units not required)

<< >>

Question 18

On returning to Singapore after 3 months, Mei-Ling had 3,900 ZAR left. She changed this back to Singapore dollars, noting that the exchange rate had changed to:

1 SGD = 4.0 ZAR

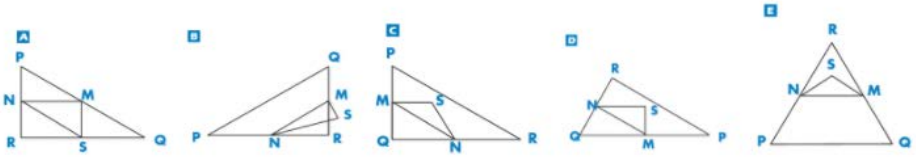
How much money in Singapore dollars did Mei-Ling get? (units not required)

<< >>

Question 19

Choose the one figure below that fits the following description.

Triangle PQR is a right triangle with right angle at R. The line RQ is less than the line PR. M is the midpoint of the line PQ and N is the midpoint of the line QR. S is a point inside the triangle. The line MN is greater than the line MS.



○ ○ ○ ○ ○

<< >>

Question 20

In a pizza restaurant, you can get a basic pizza with two toppings: cheese and tomato. You can also make up your own pizza with **extra** toppings. You can choose from four different extra toppings: olives, ham, mushrooms and salami.

Ross wants to order a pizza with two different **extra** toppings.

How many different combinations can Ross choose from? (units not required)

<< >>

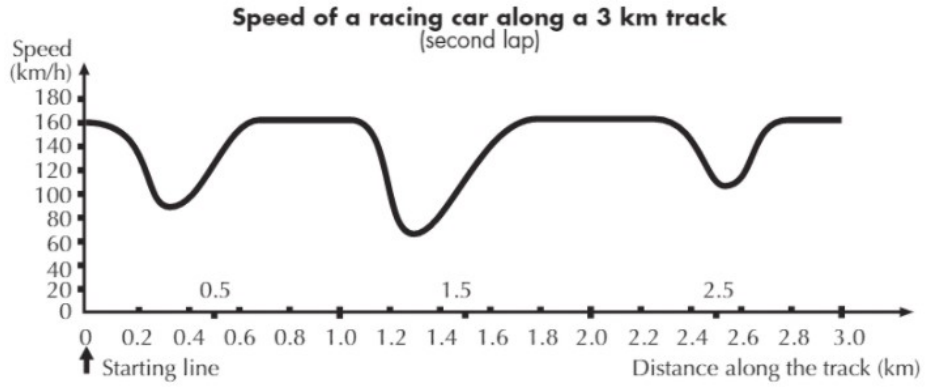
Question 21

In Mei Lin's school, her science teacher gives tests that are marked out of 100. Mei Lin has an average of 60 marks on her first four Science tests. On the fifth test she got 80 marks.

What is the average of Mei Lin's marks in Science after all five tests? (units not required)

Question 22

This graph shows how the speed of a racing car varies along a flat 3 kilometre track during its second lap.

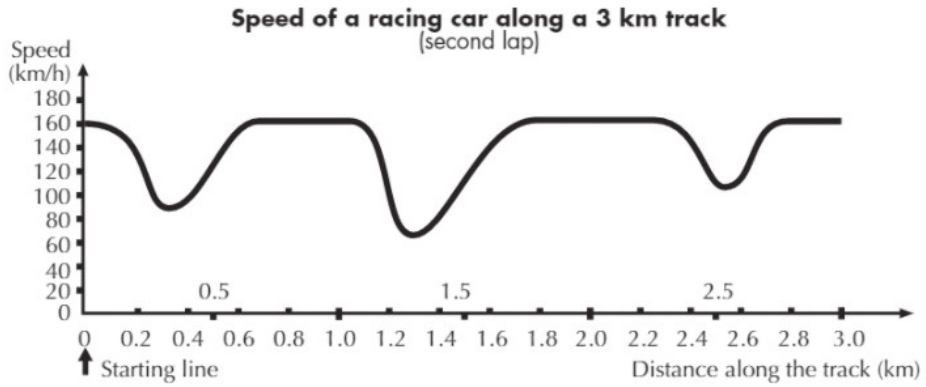


What is the approximate distance from the starting line to the beginning of the longest straight section of the track?

