char·asse·in
THE CHARACTER ASSESSMENT INITIATIVE

The role of student effort on performance in PISA: Revisiting the gender gap in achievement

Lina Anaya*
University of Arkansas

Gema Zamarro
University of Arkansas

February, 2021

*Corresponding Author: Lina Anaya, University of Arkansas, 211 Graduate Education Building, Fayetteville, Arkansas, 72701, email: lanaya@aurk.edu

Abstract:

International assessments are important to benchmark the quality of education across countries. However, on low-stakes tests, students' incentives to invest their maximum effort may not be optimal. Research stresses that ignoring students' effort when interpreting results from low-stakes assessments can lead to biased interpretations of test performance across groups of examinees. We use data from the Programme for International Student Assessment (PISA), a low-stakes test, to analyze the extent to which student effort helps to explain test scores heterogeneity across countries and by gender groups. We build two measures of effort based on the response time of questions (i.e., rates-guessing rates in the test) and on the effort of the survey that students take after the test (i.e., item non-response rates). Our results highlight the importance of accounting for differences in student effort to understand cross-country heterogeneity in performance and variations in gender achievement gaps across nations. We find that, once we account for differential student effort across gender groups, the estimated gender achievement gap in math and science could be up to 36 and 40 percent of a standard deviation wider, respectively, and up to 39 percent of a standard deviation narrower in reading, in favor of boys. In math and science, the gap widens in 50 and 45 out of 55 countries, respectively. Altogether, our effort measures on average explain between 43 and 48 percent of the cross-country variation in test scores.

JEL Codes: I20, J16, C83

Keywords: Student Effort, Gender Gaps, Rapid guessing, PISA 2015.

1. Introduction

Understanding how well a school or an educational system educates its students is important for stakeholders such as parents, teachers, and governments. Standardized assessments help policymakers to benchmark the quality of schools or a country's educational system relative to other nations. However, when students do not face the consequences for high or low performance, their incentives to invest their maximum effort on the test may not be optimal. Thus, differences in test performance may not just reflect variations in actual content knowledge but also differences in other non-content-knowledge factors, such as student effort. One such example is low-stakes international assessments, such as PISA (Programme for International Student Assessment) or TIMSS (Trends in International Mathematics Science Study), in which differences in student effort may be essential for explaining part of the observed differences in student achievement across and within countries by gender.

Several studies find that ignoring student effort may lead to biased conclusions about the test performance of a group of examinees (Demars, 2007; Swerdzewski et al., 2011; Wise & DeMars, 2010; Wise & Kong, 2005). This problem can worsen when making international comparisons of achievement. Evidence from international assessments shows that student effort is essential to understand differences in test performance within and across countries (Boe et al., 2002; Debeer et al., 2014; Zamarro et al., 2019).

In this paper, we revisit this prior literature studying the role of effort in explaining differences in test scores to analyze the extent to which student effort contributes to explain variation in test performance in math, reading, and science, across countries, as well as within countries by gender. We use data from the PISA 2015 computer assessment and student computer-

based survey to construct measures of student effort based on the instances of rapid-guessing responses in the test and the effort students put forward in the survey (i.e. item non-response rates), respectively. Prior research from PISA suggests that student item non-response rates contribute to explain a significant part of the variation across countries in test scores (Zamarro et al., 2019).

To compute student rapid-guessing rates, we use the inverse response-time-effort (RTE) score as introduced by Wise & Kong (2005). Following Wise & Kong (2005), we use the information on response times for each question to calculate the proportion of questions of the assessment in which the examinee does not engage in solution behavior (i.e., the examinee does not take the time to analyze the question [Schnipke, 1995; Schnipke & Scrams, 1997]).

Differences in student effort could help explain differences in student performance across countries, as well as test score gender gaps within countries. Obtaining a better understanding of the role of effort on gender achievement gaps is important given women's underrepresentation in science occupations (Anaya et al., 2017; Ceci et al., 2014; Nix et al., 2015; Perez-Felkner et al., 2017).

If student effort varies by gender, differences in effort could affect our understanding of gender gaps in test performance. Along these lines, Balart & Oosterveen (2019) use measures of decline in performance throughout the PISA test and find that girls are better at sustaining test performance than boys. According to the authors, this result has consequences for the measurement of the gender achievement gap because in longer assessments, the gap in math and science is smaller compared to shorter assessments. Using data from the U.S., Soland (2018a, 2018b) obtains similar findings; the author measures effort based on response times of test questions and finds

that after removing the effect of effort in test scores, the gender gap in math achievement would be wider, and it is more sensitive to effort-adjustment than the reading gap.

