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

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

April, 2019

Abstract

U.S. 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 U.S. high schools and four Shanghai high schools. U.S. 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 classification: C91, D12, D81

Keywords: Education, Test Effort, Motivation, Field Experiment

[*] Gneezy: University of California-San Diego and CREED, University of Amsterdam, ugneezy@ucsd.edu. List: University of Chicago and NBER, jlist@uchicago.edu. Livingston: Bentley University, jlivingston@bentley.edu. Qin: Shanghai Jiao Tong University, xdqin@sjtu.edu.cn. Sadoff: University of California San Diego, ssadoff@ucsd.edu. Xu: University of Chicago, yangxu@uchicago.edu. 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. Please direct correspondence to Sally Sadoff.

1 Introduction

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 (NCES, 2015). Similarly, on the 2012 Programme for International Student Assessment (PISA), among the 65 countries and economies that participated, U.S. high school students ranked 36th for mathematics performance, with scores declining since 2009 (OECD, 2014).

In response to poor U.S. performance on such assessments, then 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."[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 U.S. perform so poorly relative to other countries despite its wealth and high per pupil expenditures? Examples of answers discussed in the literature

[1] See S. Dillon, Top test scores from Shanghai stun educators. *The New York Times* (2010; http://www.nytimes.com/2010/12/07/education/07education.html).

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 et al., 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 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.

In this study, we present an experimental methodology for comparing test effort across student groups. 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 at four schools in Shanghai and two schools in the U.S. 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 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 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. These effects are concentrated among students whose baseline performance is near the U.S. 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 assessments tests, which complicates the challenges of international test comparisons.

2 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,

[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 Appendix Figure A.1 for an illustration.

2010; DeMars and Wise, 2010; Borghans and Schils, 2013; Zamarro et al., 2016; Borgonovi and Biecek, 2016; Balart et al., 2017; Akyol et al., 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 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.[4] These studies demonstrate that rewards (both financial and non-financial) as well as how the test is framed can increase effort and improve performance (Duckworth et al., 2011; Braun et al., 2011; Levitt et al., 2016b; 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 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

[3] For example, Borghans and Schils (2013) and Zamarro et al. (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 et al., 2016a; Barrow and Rouse, 2018, provide a review).

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 U.S. (Zamarro et al., 2016). There is also evidence from descriptive studies showing 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 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.

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 36th on the same test. The PISA is conducted by the Organisation for Economic Cooperation 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 policy makers 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 U.S. 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 consists of 13 multiple-choice questions and 12 free answer fill-in-the-blank questions (see 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 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 ($\rho = 0.14$). 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 four 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 two schools with performance that is well above average.[7]

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

[6] 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\textunderscoremath2012.pdf`.

[7] School performance is rated compared to the average Shanghai 2015 Senior High School Entrance exam 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.

In the U.S., 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 U.S. 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, 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 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 Appendix C for instructions). Accordingly, we are assured that the incentives only influence effort on the test itself, not preparation for the exam.

We randomized at the class level in the lower performing school in the U.S. and in the 2016 sessions in Shanghai. We randomized at the individual level in the higher performing school in the U.S. and in the 2018 sessions in Shanghai.[12] In the U.S., we

[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 U.S. school, eleven 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 U.S. 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 obtained from `http://www.economist.com/content/big-mac-index` to determine currency conversion. 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.

[12] Differences in the randomization across schools and waves of the experiment were driven by

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.

4 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 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 U.S. or Shanghai samples (standard errors are clustered at the level of randomization). We note that our U.S. sample includes a slightly lower proportion of white students and slightly higher proportion of minority students (Asian, black and Hispanic) than the national average.

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.

[14] The standardized tests for which we have data are the 7th grade Massachusetts Comprehensive Assessment System mathematics assessment for U.S. school 1, the mathematics Secondary School Admissions Test for U.S. school 2, and the 2015 Shanghai Senior High School Entrance exam for all Shanghai schools.

4.1 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, U.S. 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 U.S. 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 U.S. 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 U.S. 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 Appendix Figure A.3, the financial incentives shift the entire distribution of U.S. 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 U.S. and Shanghai by Ordinary Least Squares (OLS), estimating the following equation

[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 nine) 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.

separately in each country:

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

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 U.S., race/ethnicity (Asian, black, 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 (U.S.) 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 U.S. and Shanghai, which we also calculate using a randomization test.[20]

In response to incentives, the performance of Shanghai students does not change while the scores of U.S. students increase substantially. The estimated treatment ef-

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

[18] 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. 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. Finally, the results are robust to including controls for baseline student standardized exam score rather than school-track fixed effects. See Appendix Table A.1 for results.

[19] See Young (forthcoming) 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 U.S. and Shanghai samples and estimate an OLS regression on test score that controls for a treatment assignment indicator, a U.S. 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.

fect in the U.S. 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 U.S. and Shanghai are significantly different at the 1% level. These results are consistent with our hypothesis that U.S. students are more responsive than Shanghai students to incentives for effort because they are less motivated at baseline.[21]

4.2 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 et al., 2016).

In Table 3, we report regression results for effort proxies, using the following

[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 1.8RMB 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.

equation estimated by OLS:

$$Y_{qicsw} = \alpha + \beta_1 Z_c + \beta_2 X_i + Q_q + \mu_s + \gamma_\omega + \epsilon_{qicsw} \qquad (2)$$

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 U.S., column 4 for Shanghai) reports the results using responses to all 25 questions. The next two columns (columns 2 and 3 for the U.S., 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 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 U.S.; columns 5 and 6 for Shanghai).

We first estimate the effect of incentives in the U.S. 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 U.S. 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% in the first half of the test to 64% in the second half.[22]

In Table 3 Panel B (columns 1 through 3), we estimate the effects of incentives in

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

the U.S. on the percentage of attempted questions answered correctly. 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 increase the share of attempted questions answered correctly by U.S. 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 U.S. 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 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.

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.

4.3 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 U.S.) 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 U.S. student. The Treatment and Control lines are estimated by performing 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.[24]

As shown in Figure 2 Panel A, 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 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 Appendix Table A.2, which de-

[23] In the U.S., 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.

[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.

tect a split at a predicted test score of 11 when the dependent variable is test score or proportion correct, 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]

5 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 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 method-

[25] We also examine treatment effects by gender (Appendix Table A.3). Incentives significantly increase both male and female scores in the U.S., 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.

ology 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 U.S. 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 U.S., 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 U.S. 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 between ability, intrinsic motivation and long-term outcomes (e.g., Borghans and Schils, 2013; Balart et al., 2017; Segal, 2012). Future work could also explore how best to interpret or perhaps even redesign low stakes assessment tests so that policy makers 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

[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 U.S. mathematics performance by 22 to 24 points, equivalent to moving the U.S. from 36th to 19th in the 2012 international mathematics rankings. While this gives a sense of magnitudes, we note that it is based on out of sample calculations for a non-representative sample, and holds constant the effort level exerted in all other countries.

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, S.P., K. Krishna, and J. Wang, "Taking PISA seriously: How accurate are low stakes exams?," 2018. NBER Working Paper No. 24930.

Aoki, K., "Confucious vs Socrates: The Impact of Educational Traditions of East and West in a Global Age," *International Journal of Learning*, 2008, *14*, 11.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar, "Large stakes and big mistakes," *The Review of Economic Studies*, 2009, *76* (2), 451–469.

Balart, P., M. Oosterveen, and D. Webbink, "Test scores, noncognitive skills and economic growth," *Economics of Education Review*, 2017. forthcoming.

Barrow, L. and C.E. Rouse, "Financial incentives and educational investment: The impact of performance-based scholarships on student time use," *Education Finance and Policy*, 2018, *14*, 419–448.

Borghans, L. and T. Schils, "The leaning tower of Pisa: decomposing achievement test scores into cognitive and noncognitive components," 2013. Unpublished manuscript. Draft version: July 22, 2013.

Borgonovi, F. and P. Biecek, "An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test," *Learning and Individual Differences*, 2016, *49*, 128–137.

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

Carnoy, M. and R. Rothstein, "What do international tests really show about U.S. student performance?," Technical Report, Economic Policy Institute 2013.

DellaVigna, S. and D. Pope, "What motivates effort? Evidence and expert forecasts," *The Review of Economic Studies*, 2018, *85* (2), 1029–1069.

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

Duckworth, A. L., P. D. Quinn, D. R. Lynam, R. Loeber, and M. Stouthamer-Loeber, "Role of test motivation in intelligence testing," in "Proceedings of the National Academy of Sciences 108" 2011, pp. 7716–7720.

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

Finn, B., "Measuring Motivation in Low-Stakes Assessments," *ETS Research Report Series*, 2015, pp. 1–17.

Fryer, R., “Financial incentives and student achievement: Evidence from randomized trials,” *Quarterly Journal of Economics*, 2011, *126*, 1755–1798.

Gneezy, U., J.A. List, J.A. Livingston, S. Sadoff, X. Qin, and Y. Xu, “Measuring success in education: the role of effort on the test itself,” 2017. NBER Working Paper No. 24004.

_, S. Meier, and P. Rey-Biel, “When and why incentives (don’t) work to modify behavior,” *The Journal of Economic Perspectives*, 2011, *25*, 191–210.

Grek, S., “Governing by numbers: The PISA effect in Europe,” *Journal of Education Policy*, 2009, *24*, 23–37.

Hanushek, Eric A. and Ludger Woessmann, “"How much do educational outcomes matter in OECD countries?",” *Economic Policy*, 2011, *26* (67), 427–491.

Hau, K. T. and I. T. Ho, “Chinese students’ motivation and achievement in The Oxford Handbook of Chinese Psychology, 187-204, M.H,” in “Bond,” Oxford University Press, 2010.

Hess, R. D., C. M. Chang, and T. M. McDevitt, “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*, 1987, *79*, 179–188.

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

Jacob, B., “Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools,” *Journal of Public Economics*, 2005, *89*, 761–796.

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

Levitt, S.D., J.A. List, and S. Sadoff, “The effect of performance-based incentives on educational achievement: Evidence from a randomized experiment,” 2016. NBER Working Paper No. 22107.

_, _, S. Neckermann, and S. Sadoff, “The behavioralist goes to school: Leveraging behavioral economics to improve educational performance,” *American Economic Journal: Economic Policy*, 2016, *8*, 183–219.

Murnane, Richard J., J.B. Willett, and Y. Duhaldeborde, “How important are the cognitive skills of teenagers in predicting subsequent earnings?,” *Journal of Policy Analysis and Management*, 2000, *19* (4), 547–568.

NCES, *"The Nation’s Report Card"*, 2015.

OECD, "PISA 2012 Results: What Students Know and Can Do–Student Performance in Mathematics, Reading and Science (Volume I, Revised Edition, February 2014)," 2014.

Schmid, F. and M. Trede, "Testing for first-order stochastic dominance: A new distribution-free test," *Journal of the Royal Statistical Society. Series D (The Statistician)*, 1996, *45* (3), 371–380.

Segal, C., "Working when no one is watching: Motivation, test scores, and economic success," *Management Science*, 2012, *58* (8), 1438–1457.

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

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

_, S. Lee, C. Chen, J. W. Stigler, C. Hsu, and S. Kitamura, "Context of achievement: A study of American, Chinese, and Japanese Children," *Monograph of the Society for Research in Child Development*, 1990, *55*, 221.

Tao, Y. K. V., "Understanding Chinese Students' Achievement Patterns: Perspectives from Social-Oriented Achievement Motivation," *The Psychology of Asian Learners*, 2016, *46* (4), 621–634.

Westfall, P.H. and S.S. Young, *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*, New York: John Wiley and Sons, 1993.

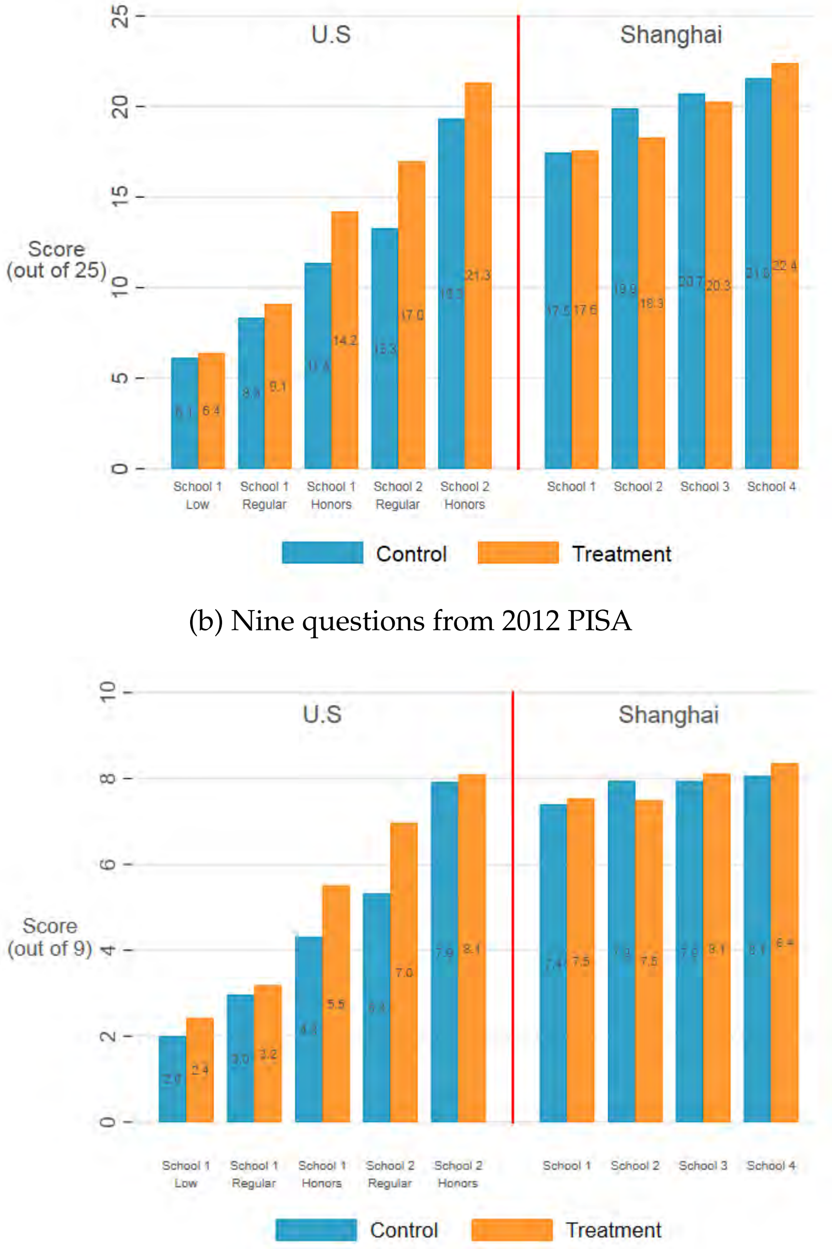
Wise, S. L. and C. E. DeMars, "Low examinee effort in low-stakes assessment: Problems and potential solutions," *Educational Assessment*, 2005, *10*, 1–17.