- ○ 0.5 km
- ○ 1.5 km
- ○ 2.3 km
- ○ 2.6 km

<< >>

Question 23

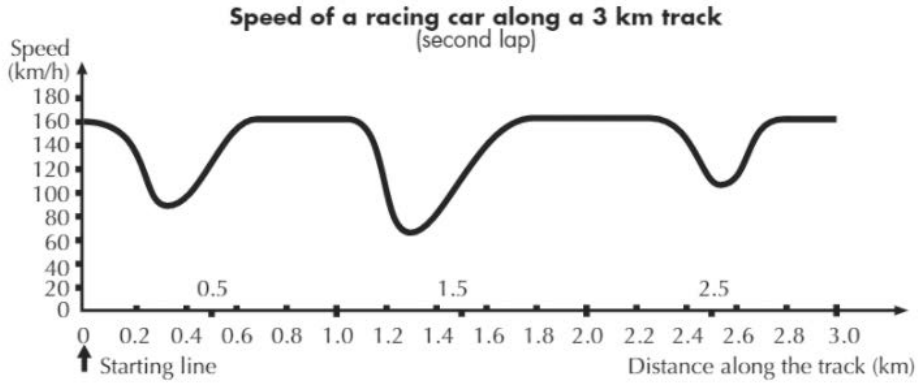


Where was the lowest speed recorded during the second lap?

- ○ at the starting line.
- ○ at about 0.8 km.
- ○ at about 1.3 km.
- ○ halfway around the track.

<< >>

Question 24

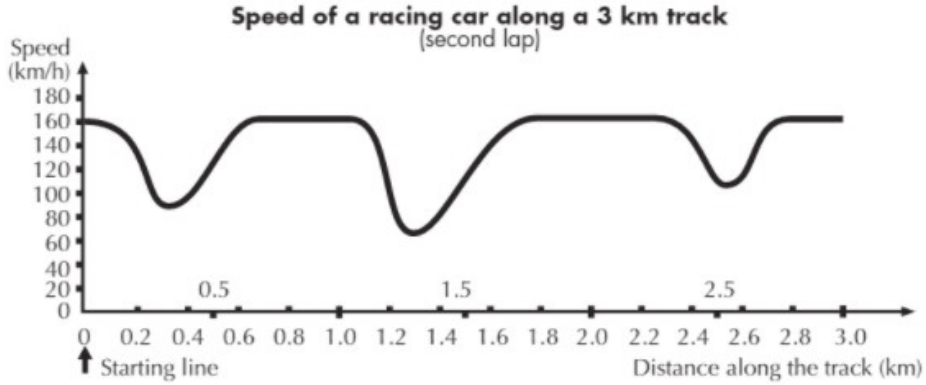


What can you say about the speed of the car between the 2.6 km and 2.8 km marks?

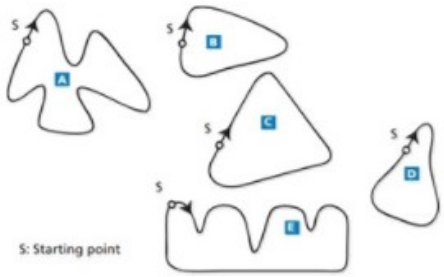
- ○ The speed of the car remains constant.
- ○ The speed of the car is increasing.
- ○ The speed of the car is decreasing.
- ○ The speed of the car cannot be determined from the graph.

<< >>

Question 25



Here are pictures of five tracks:



Along which one of these tracks was the car driven to produce the speed graph shown earlier?

- A
- B
- C
- D
- E