We find evidence of significant variation of rapid-guessing behavior in PISA. In line with prior research, we find that student effort explains a significant part of the variation in PISA scores across countries. Altogether, our effort measures represent, on average, between 43 and 48 percent of the variation in test performance across countries. Also, the probability of engaging in rapid-guessing behavior is higher for boys than for girls, which has implications for estimated gender gaps in performance. Accounting for student effort affects the estimated gender gaps in achievement. We find that the gender achievement gap could be up to 36 and 40 percent of a standard deviation (SD) wider in science and math, respectively, and up to 39 percent of a standard deviation (SD) narrower in reading, in favor of boys.

The remaining parts of this document are organized as follows: Section 2 presents the literature review; section 3 explains in more detail the data we use in this study; section 4 describes the measures of student effort in PISA that we use in the paper; section 5 shows the methodology and results; section 6 illustrates some robustness checks, and section 7 presents our conclusions.

2. Literature review

Student motivation or effort is an essential element to understand student achievement in low-stakes assessments. Wise & DeMars (2005) define student motivation as the amount of effort or energy that a student invests towards achieving the highest possible score on a test. When students do not face consequences for performance, their incentives to invest their maximum effort on the test may not be optimal. As a result, ignoring the role of students' motivation in the interpretation of test scores may lead to biased conclusions given that the resulting scores may not

be an accurate indicator of students' ability (Kane, 2006; Swerdzewski et al., 2011; Wise & Kong, 2005).

A significant first step to take student effort into account when interpreting test scores is to identify who the low-effort examinees are. Researchers who analyze student effort using large representative samples from international assessments have developed several methods to calculate student effort using paper-based assessments such as the decline in performance based on the position of questions in the test, rate of decline in performance, careless answering patterns, and item non-response rates (Boe et al., 2002; Borghans & Schils, 2012; Debeer et al., 2014; Zamarro et al., 2019).

For example, Debeer et al. (2014) focuses on the reading achievement data from PISA 2009 and defines effort as the difference in test performance due to the different positions a group of questions occupies on the test. Similarly, Borghans and Schils (2012) employ the rate of decline in performance as the test progresses, while Zamarro et al. (2019) not only employ the rate of decline in performance but also measure the careless answering patterns and item non-response rates on the survey students take after the PISA 2009 test, in order to measure student effort. The authors find that item non-response in the survey has the highest predictive power in explaining differences in test scores across countries. Previous work also highlights the importance of item non-response rates, as a proxy for non-cognitive skills, to understand how differences in student effort can explain cross-country differences in achievement (Boe et al., 2002).

Computer-based assessments create a new opportunity for researchers to develop new measures of student effort. Wise and Kong (2005) propose using the response-time-effort (RTE) score, which focuses on examinees' response times in computer-based-low-stakes assessments, as

a proxy for motivation. This idea comes from Schnipke (1995) and Schnipke and Scrams (1997), who define solution behavior as the situation in which the examinee takes the time to analyze the question in order to find the right answer and, rapid-guessing behavior, when the examinee rapidly chooses a response.

Although in high-stakes evaluations, rapid-guessing may represent the hurry to answer all the questions, when examinees do not have enough time to complete the test using solution behavior (Schnipke, 1995; Schnipke & Scrams, 1997), Wise and Kong (2005) argue that in a low-stakes context, responses given within a short time represent students' low engagement in trying to find the right answer. As a result, the RTE score represents the proportion of test questions for which the examinee exhibits solution behavior (Wise & Kong, 2005). When the RTE score is close to zero, it represents a low-effort student who rapidly guesses most of the test question answers, while an RTE close to one represents a high-effort examinee who engages in solution behavior in answering most of the questions. Therefore, the rapid-guessing rate is defined as the inverse RTE score and captures the percentage of questions a student answers guessing rapidly.

To validate the RTE scores, Wise and Kong (2005) use data from a low-stakes computer test of a random sample of about 400 college students. To set the time thresholds that separate rapid-guessing from solution behavior, Wise and Kong (2005) conduct a visual inspection of response time distributions and question structure for each question separately. Wise and Kong (2005) show that RTE is then a valid measure of student motivation because of its high reliability, alpha of .97, and its correlation with other measures of motivation such as self-reported test effort. Additionally, their results show that RTE is weakly correlated with SAT scores, which exemplifies that student motivation can be differentiable from ability, a distinction not easily possible using

self-reported measures of effort. Finally, the RTE approach evinces that the rate at which rapid guessers choose the right answer is not higher than the probability of getting the question right by chance, which suggests that this method creates a reliable distinction between rapid-guessing and solution behavior.

Although other studies obtain similar findings to Wise and Kong (2005) regarding the RTE score validity (Kong et al., 2007; Swerdzewski et al., 2011; Wise, 2006), performing a question-by-question inspection to set time thresholds can be tedious and unfeasible on long assessments such as PISA. Instead, Wise & Ma (2012) propose using the normative threshold (NT) method to set the question-by-question time thresholds. In the NT method, the time threshold is a percentage of the mean response time of a given question. The authors recommend the threshold should not exceed a maximum value of 10 seconds; thresholds above 10 seconds, they argue, may not produce a reliable classification of rapid-guessing and solution behavior (Setzer et al., 2013).