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

Young, A., "Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results," *Quarterly Journal of Economics*, forthcoming.

Zamarro, G., C. Hitt, and I. Mendez, "When students don't care: Reexamining international differences in achievement and non-cognitive skills," *EDRE Working Paper No. 2016-18*, 2016.

Figure 1: Average test score by group and treatment: U.S. vs. Shanghai

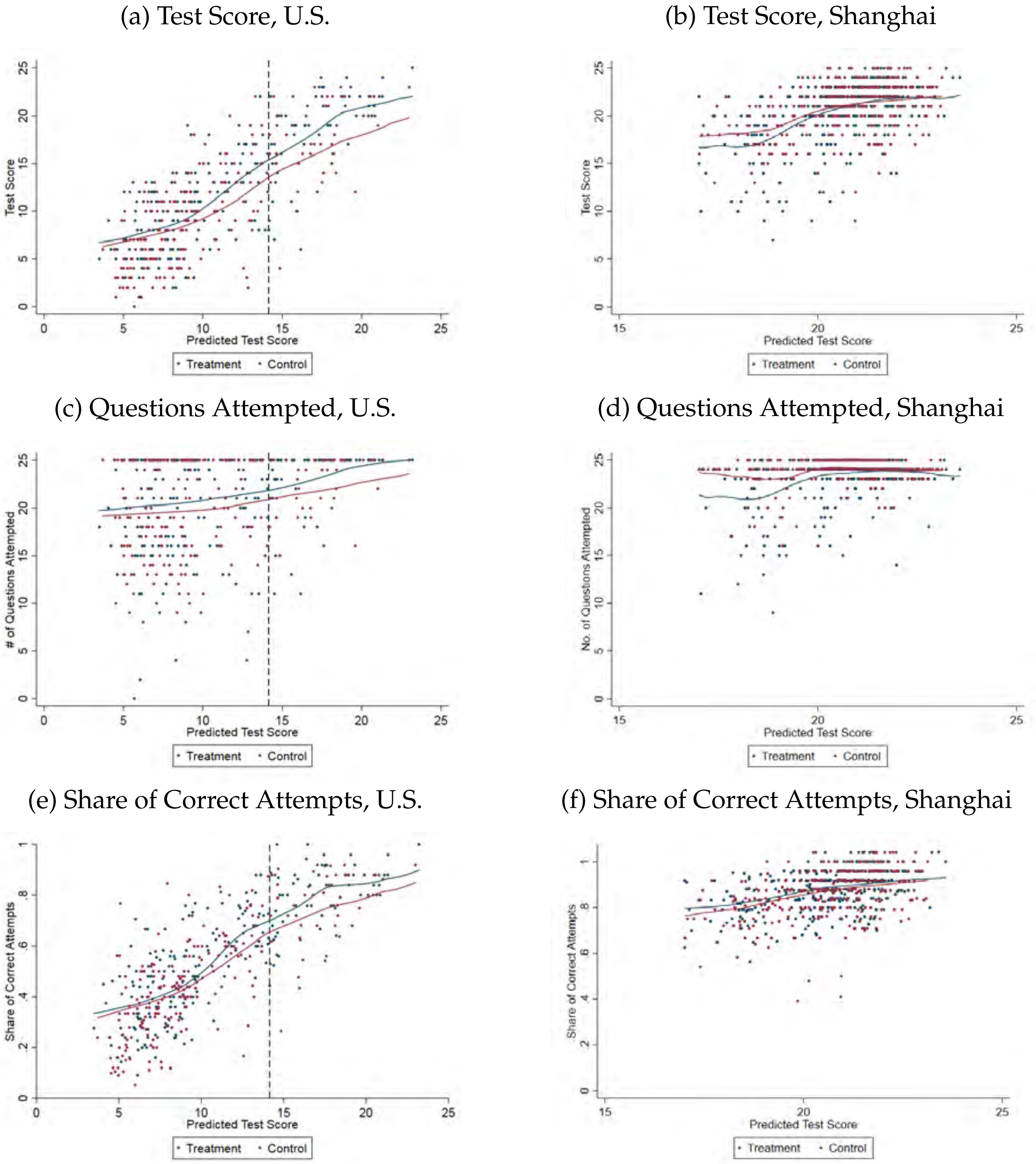


(a) Full test (all 25 questions)

(b) Nine questions from 2012 PISA

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 U.S. students when these questions were administered as part of official PISA tests is 14.15. We calculate this average using the proportion of U.S. 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 U.S. and Shanghai. The average score on these nine questions among U.S. students on the official PISA is 5.09. The average score in Shanghai is 7.37.

Figure 2: Treatment effects by predicted score



(a) Test Score, U.S.

(b) Test Score, Shanghai

(c) Questions Attempted, U.S.

(d) Questions Attempted, Shanghai

(e) Share of Correct Attempts, U.S.

(f) Share of Correct Attempts, Shanghai

Notes: For the U.S., we predict score using age, gender, race/ethnicity and baseline exam score in the U.S. control group. The vertical line at 14.15 in the U.S. panels is the U.S. national average. We calculate this average using the proportion of U.S. 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.

Table 1: Sample characteristics by treatment group

		U.S.			Shanghai	
	Control	Treatment	Nat'l avg.	Control	Treatment	Nat'l avg.
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	0.07	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)	
N (students)	227	220		333	323	

Notes: The table reports group means. Standard deviations in parentheses. U.S. national 10th grade averages for gender and ethnicity categories are computed from enrollment numbers from the U.S. 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 U.S. 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 U.S. school 1, the Quantitative Secondary School Admissions Test (SSAT) for U.S. 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.

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	1.34	-0.26	-0.28	0.0002
(Std. error)	(0.40)	(0.34)	(0.27)	(0.26)	
[p-value]	[0.001]	[0.001]	[0.399]	[0.367]	
Control mean	10.22 (5.64)		20.50 (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 U.S. 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 (forthcoming). The dependent variable is the student's score on the full 25 question test. All regressions control for school-track (U.S.) or school (Shanghai) fixed effects and a wave fixed effect (Shanghai only). Columns 2 and 4 add controls for race/ethnicity (U.S. only), gender, and age. 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. To conduct this test, we pool the U.S. and Shanghai samples and estimate an OLS regression on test score that controls for a treatment assignment indicator, a U.S. 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 (forthcoming). Effect sizes are standardized using the full sample.

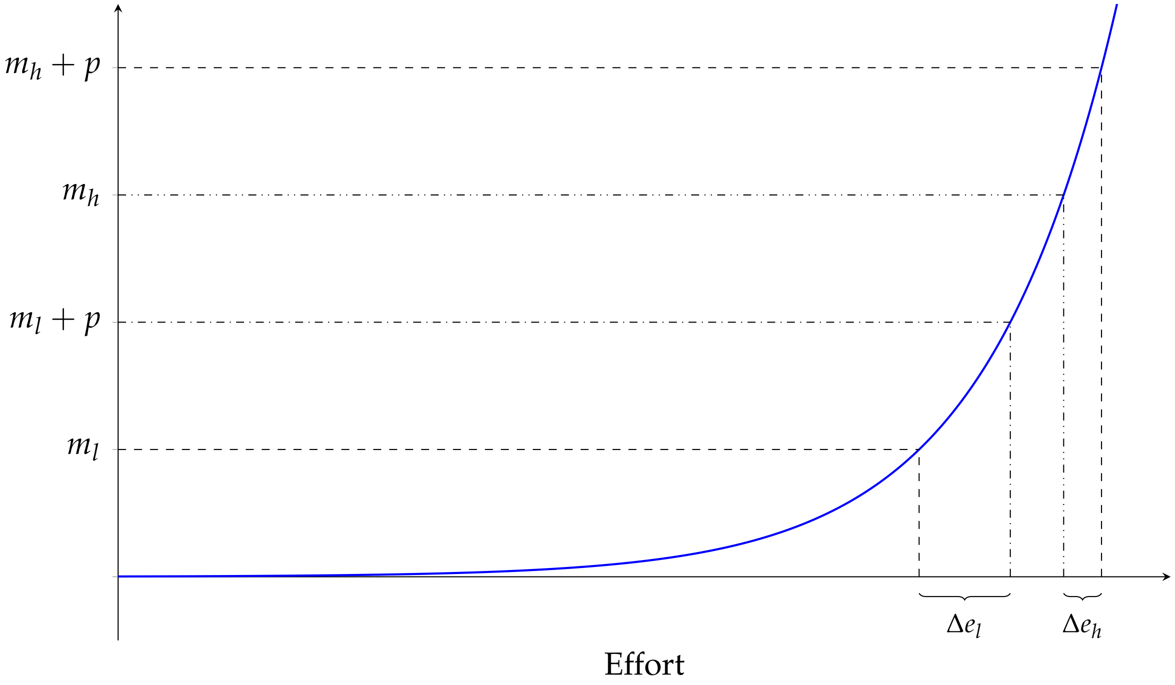
Table 3: Treatment effects on questions attempted and questions correct

	U.S.			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
(Std. error)	(0.017)	(0.016)	(0.028)	(0.008)	(0.002)	(0.017)
[*p*-value]	[0.057]	[0.324]	[0.023]	[0.001]	[0.022]	[0.017]
Control mean	0.807	0.962	0.640	0.970	0.998	0.940
(Std. deviation)	(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
(Std. error)	(0.012)	(0.013)	(0.019)	(0.007)	(0.008)	(0.009)
[*p*-value]	[0.004]	[0.017]	[0.119]	[0.127]	[0.801]	[0.012]
Control mean	0.515	0.494	0.549	0.852	0.856	0.848
(Std. deviation)	(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	0.054	0.030	0.079	-0.013	-0.007	-0.020
(Std. error)	(0.013)	(0.015)	(0.019)	(0.014)	(0.011)	(0.021)
[*p*-value]	[0.001]	[0.086]	[0.002]	[0.338]	[0.605]	[0.605]
Control mean	0.416	0.475	0.351	0.827	0.854	0.797
(Std. deviation)	(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 U.S. 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 (forthcoming) 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 U.S., and a wave fixed effect in Shanghai.

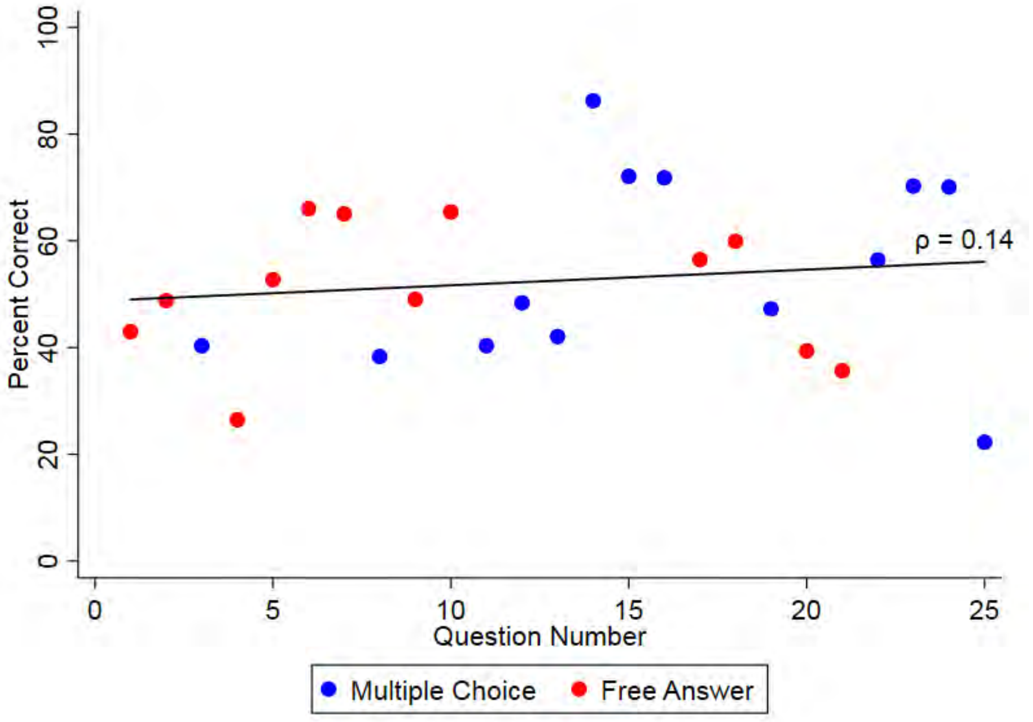
A Appendix Figures and Tables

Figure A.1: Marginal Benefits and Cost Curves



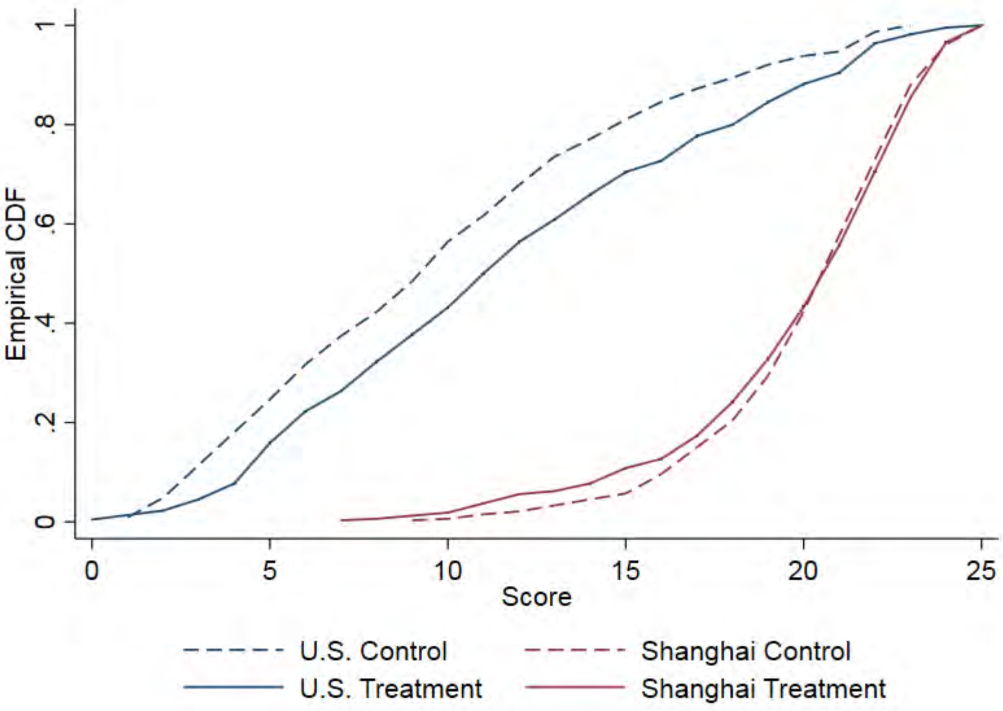
The figure plots the determination of effort at the intersection of marginal benefits (dashed lines) and marginal cost (solid line). Following DellaVigna and Pope (2018), a student chooses effort, e, to maximize: $\max_{e \geq 0} (m+p)e - c(e)$ where m is intrinsic motivation (i.e., motivation absent extrinsic incentives), p is the extrinsic incentive and $c(e)$ is the cost of effort with $c'(e) > 0$ and $c''(e) > 0$. We compare the changes in effort Δe in response to the same extrinsic incentive p for a student with high intrinsic motivation, m_h, compared to a student student with low intrinsic motivation, m_l, where $m_h > m_l$. Due to the convexity of the cost function, the student with low intrinsic motivation will experience larger effort responses than the student with high intrinsic motivation, $\Delta e_l > \Delta e_h$

Figure A.2: PISA worldwide percentage correct



The figure plots 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.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`. The percentage correct ranges from 25.7 to 87.3. The correlation between question difficulty and question position on the test is $\rho = 0.14$.