Wise and Ma (2012) evaluate the performance of three thresholds, 10, 15, and 20 percent of the mean question-specific response time, on identifying rapid-guessing responses. Using data from a large-scale computer-based assessment that has more than 200 thousand students from the third to the ninth grades in the U.S., the authors find that only the NT at 10 percent of the mean shows accuracy in classifying solution and rapid-guessing behavior. In contrast, the NT at 15 and 20 percent provide evidence of classifying effortful responses as rapid-guessing. The authors recommend using the NT at 10 percent of the mean given its better accuracy in classifying effortful and non-effortful responses.

Concerning how low student effort can potentially distort average test score results, as well as proficiency rates for a group of examinees, Wise & DeMars (2010) exclude from the calculation

of group test performance the test score data of low-effort students in order to obtain an effort-corrected measure of overall achievement. The authors use a sample of about 300 college students who take a low-stakes computer test and then remove from the sample the test scores of low-effort examinees whose RTE score is below 90 percent. Their findings show that the mean test score gains almost doubled after effort-corrections, and the percentage of students scoring at or above the proficiency score increased approximately by eight percentage points after adjusting test scores by effort. Our paper contributes to this literature by studying patterns of rapid-guessing in PISA and studying their importance on observed differences in test performance across countries, as well as differences in test score gender gaps within each country.

There is little research available that explicitly studies the effect of student effort on gender differences in test performance (DeMars et al., 2013; Soland, 2018a, 2018b; Wise et al., 2009). In this respect, this paper contributes to an emerging literature on this topic. DeMars et al. (2013) study gender differences in test effort using RTE scores of a random sample of about 2,000 college students. The authors find that, on average, male students have a lower RTE score than their female peers. At the lower tail of the RTE score distribution, the gender differences are more significant given that a higher percentage of male students engage in rapid-guessing behavior. However, the limitation of this study is that the sample size hinders generalizing the findings.

Along these lines, Soland (2018a) and Soland (2018b) extend the analysis from DeMars et al. (2013) and Wise et al. (2009) by not only studying gender differences in the RTE scores but also assessing how accounting for student effort may change the measured achievement gaps in math and reading. Soland (2018a) and Soland (2018b) use student data from five and seven states in the U.S., respectively, that come from the Measures of Academic Progress (MAP) test. The

findings suggest that although the male-female differences in rapid-guessing rates do not change the interpretations of achievement gaps in a significant way, the gender gap in math increases after corrections and it is more sensitive to effort-adjustment than the reading gap. Soland (2018a) calls into question whether or no recent progress in narrowing the gap in math may reflect differences in effort rather than test score gains by female students.

A related work that connects student effort with gender achievement gaps, but using data from international assessments, also highlights the implications of effort in the measurement of gender gaps in test scores. Balart and Oosterveen (2019) employ the rate of decline in performance throughout the PISA 2015 test to study gender differences in sustaining performance and its implications for the gender achievement gap. The authors find that in longer assessments, the gender gap in math and science decreases, which occurs because, in most countries, girls are better able to sustain performance throughout the test relative to boys, even in math and science subjects.

In this paper, we use data from the computer-based assessment PISA 2015 to examine to what extent student effort helps explain cross-country variation in test performance, as well as gender gaps in achievement, within each country, in the subjects of math, reading, and science. Our study builds upon the previous work we present in this literature review, especially on previous work from Balart & Oosterveen (2019), Debeer et al. (2014), DeMars et al. (2013) Soland (2018a, 2018b), Wise & Ma (2012), and Zamarro et al. (2019). Our study advances the current state of knowledge in two ways:

First, we contribute to the student effort literature in international assessments such as PISA (Balart & Oosterveen, 2019; Debeer et al., 2014; Zamarro et al., 2019) by using the NT method and RTE approach to measure student motivation. To our knowledge, this method has not been

applied to the full PISA achievement sample given that assessments before 2015 are paper-based assessments. Therefore, studies that use earlier versions of PISA adopt other approaches to define student effort because it is not possible to obtain response times for a paper-based test.

We find two studies that use the NT, or a similar method, to identify low-effort examinees in PISA 2015; however, they focus on only one subject or a subsample of students and do not analyze the consequences of low-effort on gender achievement gaps (Akyol et al., 2018; Michaelides et al., 2020). In contrast, Balart and Oosterveen (2019)'s work focuses on gender achievement gaps, but it uses a different measure of effort.