Figure A.3: Distribution of test scores, by treatment group



Tests of equality of Treatment and Control distributions within country: U.S. $p < 0.01$, Shanghai $p = 0.6869$. We estimate p-values using the following non-parametric permutation tests of differences between the test score distributions in each country. We construct test statistics using permutation methods based on Schmid and Trede (1996) and run one-sided tests for stochastic dominance and separatedness of the distributions (see also Imas, 2014). The test statistic identifies the degree to which one distribution lies to the right of the other, and takes 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.

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

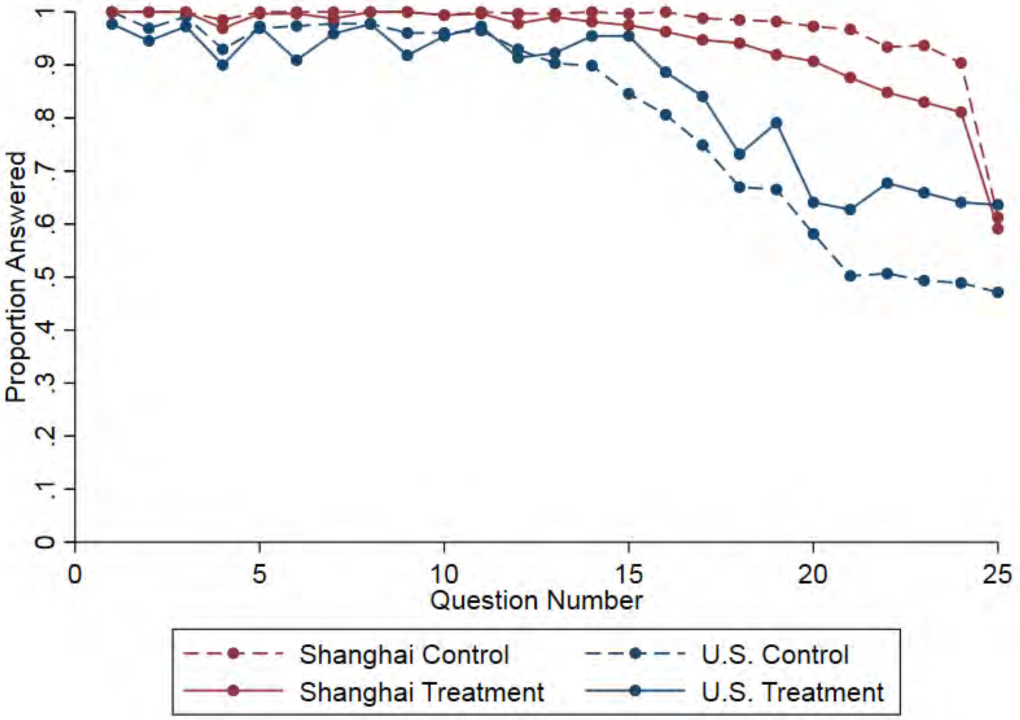


Table A.1: Sensitivity of U.S. treatment effect to sample changes

	Main sample	drop if missing age	keep non-10th	keep ELL	Control for baseline exam score
	(1)	(2)	(3)	(4)	(5)
Treatment	1.34	1.34	1.37	1.37	1.38
(Std. error)	(0.34)	(0.33)	(0.32)	(0.33)	(0.62)
[*p*-value]	[0.001]	[0.001]	[0.001]	[0.001]	[0.007]
Control mean	10.22	10.22	9.59	9.91	11.09
(Standard dev.)	(5.64)	(5.64)	(5.58)	(5.69)	(5.71)
School-track FE	Yes	Yes	Yes	Yes	Yes
Covariates	Yes	Yes	Yes	Yes	Yes
Students	447	446	534	469	348
Clusters	133	132	132	135	123

Notes: OLS estimates of equation (1). Robust standard errors clustered by class (except U.S. school 2, which was 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 (forthcoming).

Table A.2: Treatment effects by predicted test score: Threshold regressions, 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
(Std. error)	(0.59)	(0.73)	(0.48)	(0.025)	(0.023)
[*p*-value]	[0.217]	[0.010]	[0.049]	[0.288]	[0.052]
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	31	120	133	31	121

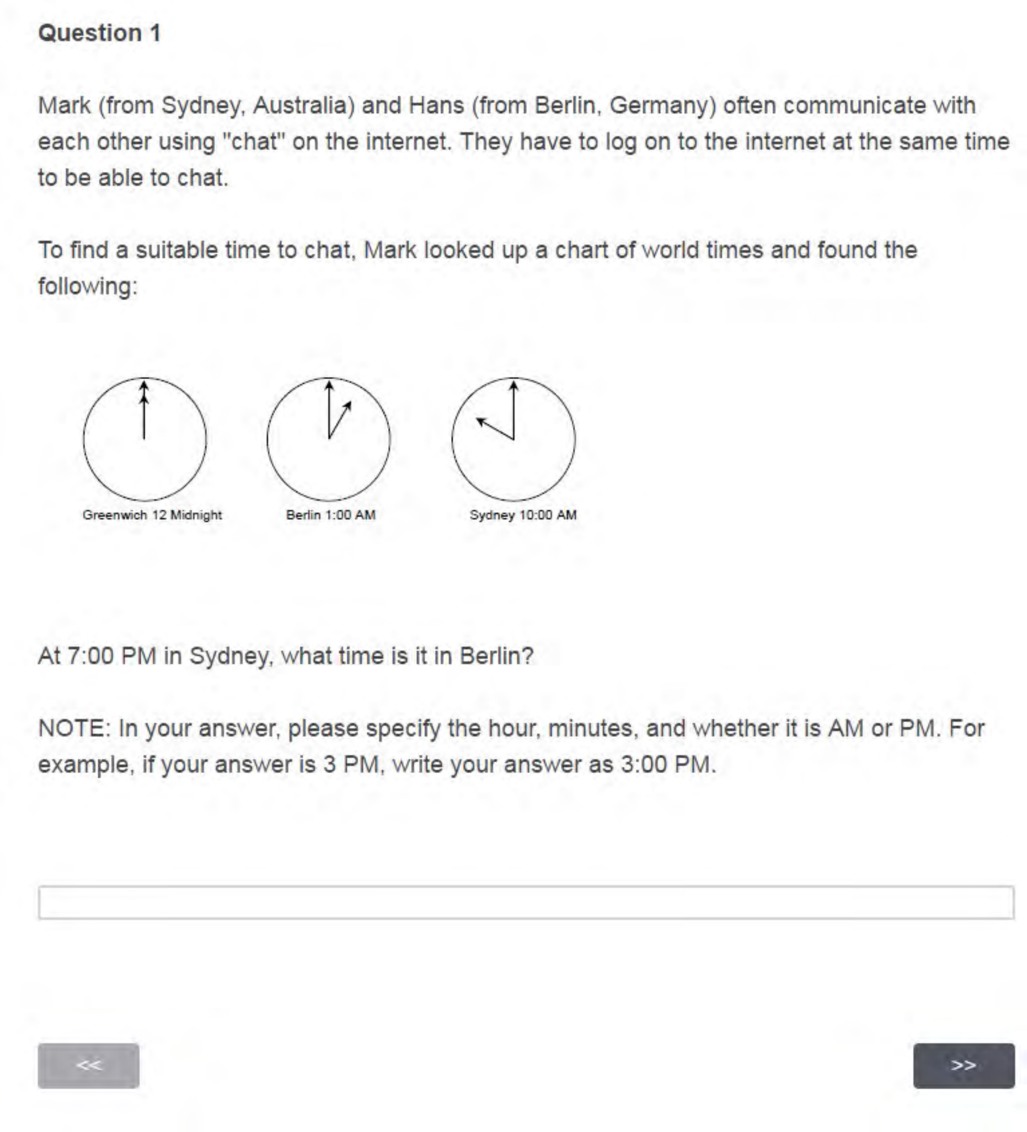
Notes: The table reports results from threshold regressions where the number of thresholds is estimated by minimizing the Bayesian Information Criterion and the threshold is the value of γ that minimizes the sum of squared residuals $S_{T1}(\gamma) = \sum_{t=1}^{T}\{Y_{ics} - (\alpha^1 + \beta_1^1 Z_c + \beta_2^1 X_i)I((-\infty < w_{ics} \leq \gamma) - (\alpha^2 + \beta_1^2 Z_c + \beta_2^2 X_i)I(\gamma < w_{ics} < \infty)\}^2$. Robust standard errors clustered by class (except U.S. school 2, which was randomized at the individual level) in parentheses. *p*-values in brackets. Inference 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 1-2 and columns 4-5. Inference in column 3 is based on a randomization test using the procedure of Young (forthcoming). All columns include the following covariates: age, gender, race/ethnicity. School-track fixed effects are not controlled for because they are collinear with predicted score.

Table A.3: Effect of incentives on test scores, by gender

	U.S.		Shanghai	
	Male	Female	Male	Female
	(1)	(2)	(3)	(4)
Treatment	1.67	0.97	0.07	-0.41
(Std. error)	(0.55)	(0.37)	(0.38)	(0.34)
[*p*-value]	[0.019]	[0.019]	[0.870]	[0.477]
Control mean	10.36	10.08	20.51	20.50
(Std. deviation)	(5.97)	(5.31)	(2.93)	(2.96)
Observations	226	221	308	348
Clusters	86	71	179	213

Notes: OLS estimates of equation (1). Robust standard errors clustered by class (except U.S. school 2 and Shanghai schools visited in 2018, which were randomized at the individual level) in parentheses. *p*-values in brackets. We adjust *p*-values within each country 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). All columns include school-track fixed effects and the following covariates: age, race/ethnicity (U.S. only), and a wave fixed effect (Shanghai only).

B Test Questions



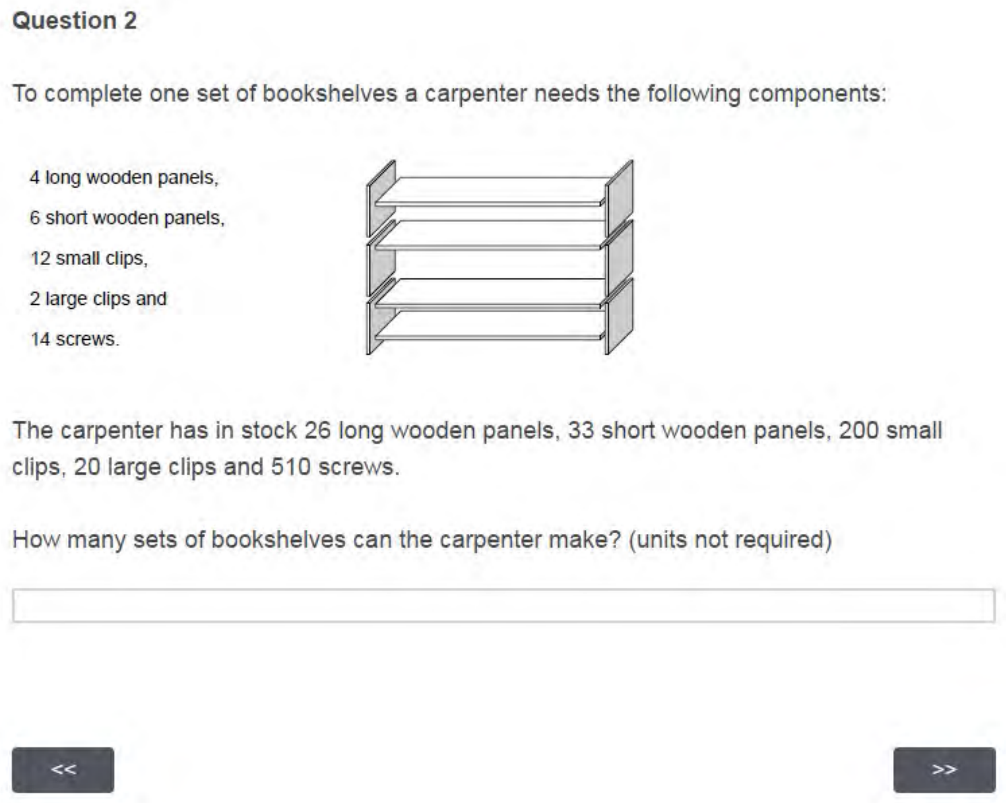
Question 1

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,

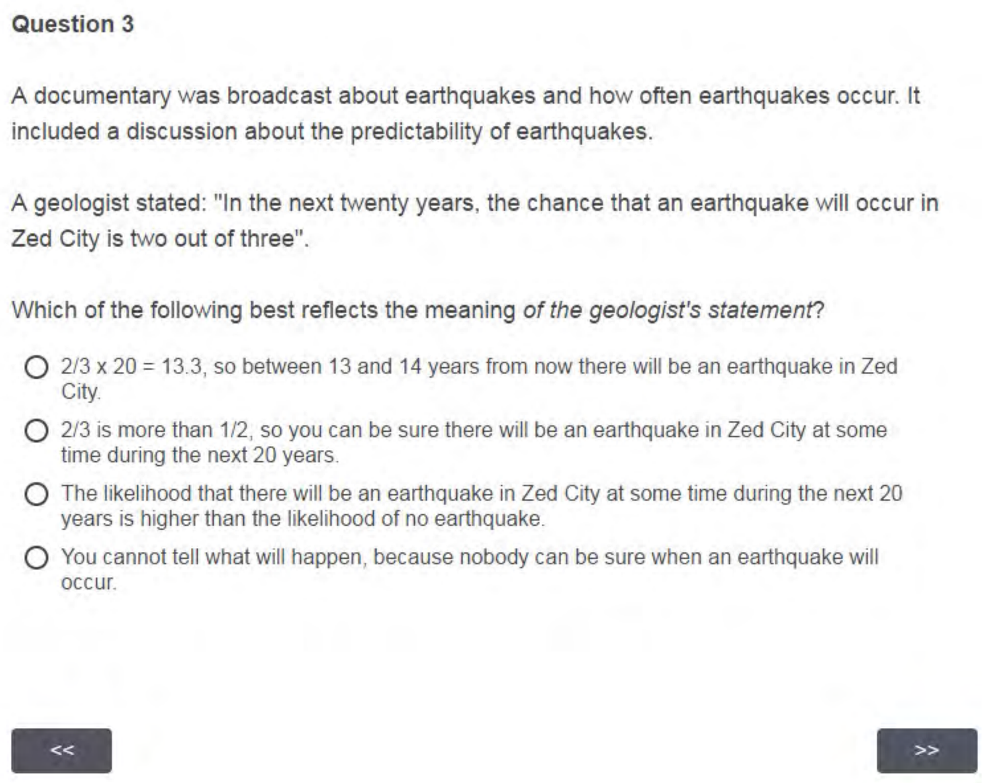
12 small clips,

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".

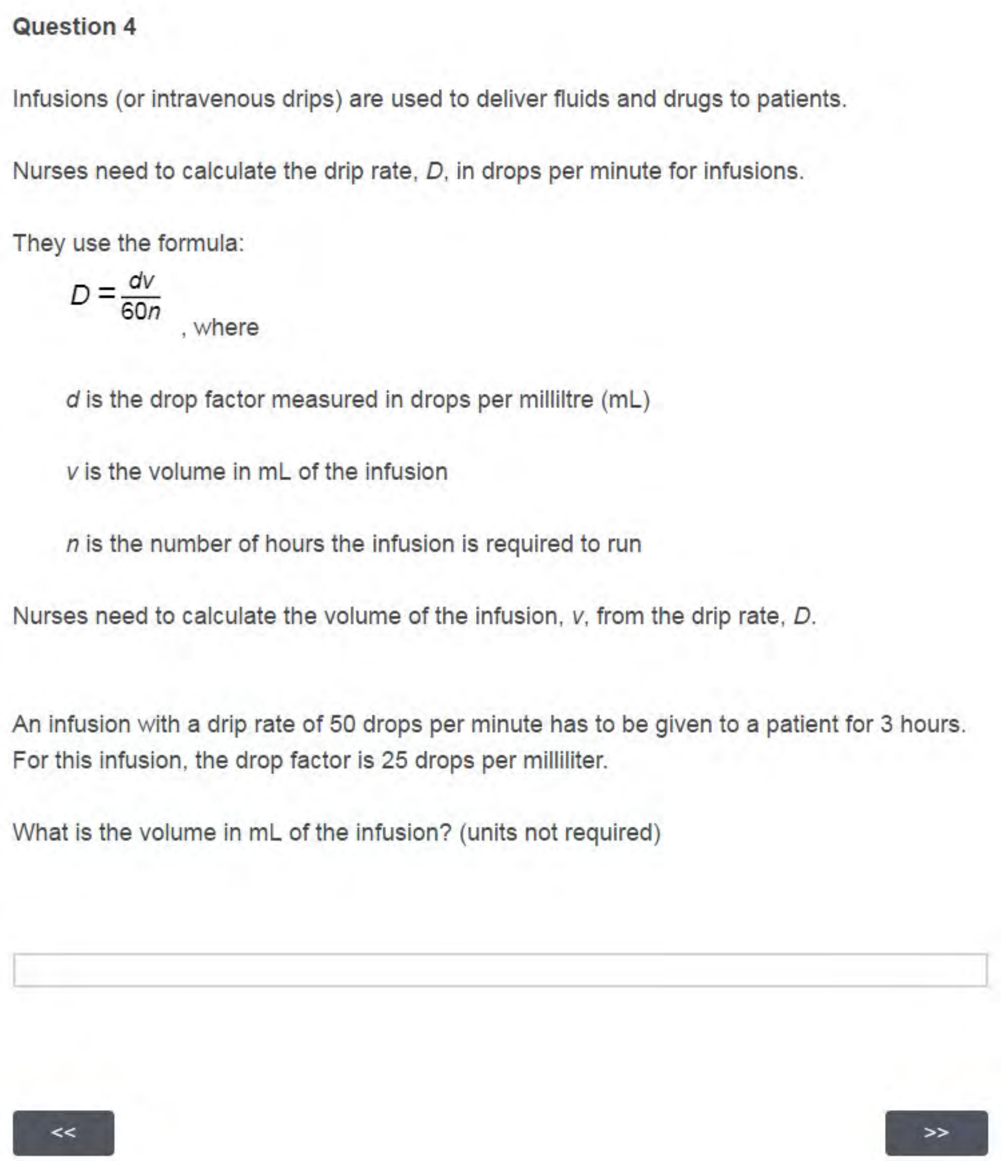
Which of the following best reflects the meaning *of the geologist's statement*?

○ 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)

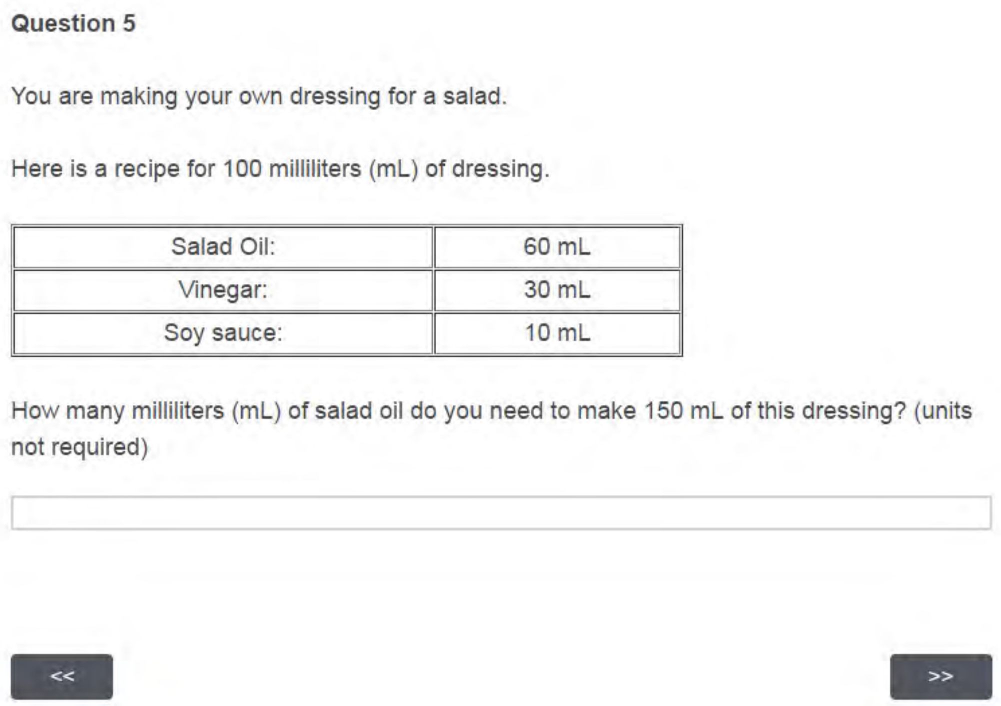
v is the volume in mL of the infusion

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)



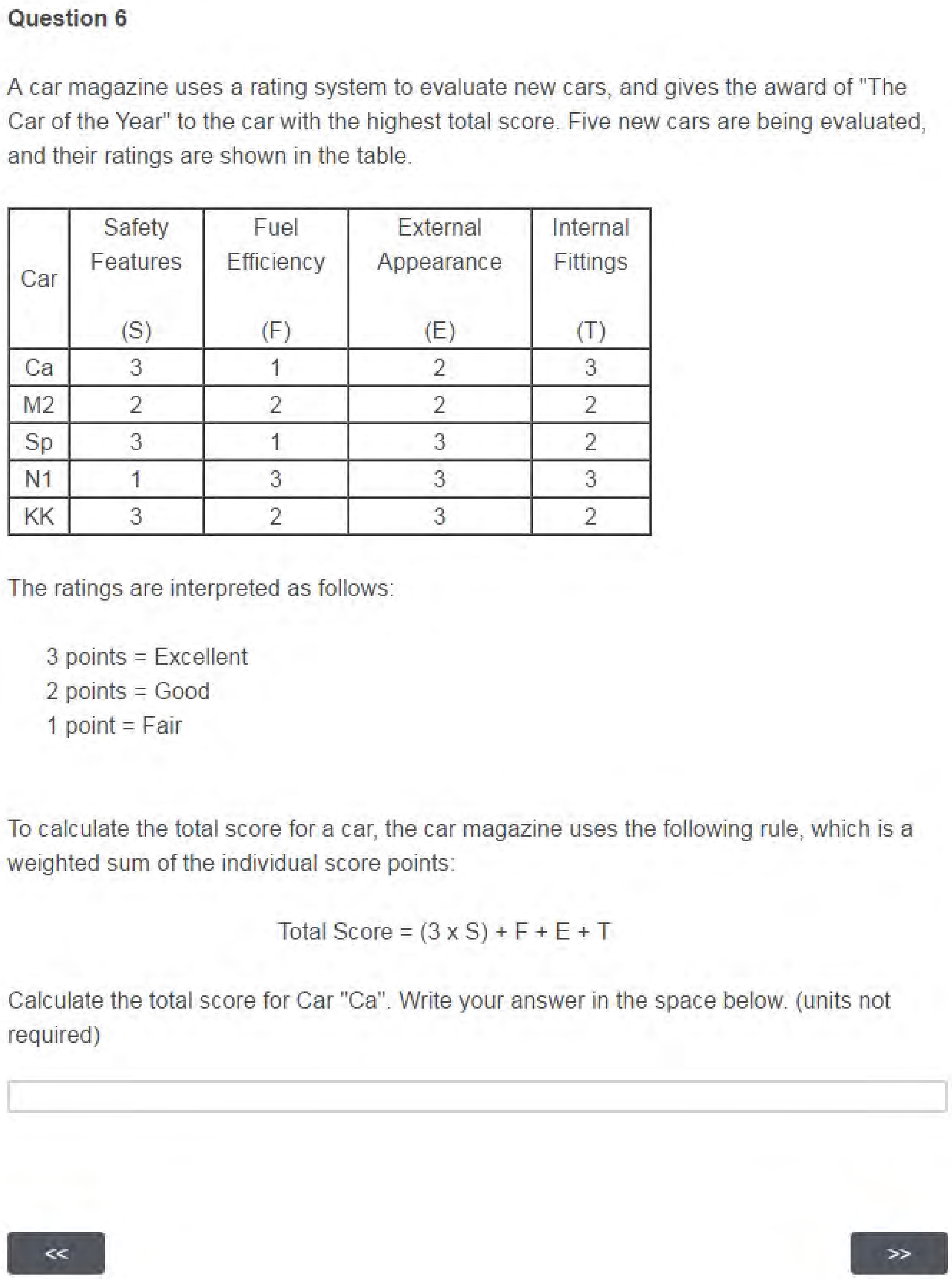
Question 5

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

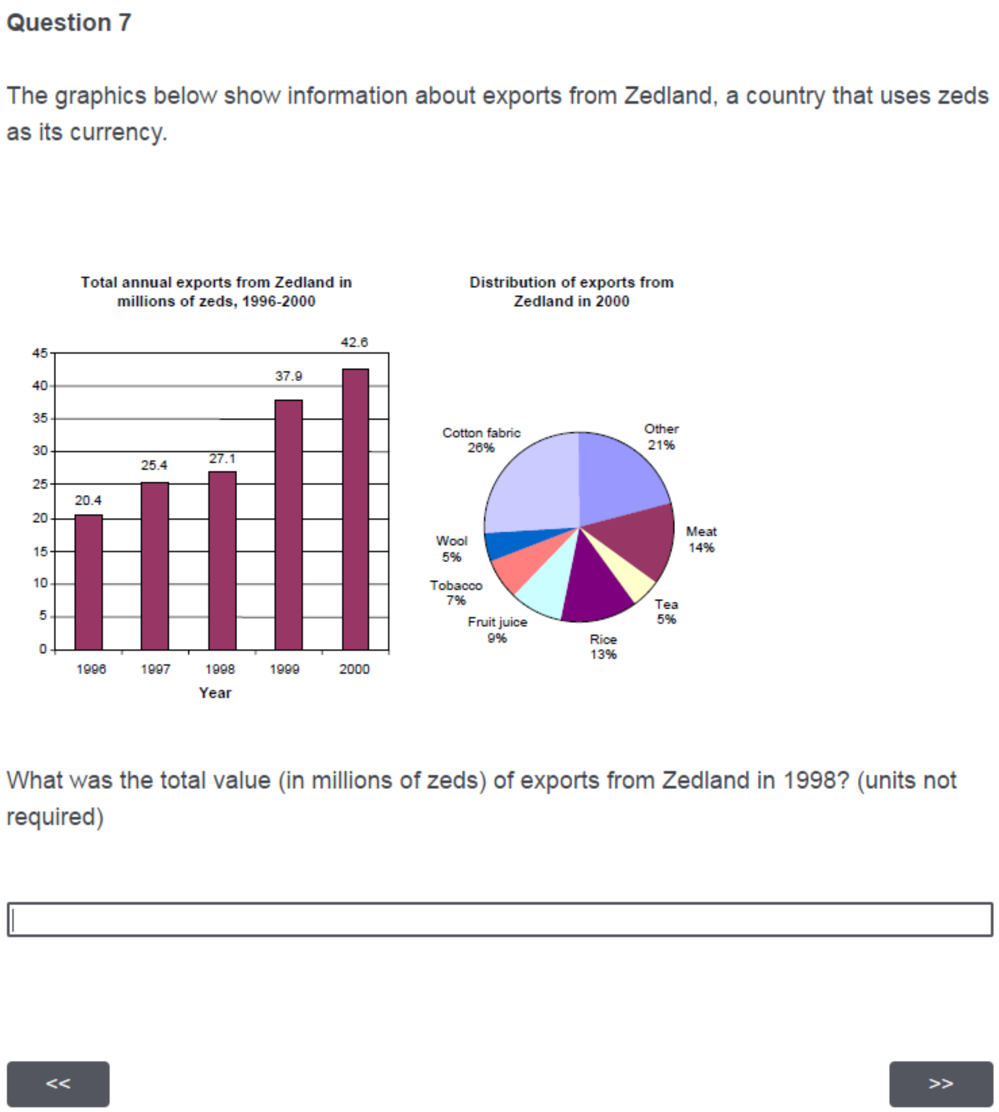
The ratings are interpreted as follows:

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:

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

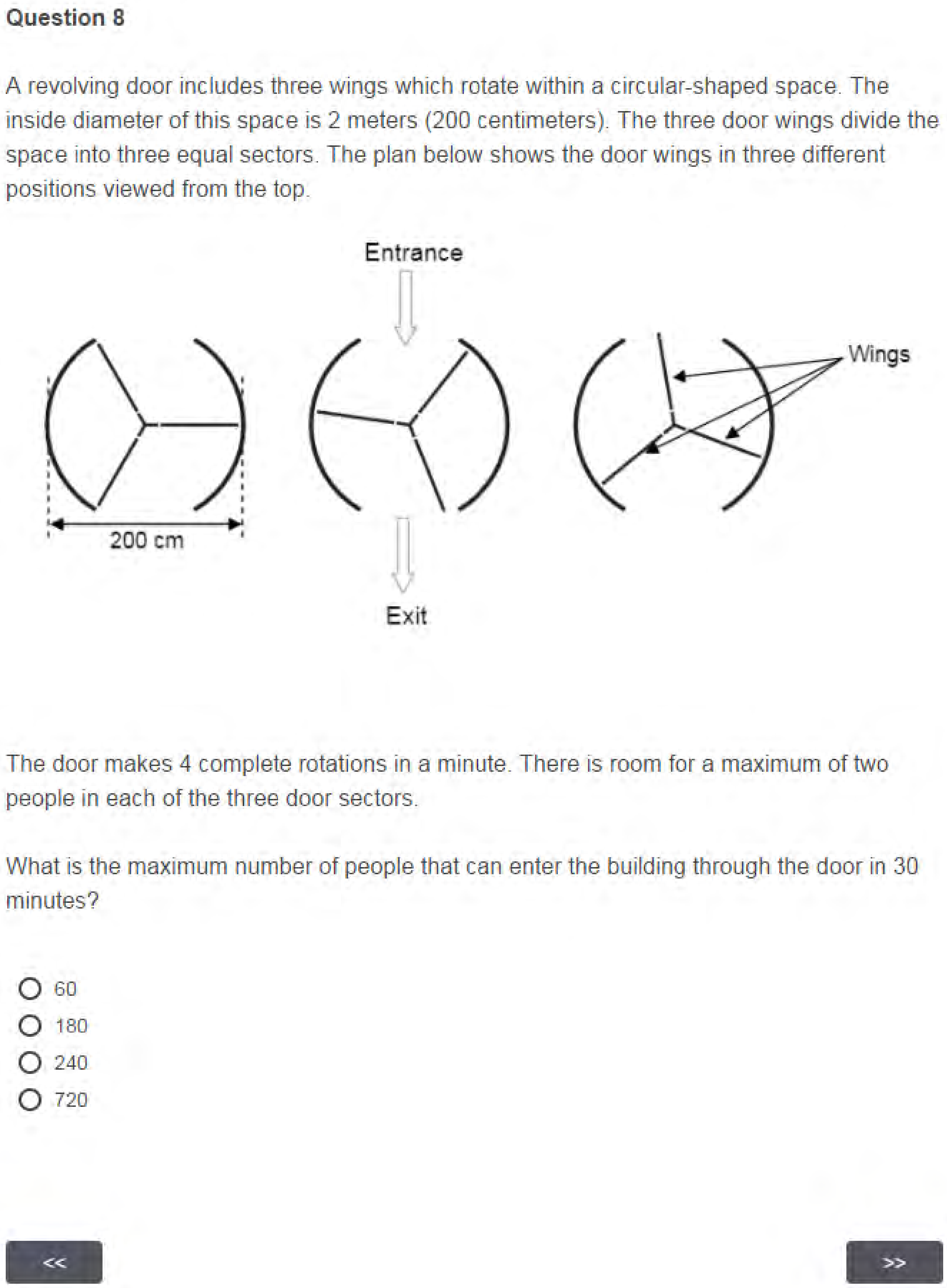
Calculate the total score for Car "Ca". Write your answer in the space below. (units not required)



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)



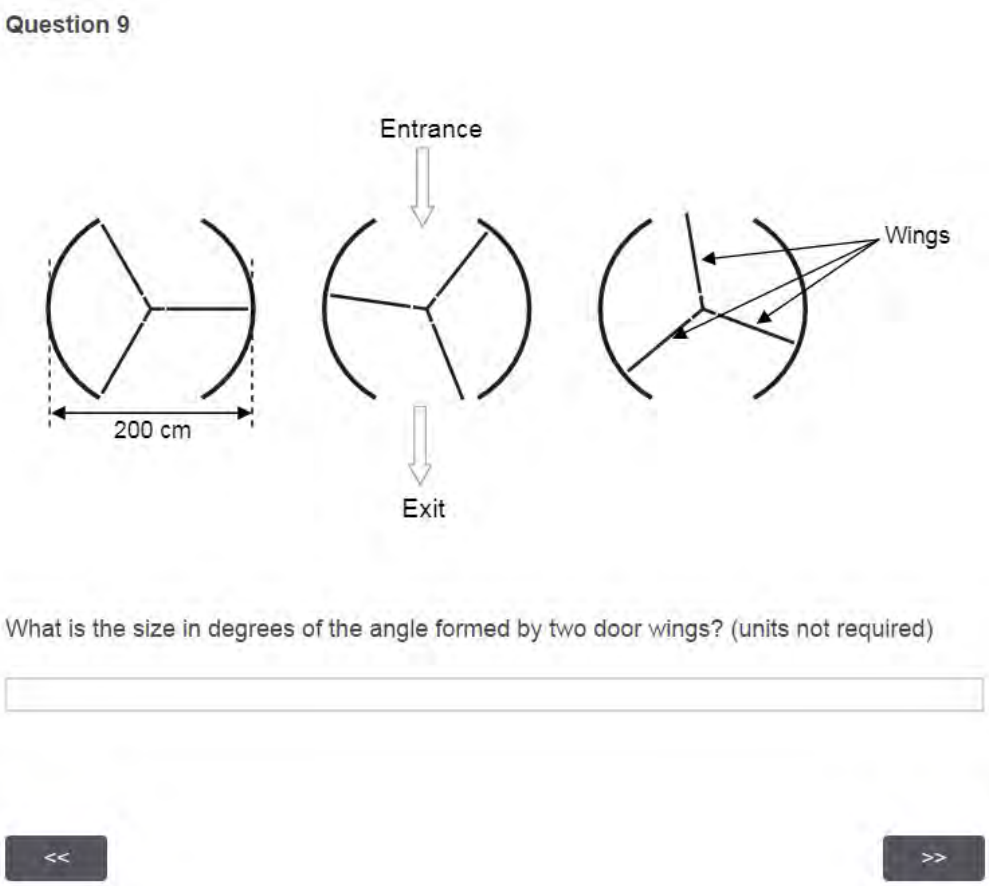
Question 8

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.

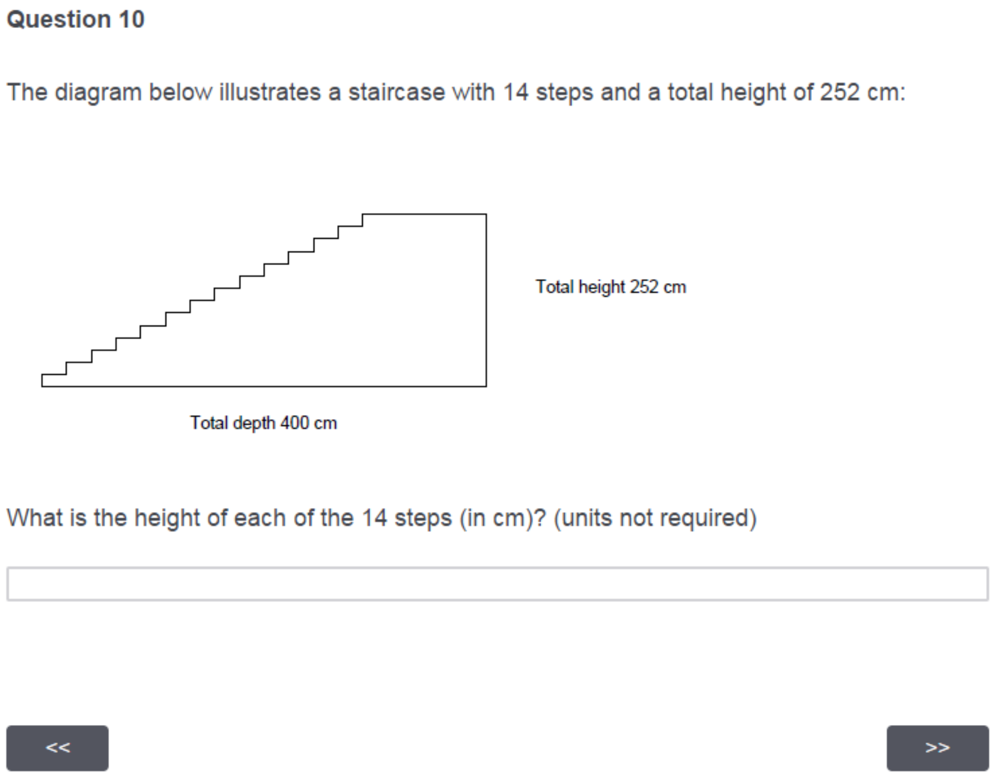
What is the maximum number of people that can enter the building through the door in 30 minutes?

○ 60
○ 180
○ 240
○ 720



Question 9

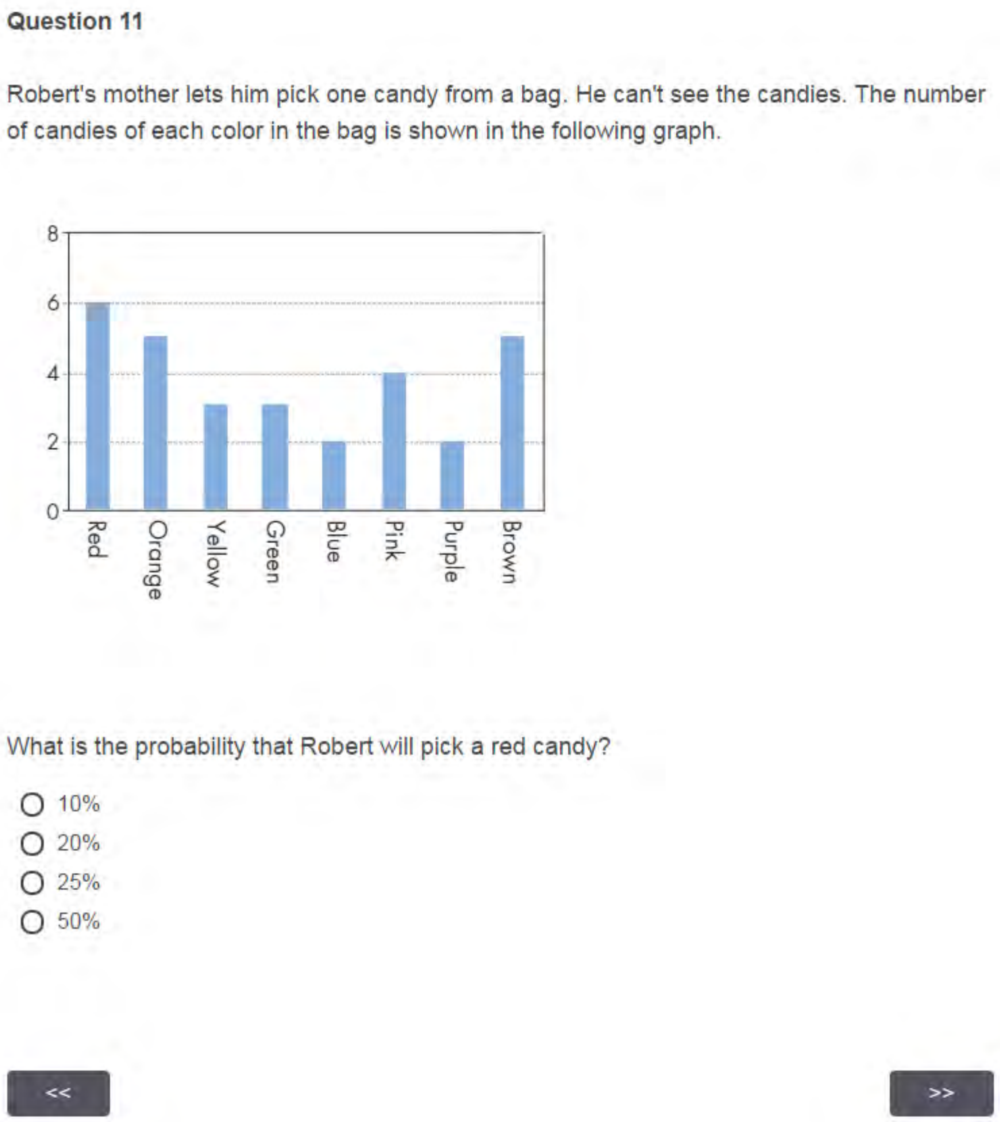
What is the size in degrees of the angle formed by two door wings? (units not required)