Second, we contribute to the RTE literature by replicating the RTE approach and the NT method in a large international representative sample. Most of the research using this technique focuses on U.S. samples, and some of them are based on small convenient samples (DeMars et al., 2013; Soland, 2018a, 2018b; Soland et al., 2019; Swerdzewski et al., 2011; Wise et al., 2009; Wise & DeMars, 2005; Wise & Ma, 2012). Besides, few studies analyze gender differences in student effort using the RTE approach (DeMars et al., 2013; Soland, 2018a, 2018b; Wise et al., 2009) and the implications for gender achievement gaps. Only Soland (2018a) and Soland (2018b) assess the effects of rapid-guessing behavior on the measurement of gender achievement gaps in math and reading; however, these studies only use a sample of students from the U.S.

3. Data

The Programme for International Student Assessment (PISA) is a triannual survey, managed by the Organization for Economic Co-operation and Development (OECD), which evaluates how well 15-year-old students are capable of using their knowledge and skills to meet real-life challenges in the areas of mathematics, reading, and science. The number of participants

in 2015 was about 540,000 students from 72 countries and economies[1]. In addition to the three core evaluation subjects, PISA 2015 evaluated students on collaborative problem solving and financial literacy. These last two subjects were optional for the participant countries. Every PISA wave focuses on a subject; in 2015, the primary area of assessment was science, and therefore, the evaluation included more questions about this topic.

For the first time, the main form of assessment in PISA 2015 was computer-based. Paper-based assessments were available to countries that had limited access to computers. These two forms of assessments lasted about two hours. After the completion of the test, students answered a background questionnaire about 30 minutes long that collected information about home environment, school, and learning experiences.

For this study, we restrict our sample to those countries and economies that took the computer-based test. We also exclude the test booklets that have clusters about cooperative problem solving, financial literacy, or that were designed for students with special needs. Our final sample contains 55 countries/economies[2]. We only focus on the computer-based assessment because this form includes response times for each student on each question, which we use later in order to define rapid-guessing behavior.

In the PISA 2015 assessment, the test booklets are randomly assigned to students within each country. The total number of questions in these booklets ranges from 47 to 71 questions with an average of 60 total questions.

[1] To simplify, in the rest of this document we use the term countries to refer to countries and economies. See Table 1 for the list of countries and abbreviations.

[2] We restrict our analytical sample to countries and economies. We exclude the adjudicated regions of USA Massachusetts, USA North Carolina, and the adjudicated regions from Spain.

3.1. Measuring student effort in PISA

3.1.1. Rates of rapid-guessing in the entire assessment

Defined as the inverse RTE score ($1 - RTE$), our measure of rapid guessing represents the proportion of responses, out of all test questions, in which an examinee engages in rapid-guessing behavior. To identify rapid-guessing behavior, we first calculate the average response time for each question across all test booklets within each country. Second, we use the NT method at 10 percent of the mean to set time thresholds for each question within each country; responses given at a smaller time than these time thresholds are considered instances of rapid-guessing. We focus on 10 percent of the mean response time because prior evidence suggests that this threshold has better accuracy in classifying rapid-guessing and solution behavior (Wise & Ma, 2012). Finally, we identify the number of questions in which an examinee's response time is below the 10 percent of the mean[3] to calculate the inverse RTE score (i.e., the proportion of rapid-guessing responses) on the complete test for each student within each country.

When calculating the rapid-guessing rate on the test, we exclude response times from students whose total time in completing the test exceeds 120 minutes[4], which represents 5,311 observations. Although the test was expected to last two hours, we are unsure of whether or not some students obtained extra time. Total time above 120 minutes could also occur because test proctors had to log off the computer assessment one by one. According to what we see in the data, it seems that in some cases, the proctor did not end the session, or there was a technical problem

[3] We also performed a sensitivity analysis using a more conservative threshold of 5% of the mean response time and our findings do not change significantly. Results are available from the authors upon request.

[4] We also conducted our estimations without excluding outliers in total time and the results do not change meaningfully. Estimates excluding outliers are the ones presented in the paper since they are more conservative. The results that did not exclude outliers are available upon request.

in the data collection because we find some records of total time spent on the assessment of up to 171 hours.

Tables 2 and 3 show descriptive statistics of rapid-guessing behavior[5] in the complete assessment, as well as other variables of interests that we describe in the following sections. Students in the estimation sample take, on average, 79 minutes to complete the assessment (see table 2). Approximately 185 observations have total times of less than two minutes, which may occur because of a technical problem in the data collection or because the students decided not to complete the assessment. The variation in total time is lower between countries than within countries, which suggests that the distribution of total time across countries probably does not vary considerably, ruling out meaningful country differences in the total time allocated to the test.

Although the proportion of rapid-guessing on the test ranges from 0 to 100 percent, students across countries on average rapidly guess 3 percent of all test questions (see table 2). Since the average number of questions in PISA booklets is 60, a 3 percent rapid-guessing rate on the test is equivalent to rapidly guessing about 2 questions on average. Table 2 also shows that the variation in rapid-guessing behavior is higher across all students, regardless of country, and within countries rather than between countries. The standard deviations for the whole sample show that, overall, the average dispersion in the proportion of rapid-guessing responses is about 8 percentage points.