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)

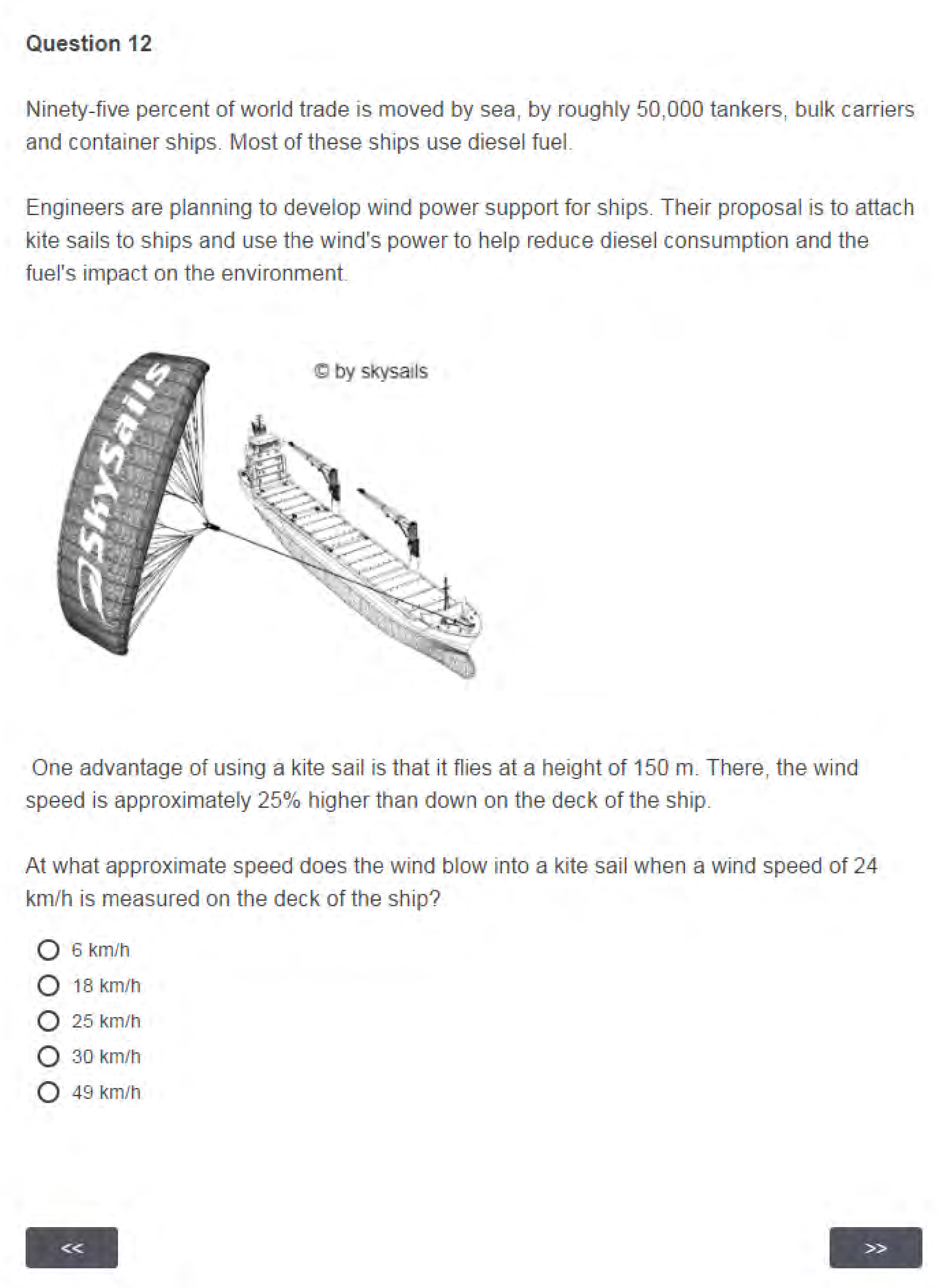


Question 11

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.

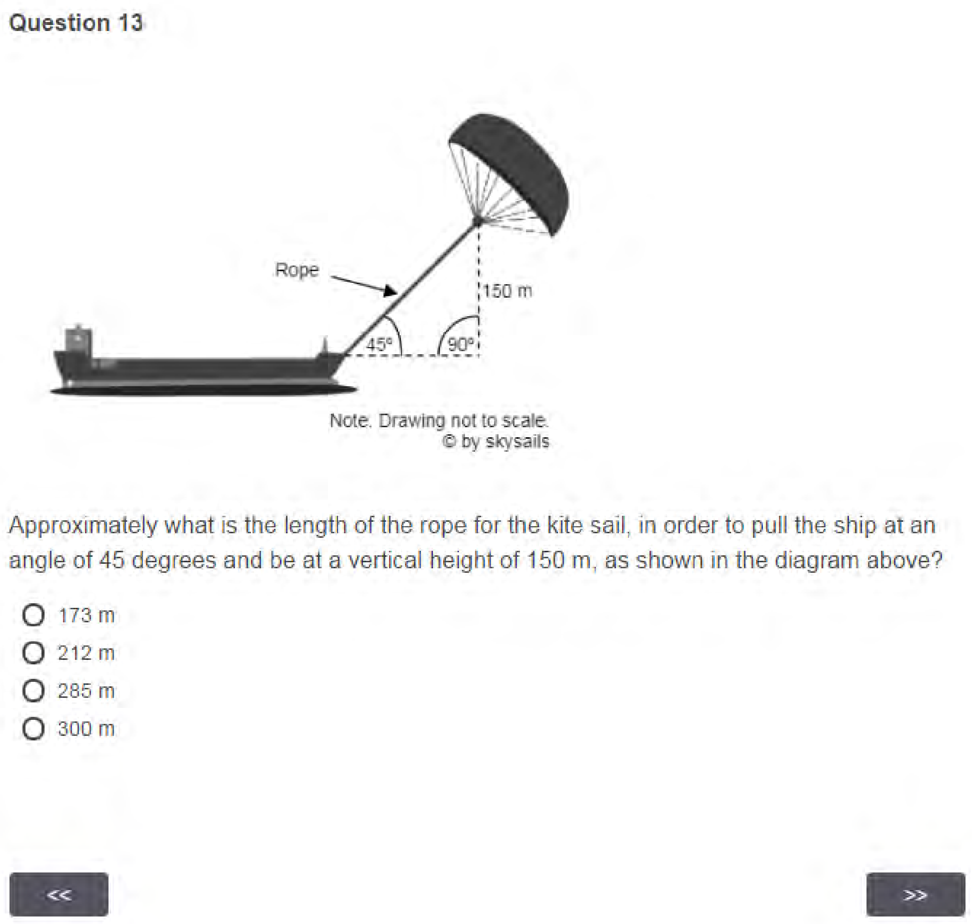
Engineers are planning to develop wind power support for ships. Their proposal is to attach kite sails to ships and use the wind's power to help reduce diesel consumption and the fuel's impact on the environment.

© 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

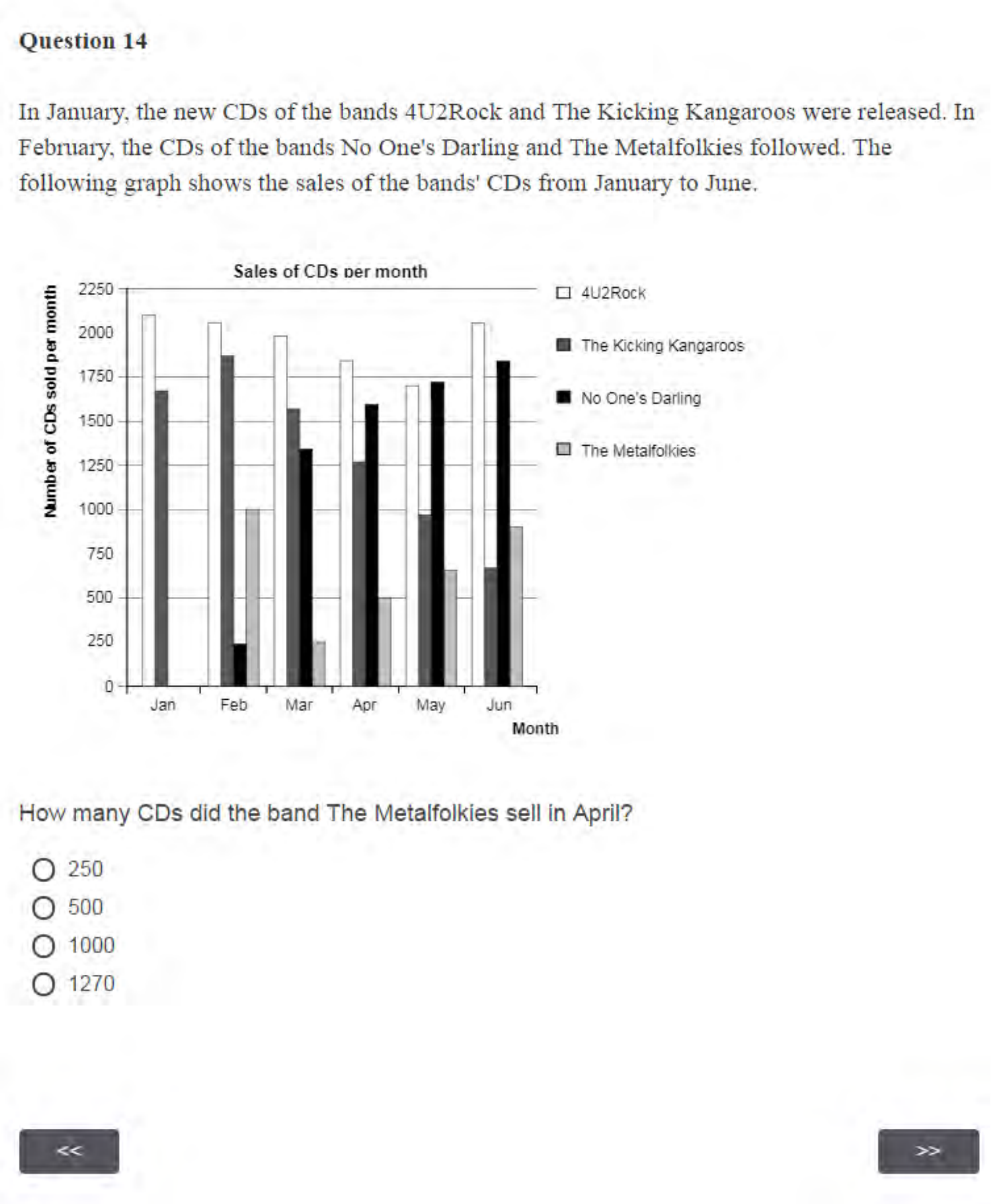
Approximately what is the length of the rope for the kite sail, in order to pull the ship at an angle of 45 degrees and be at a vertical height of 150 m, as shown in the diagram above?

○ 173 m

○ 212 m

○ 285 m

○ 300 m

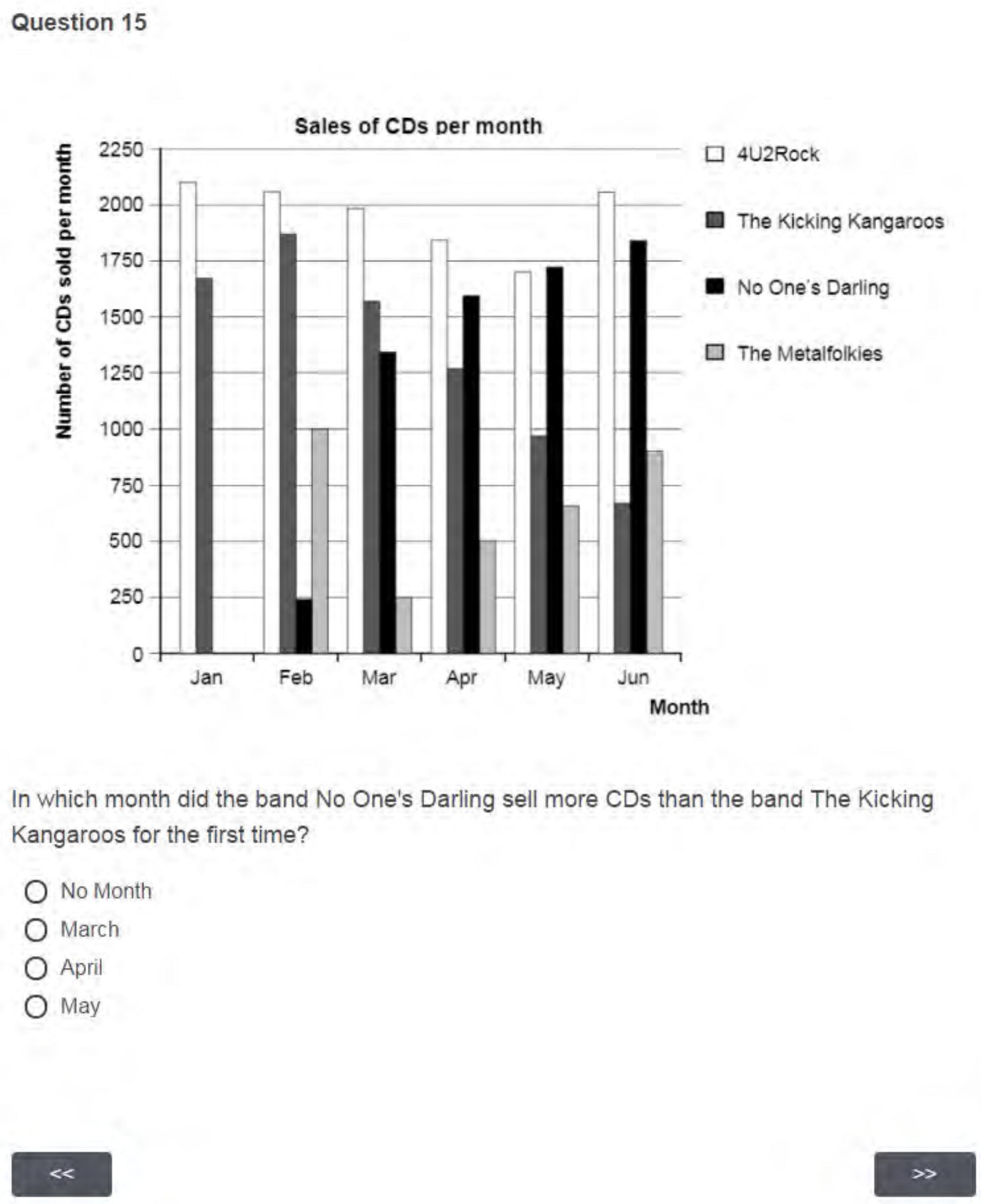


Question 14

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?