[5] Due to a technical issue in the timing variables, as of December 2020, PISA re-issued the time data for 2015 so that they capture the total time students spent on a question. Before, the timing variables captured the total time spent on a question the last time a student visited that question's screen, which means that if a student went back and forth to revise a question several times, the time variable of that question would only capture the total time spent on the question in the last visit. Although this behavior is limited because students can only go back and forth within screens of questions that belong to a given test module, measures of the total time spent on a question would lead to more accurate identification of rapid-guessing instances. As a result, in this paper we construct the measure of rapid-guessing using the most recent data available consisting on total response times for each test item. However, our findings using the old and new timing information provided by PISA do not affect our main conclusions. The results using the old variables are available upon request.

When comparing students within each country, the variation is slightly lower, showing that the dispersion of rapid-guessing proportions is, on average, 7 percentage points above or below the mean. In contrast, the variation between countries is roughly a third lower, with a standard deviation of about 2 percentage points.

When we look at the average rapid-guessing rate for boys and girls (see table 3), their rates differ roughly by one percentage point. Girls have a slightly lower probability of engaging in rapid-guessing behavior than boys. This result is similar to prior research which finds that female students, on average, have lower rapid-guessing rates than boys have (DeMars et al., 2013; Soland, 2018a, 2018b). This result is consistent with the difference in total time between girls and boys. Girls, on average, take 5 minutes longer than boys do in completing the assessment.

In summary, we find descriptive evidence of rapid-guessing behavior in PISA 2015. The dispersion of this variable is higher when we compare all students, regardless of country, and when we compare examinees within each country. The variation is lower across countries, which suggests that across countries, the distributions of rapid-guessing behavior probably are not very different from each other. The latter does not necessarily imply that student effort is not relevant to explain cross-country variations in achievement. Zamarro et al. (2019) find that even though cross-country variation in student effort is lower than the within-country variation, the differences in student effort across countries are still relevant in explaining cross-country heterogeneity in test scores. Finally, we observe that girls, on average, exhibit more effort and take more time to complete the test than boys do.

3.1.2. Item non-response rates on the student background survey

We replicate the Zamarro et al. (2019) approach by calculating the item non-response rate in the student survey, but this time by using a computer-based survey from PISA. This rate corresponds to the proportion of questions that a student skips or does not complete on the PISA survey that follows the test.[6] We focus on the item non-response rate since previous research finds that this indicator has the highest predictive power in explaining cross-country variation in performance on paper-based assessments (Boe et al., 2002; Zamarro et al., 2019). According to table 2, students do not respond to between 0 and 98 percent of survey items, and on average, they leave blank 7 percent of the questions. The variation between and within countries on the item non-response rate is almost twice the variation on the rapid-guessing rate on the test. Girls on average have a roughly 2-percentage-points lower item non-response rate than boys have (see table 3). Overall, girls consistently show higher levels of effort than boys do both in the test and the survey.

4. Estimating the role of student effort in explaining cross-country differences in achievement and within-country differences in gender achievement gaps

We follow a similar methodological approach to that of Zamarro et al. (2019) and conduct a country-random-effects estimation for each tested subject in PISA to assess the role that student effort may have in explaining cross-country differences in performance and within-country gender achievement gaps. Our dependent variable in the model (1) below corresponds to the plausible value j (i.e., test score) that student i from country c obtained on the subject s. The variables $INRsurvey$ and $RGtest$ represent the item non-response rate on the student background survey

[6] Although we have response times for this questionnaire, we do not construct rapid-guessing rate for the background survey because PISA does not report response times for each question but for a group of items.

and the proportion of rapid-guessing responses on the entire assessment, respectively. The terms α and ε represent the country random-effect for the subject s and the error term, respectively.

$$TestScore_{ic}^{s_j} = \beta_0^{s_j} + \beta_1^{s_j} INRsurvey_{ic} + \beta_2^{s_j} RGtest_{ic} + \alpha_c^{s_j} + \varepsilon_{ic}^{s_j} \quad (1)$$

PISA reports test scores as plausible values. These scores are calculated using a multiple imputation method that aims to increase accuracy in measuring students' skills[7]. Each student has 30 possible values in total; ten plausible values for each subject. We estimate model (1) using as a dependent variable each of the 10 plausible values on each subject, and we report the average estimated coefficients for each subject in table 4. We first examine effort measures separately and estimate equation (1) for each effort measure. We replicate Zamarro et al. (2019) results and find that item non-response is also a statistically significant predictor of test performance in this computer-based assessment.