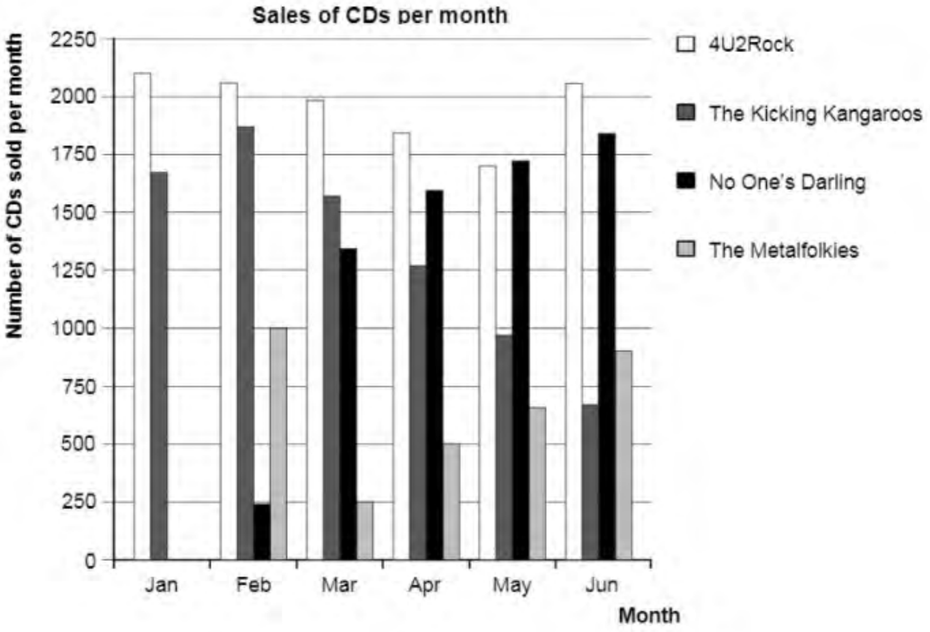
○ No Month

○ 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?

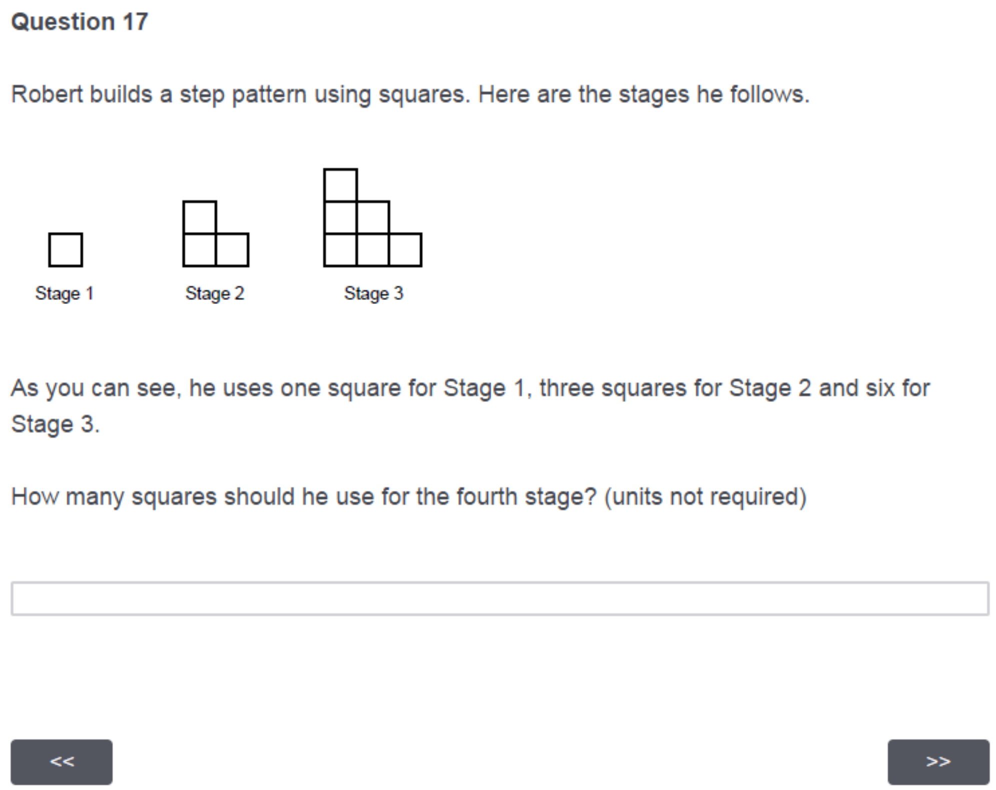
○ 70 CDs

○ 370 CDs

○ 670 CDs

○ 1340 CDs

<< >>

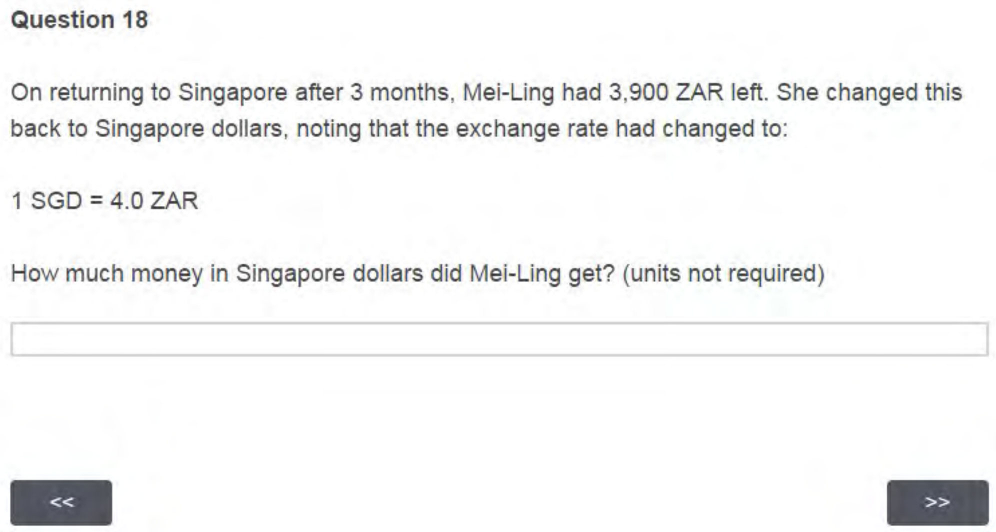


Question 17

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)

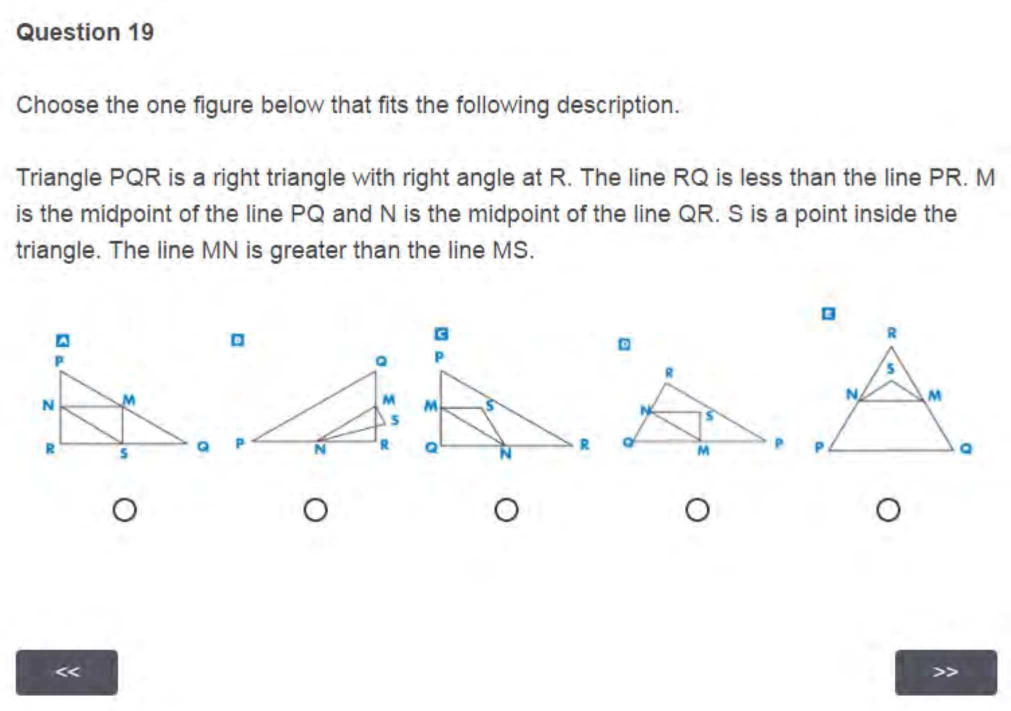


Question 18

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)

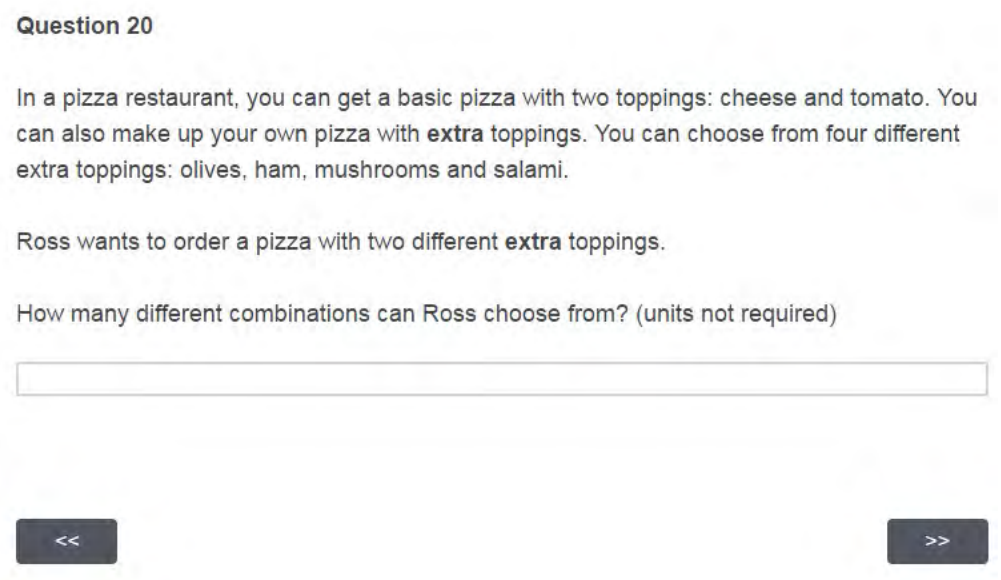


Question 19

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.

○ ○ ○ ○ ○

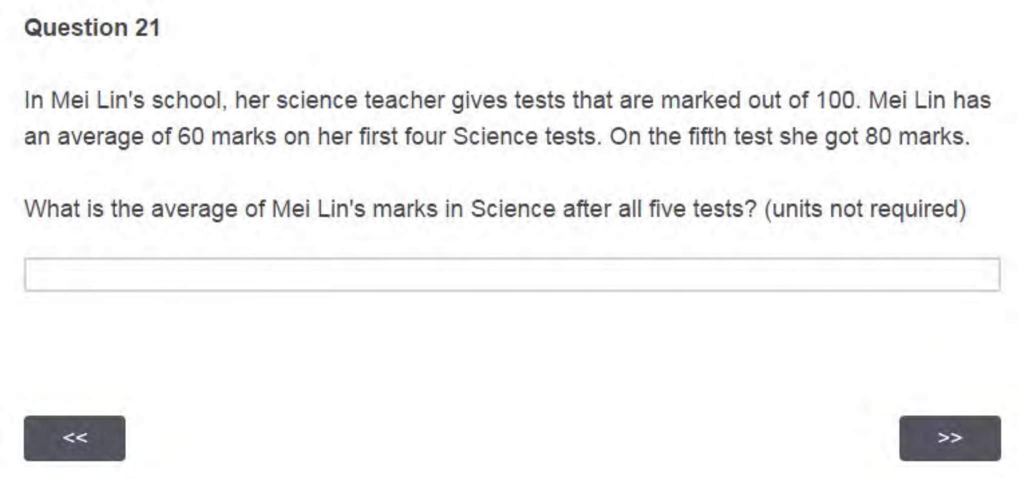


Question 20

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.