From equation (1), we follow Zamarro et al.'s (2019) approach and obtain effort-adjusted test scores ($\overline{Adjusted\ Score_{ic}^{s}}$) for each student and subject by obtaining the average of the sum of the estimated coefficients of the intercept, the country random-effect, and the residuals ($\hat{\beta}_0^{s_j} + \hat{\alpha}_c^{s_j} + \hat{\varepsilon}_{ic}^{s_j}$). We then compute the average adjusted score for each subject across the 10 plausible values using the following formula:

$$\overline{Adjusted\ Score_{ic}^{s}} = \sum_{j=1}^{10} \frac{\hat{\beta}_0^{s_j} + \hat{\alpha}_c^{s_j} + \hat{\varepsilon}_{ic}^{s_j}}{10} \quad (2)$$

[7] For further information about plausible values and multiple imputation method, see chapter 9 of the PISA 2015 technical report.

We next calculate the average effort-adjusted gender gap $\overline{\widehat{GAP_c^s}}$ for each country and subject by subtracting the average effort-adjusted test score of girls minus the score of boys using the formula:

$$\overline{\widehat{GAP_c^s}} = \sum_{g=1}^{G_c} \frac{\overline{Adjusted\ Score_{G_c}^s}}{G_C} - \sum_{b=1}^{B_C} \frac{\overline{Adjusted\ Score_{B_c}^s}}{B_c} \quad (3)$$

Where G_C and B_c represent the sample sizes of girls (G) and boys (B) from country c, respectively.

Our effort-unadjusted test scores correspond to the average of the actual plausible test score values that each student on the estimation sample obtained on each subject. Then we calculate the average effort-unadjusted achievement gap $\overline{GAP_c^s}$ for each subject and country using formula 3 but replacing the numerator with the effort-unadjusted score that boys and girls in the estimation sample obtained on each subject. On average, students score before effort-adjustment 471, 474, and 476 points on the subjects of math, reading, and science, respectively (see table 2). Before effort-adjustment, girls score on average, 25 points higher on reading than boys do, whereas in math and science, girls score 9 and 4 points lower than boys do, respectively (see table 3).

After calculating the average gender achievement gap for each subject and country using test scores, we compare the effort-adjusted and unadjusted gap using the Glass's Δ effect size (Smith & Glass, 1977) formula:

$$\Delta\%GAP_c^S = \frac{\overline{\widehat{GAP_c^s}} - \overline{GAP_c^s}}{SD_c^s} * 100 \quad (4)$$

Where SD_c^s represents the standard deviation (SD) of the effort-unadjusted test score of subject s in country c. Formula (4) represents the change of the achievement gap relative to the effort-unadjusted test score, measured as a percentage of one standard deviation. In other words, formula (4) shows, compared to the unadjusted test score, what would be the expected change in the average gender achievement gap for each country, and subject, in the absence of student effort heterogeneity. We adjust the signs of the calculated changes such that negative signs represent a widening of the gender achievement gap, and positive signs represent a reduction of the gap.

5. Results of the role of student effort in explaining cross-country differences in student achievement

When we analyze to what extent our effort measures explain the variation in performance in the PISA test, we find that both item non-response rates and rapid-guessing are relevant predictors of test scores (see table 4). A one standard deviation increase in the proportion of rapid-guessing responses in the test is associated with a decrease of 0.26, 0.29, and 0.3 SDs on the math, science, and reading test scores, respectively (see columns 3, 6, and 9). Regarding the item non-response variable, a one SD increase on this variable is associated with a decrease of 0.12, 0.13, and 0.16 SDs on the math, science, and reading test scores, respectively (see columns 3, 6, and 9). These findings suggest that low-effort students often experience lower test performance.

Additionally, we find that our effort measures have more explanatory power across countries than within countries. Altogether, our effort measures explain between 43 and 48 percent of the variation in test performance across countries, which is similar to Zamarro et al.'s (2019) findings, versus about 12 to 16 percent of the within-country variation in test scores (see table 4). This finding is not very surprising. Previous work by Wise et al. (2020) examine the distortive

effect of effort heterogeneity in test scores at the school level using data from a pilot computer-based assessment from PISA in the U.S. Although the authors find variation in effort across schools, the mean test scores for each school after effort-adjustment do not significantly change compared to the effort-unadjusted scores. These effort measures may perform better at capturing differences in effort across different contexts or cultures than within similar environments, such as schools or countries.

6. Results of the role of student effort on gender achievement gaps

In this section, figures 1, 2, and 3 present the change of the gender achievement gap in the absence of student effort heterogeneity as the percentage of one SD. Countries in the green color correspond to a reduction of the gap, represented by a positive change after adjustments for student effort. In contrast, the remaining colors correspond to a widening of the gap represented by a negative sign; the darker the color of a country is, the wider the gap becomes. Tables 5, 6, and 7 show the effort-adjusted and unadjusted gaps, as well as the change for each country and subject as a percent of one SD.