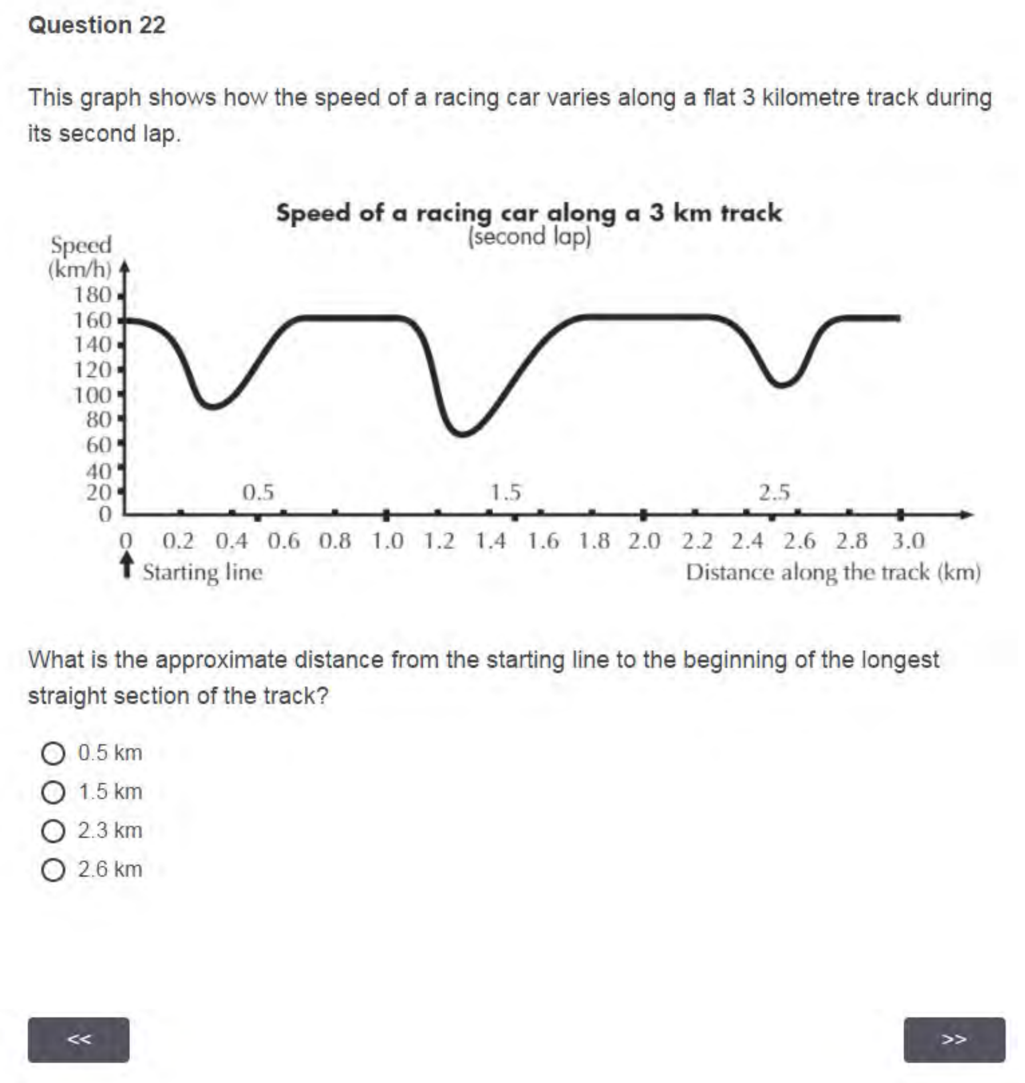
How many different combinations can Ross choose from? (units not required)



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)

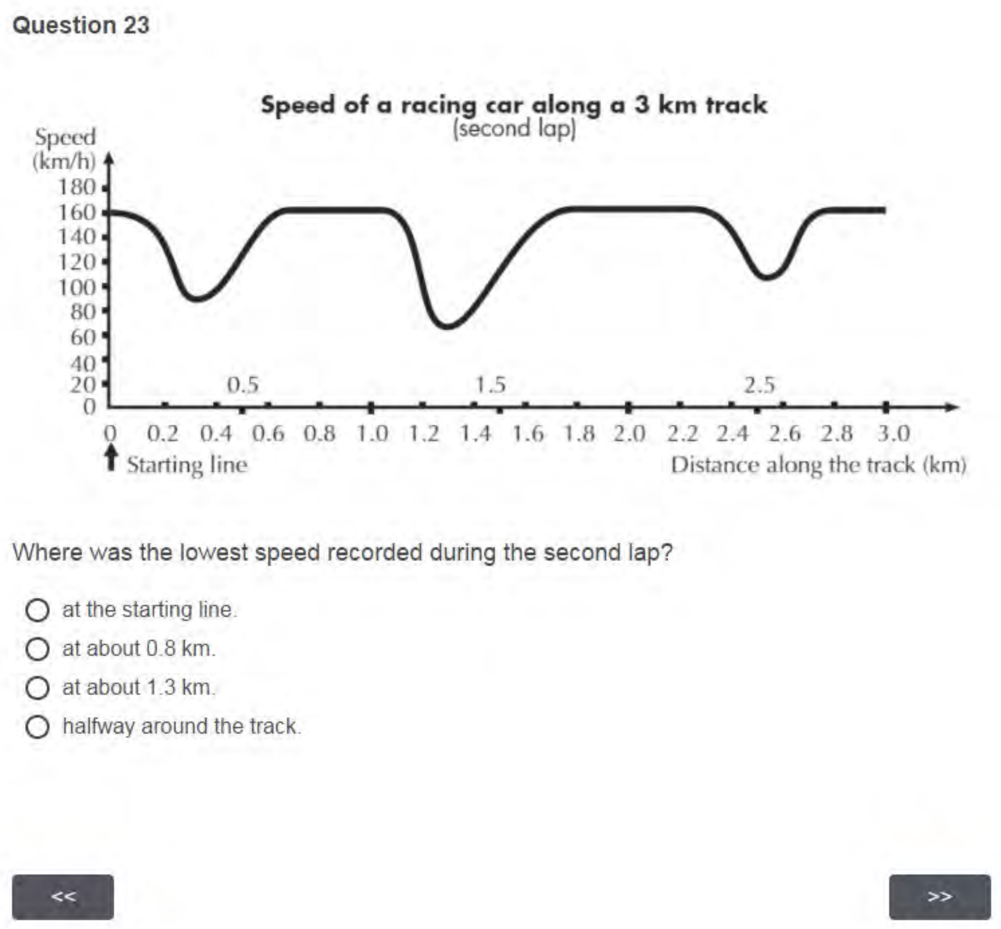


Question 22

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

Speed of a racing car along a 3 km track
(second lap)

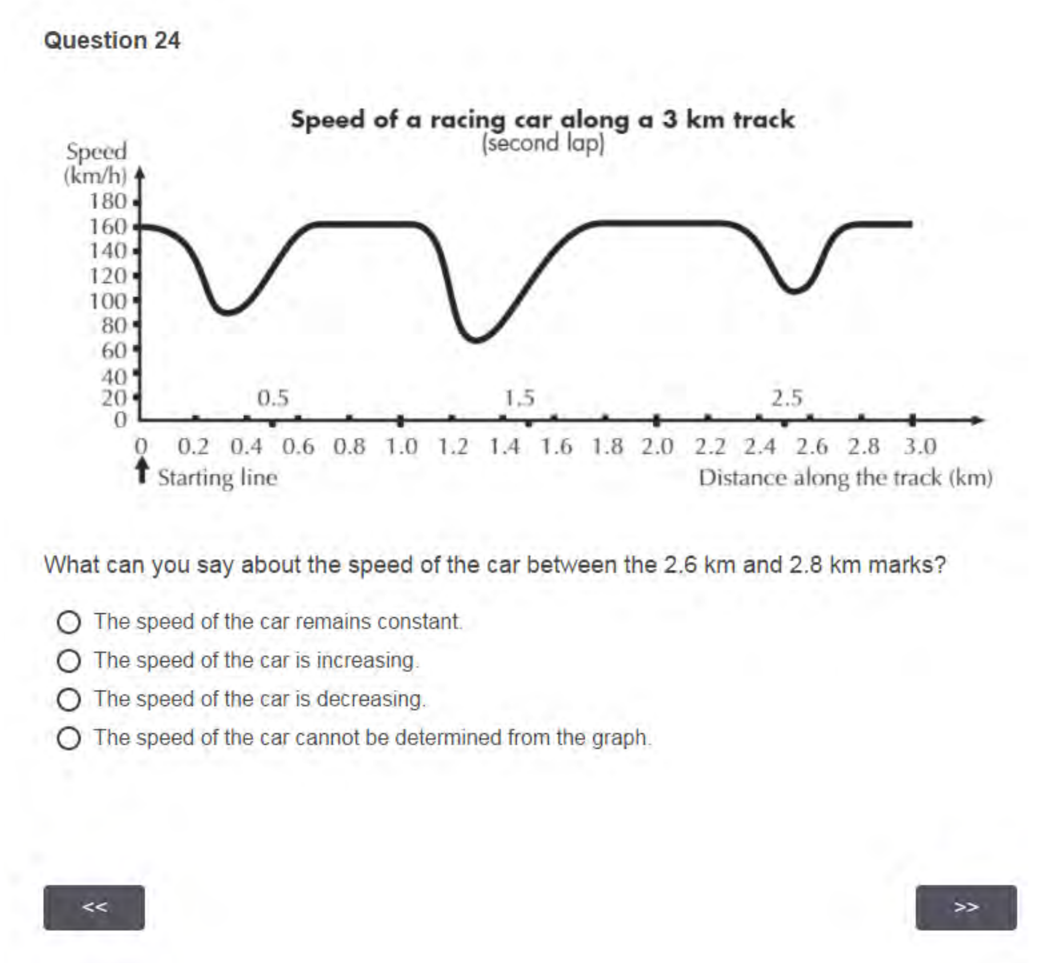
Where was the lowest speed recorded during the second lap?

○ at the starting line.

○ at about 0.8 km.

○ at about 1.3 km.

○ halfway around the track.



Question 24

Speed of a racing car along a 3 km track
(second lap)

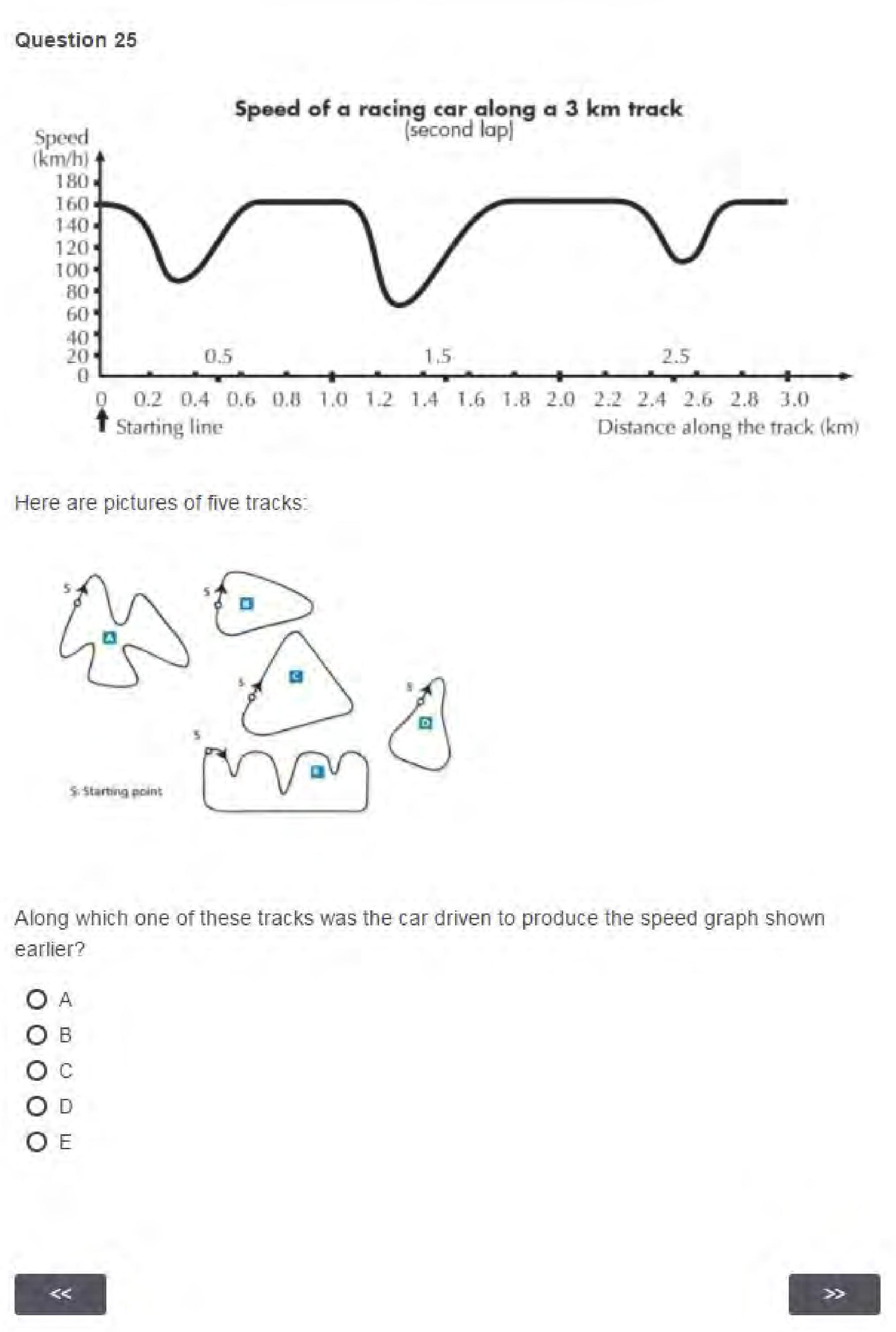
What can you say about the speed of the car between the 2.6 km and 2.8 km marks?

○ 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

C Instructions

C.1 Control Instructions

Hello and thank you for participating in our study. Today you will be asked to complete a 25-question math quiz. Your payment today will not depend on your performance on the quiz.

You may use the pen, paper and calculator provided, but all answers must be entered on the computer. You will have 25 minutes to complete the math quiz. Do not start the quiz until you are told to do so, and stop when you are asked to. During the quiz you can go back and change your answers, but after the last question your answers will be submitted and you will be unable to change them.

If you have questions about these instructions, please raise your hand and a test administrator will come to you. However, once the exam begins, you will be unable to ask questions until it is over.

When you are finished, the final screen will show you your score on the quiz and the ID number that you were assigned. When called upon to do so, please write your score and ID number on a piece of paper and bring this piece of paper to the front of the room so we can make sure your score is recorded correctly.

Please enter your ID number below and press the button at the bottom right of the screen to continue.

C.2 Treatment Instructions

Hello and thank you for participating in our study. Today you will be asked to complete a 25-question math quiz. Your payment today will depend on your performance on the quiz. You are being given an envelope that contains $25. Please open the envelope to make sure there is $25 inside. Then write your ID number that was assigned to you on the outside of the envelope.

While this money is yours, we will take away $1 for each question you answer incorrectly. Unanswered questions count as questions answered incorrectly. At the end of the quiz, we will subtract the number of questions you answered incorrectly from 25 and that will be your final payment.

Please sign the form that says this is your $25, but that you may have to give some back depending on your score on the quiz.

You may use the pen, paper and calculator provided, but all answers must be entered on the computer. You will have 25 minutes to complete the math quiz. Do not start the quiz until you are told to do so, and stop when you are asked to. During the quiz you can go back and change your answers, but after the last question your answers will be submitted and you will be unable to change them.

If you have questions about these instructions, please raise your hand and a test administrator will come to you. However, once the exam begins, you will be unable to ask questions until it is over.

When you are finished, the final screen will show you your score on the quiz along with the ID number that you were assigned. When asked to do so, write your score and ID number on a piece of paper and bring it to the test administrators at the front of the room so we can make sure your score is recorded correctly.

Please enter your ID number below and press the button at the bottom right of the screen to continue.