The widening of the gap in math achievement occurs in 50 out of 55 countries and ranges from 0.5 to up to 36 percent of one SD (see figure 1 and table 5). The smallest increase occurs in Brazil, whereas the highest increase occurs in Qatar. The latter means that, relative to the unadjusted test scores, in Qatar, the gap in math achievement could be up to 36 percent of one SD wider in favor of boys in the absence of variation in student effort. The size of the effort-unadjusted gap in Qatar is about 11.4 points in favor of girls, while after adjustment, girls fall behind boys by about 21.9 points, which represents a difference of about 33 points between the two gaps (see table 5). Another meaningful change occurs in Bulgaria. Before the adjustment, the gap is about 0.9

points in favor of girls, but after effort-adjustment, it becomes 6.6 points in favor of boys, which represents a widening of the gap by roughly 7.5 points, or 8.4 percent of a SD, favoring boys (see table 5).

In contrast, only in 5 out of 55 countries, the gap in math achievement narrows in the absence of student effort heterogeneity, according to figure 1. The decrease in the gap ranges from 0.8 to up to 6.2 percent (see table 5). The smallest decline occurs in the Dominican Republic, whereas the highest decline occurs in Finland. In the latter case, the size of the effort-unadjusted gap is about 7 points in favor of girls, and after adjustment, its size is about 2.3 points, which represents a reduction of 5 points (or 6 percent of a SD) in the math achievement gap.

We obtain similar results when we look at the change in the science achievement gap in figure 2. In 45 out of 55 countries, the widening of the gap ranges from 0.5 percent up to 40 percent of a SD. Again, in this case, the smallest increase in the science gap also occurs in the Dominican Republic, whereas the highest increase occurs in Qatar (see table 6). The latter means that in Qatar, the gap becomes about 40 percent of a SD wider after effort-adjustment, relative to the unadjusted test scores. The effort-unadjusted gap in Qatar is roughly 22.9 points in favor of girls, whereas after adjustment, girls fall behind boys by roughly 14.5 points, which represents a widening of the gap of about 37 points (see table 6).

When we analyze the percentage change in the reading achievement gap (see figure 3), the results are very different from those in math and science since most countries now appear in the green color indicating a narrowing of the gender gap after student effort adjustments. In 53 out of 55 countries, the reading achievement gap in the absence of variation in student effort narrows from 0.6 to up to 39 percent (see table 7). The smallest reduction of the gap occurs in Brazil,

whereas the highest reduction occurs in Qatar. In the latter country, the effort-unadjusted reading gap is about 53 points in favor of girls; after adjustment, it is about 12 points. Although the effort-adjusted gap in Qatar still favors girls, the gap experiences a reduction of roughly 40 points, or 39 percent of a SD, favoring boys relative to the unadjusted scores. In contrast, in Peru, the reading gap widens by 2.5 percent of a SD in the absence of student effort variation.

Overall, in most PISA countries that took the computer assessment, the gender achievement gap in math and science could be up to 36 and 40 percent of a SD wider in favor of boys, respectively, in the absence of variation in student effort. In contrast, the gender gap in reading could narrow up to 39 percent in favor of boys in the absence of variation in student effort. Our findings are consistent with Soland (2018a) and Soland (2018b), who find that the male-female gap in math is more sensitive to test effort compared to the reading gap.

7. Robustness checks

One of our concerns with our item non-response and rapid-guessing variables is to what extent they capture student effort. Although there is more robust evidence from international assessments that item non-response in the student survey appears to capture relevant information on student effort (Boe et al., 2002; Zamarro et al., 2019), there is not so much robust evidence available for the measure of rapid-guessing in the context of international assessments. In this section, we aim to assess whether or not both effort measures capture student effort.

To test whether or not our measures capture student effort, we study the correlations of our two variables with other relevant educational statistics at the country level. The idea behind this analysis is that if our measures capture important components of student effort, they should be correlated with test performance and other educational indicators that should also be correlated

with student effort such as dropout rates, out-of-school rates, or repetition rates. We expect that low-effort countries have a lower performance in the test as well as higher rates in these three education statistics.

To study these relationships, we calculate the average rapid-guessing and item non-response rates for each country. Then, we merge this information with 2015 education statistics from the World Bank at the country level. We choose the education statistics of the year 2015 since they match the year of the PISA data. Additionally, The World Bank's education statistics focus on lower and upper secondary schools, when available, as these schooling levels approximately coincide with the age of 15 years old, the age at which PISA evaluates the students.

Panels a, b, and c of figure 4 represent the relationship between rapid-guessing and PISA performance in math, science, and reading respectively, whereas figure 5 presents the same graphs for item non-response. We corroborate that rapid-guessing and item non-response are correlated with test performance. Figures 4 and 5 show that countries with high levels of performance on these three subjects tend to have lower rates of rapid-guessing and item non-response. The relationship between these effort measures and test performance seems stronger for rapid-guessing since correlations range from 0.58 to 0.64, whereas for item non-response, correlations range from 0.47 to 0.53.

Regarding the relationship between effort measures and education statistics, figure 6 shows the relationship between rapid-guessing (panel a) and item non-response (panel b) with the cumulative dropout rate to the last grade of lower secondary general education[8]. Although the

[8] The cumulative dropout rate corresponds to the proportion of students enrolled at a given grade and school year who are not enrolled in the following school year. For more information, see The World Bank data catalog.

correlations in figure 6 are not as strong as the ones for test performance, we still find that countries with lower rapid-guessing and item non-response rates often have lower dropout rates. We find similar results in figures 7 and 8 that illustrate the relationship between our effort measures and the rate of out-of-school youth of upper secondary school age[9] and the repetition rate in lower secondary general education[10], respectively.

In summary, we observe that out effort measures are negatively correlated with student test performance suggesting that countries with higher average rapid-guessing and item non-response rates tend to have lower average test performance. In contrast, we generally observe a positive relationship between effort and education statistics that signal low-effort. Countries that have high average rates of item non-response and rapid-guessing often have high dropout, out-of-school, and repetition rates.

We conduct an additional check to our rapid-guessing variable to test whether or not our threshold is identifying rapid-guessers accurately. Wise & Gao (2017) propose to calculate and study the accuracy rates for rapid-guessing and solution behavior. The accuracy rate for rapid-guessing corresponds to the total correct responses under rapid-guessing behavior, divided by the total responses classified as rapid-guessing. The same formula applies to the accuracy rate of solution behavior but this time focusing on the responses classified as solution behavior. According to Wise & Gao (2017), the idea behind comparing these rates is that if the percentage

[9] The rate of out-of-school youth of upper secondary school age employs the same formula as the rate of out-of-school adolescents of lower secondary school age but this time employs the out-of-school upper secondary school age youth and the upper secondary school age population. For more information, see The World Bank data catalog.

[10] The repetition rate in lower secondary general education corresponds to the number of students who repeat a grade in lower secondary education in a given school year divided by enrolment in lower secondary education in the previous school year. For more information, see The World Bank data catalog.

of correct responses under rapid-guessing is higher than that of solution behavior, it suggests that the threshold is capturing effortful responses instead of careless answering under rapid-guessing.

We present the comparison of the accuracy rates of rapid-guessing and solution behaviors for each country in figure 9. We find that our 10 percent threshold for the rapid-guessing measure consistently shows significantly lower accuracy rates than that of responses classified as solution behavior. In all countries, the accuracy rate of rapid-guessing is less than or equal to 10 percent, and in 45 out of 55 countries, this rate is less than or equal to 5 percent. In conclusion, we are confident that our rapid-guessing measure with a 10 percent threshold performs well at capturing low-effort students.

8. Conclusions

In this paper, we use data from PISA 2015, a triannual survey that evaluates 15-year-old students from 74 countries in math, reading, and science to study the effect of student effort on cross-country differences in performance as well as within-country gender gaps in achievement. We restrict our sample to the 55 countries which take the computer-based test and use innovative measures of effort based on rapid-guessing on the test and item non-response on the survey.

Altogether, our effort measures, on average, explain between 43 and 48 percent of the variation in test scores across countries. Our results also suggest that the estimated gender achievement gap in math and science could be up to 36 and 40 percent of a SD wider, respectively, in favor of boys in the absence of variation in student effort. The gap in these two subjects widens in most of the countries in our sample. In contrast, the estimated gender gap in reading could narrow up to 39 percent of a SD in favor of boys. Our results highlight the importance of

accounting for student effort to understand not only cross-country differences in performance but also variations in the measurement of the achievement gaps across nations.

9. References

Akyol, Ş. P., Krishna, K., & Wang, J. (2018). Taking PISA Seriously: How Accurate are Low Stakes Exams? *Taking PISA Seriously: How Accurate Are Low Stakes Exams?* https://www.nber.org/papers/w24930

Anaya, L., Stafford, F., & Zamarro, G. (2017). Gender Gaps in Math Performance, Perceived Mathematical Ability and College STEM Education: The Role of Parental Occupation. *EDRE Working Paper, 2017–21*. https://doi.org/10.2139/ssrn.3068971

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications*, *10*(1), 3798. https://doi.org/10.1038/s41467-019-11691-y

Boe, E. E., May, H., & Boruch, R. F. (2002). *Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Acheievement Differences at the National, Classroom, and Student Levels*. https://eric.ed.gov/?id=ED478493

Borghans, L., & Schils, T. (2012). *The Leaning Tower of Pisa Decomposing achievement test scores into cognitive and noncognitive components*. https://www.semanticscholar.org/paper/The-Leaning-Tower-of-Pisa-Decomposing-achievement-Borghans/add9e3d2a408bf1758e5cb3774c91e7f26b8d0b9?p2df

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in Academic Science: A Changing Landscape. *Psychol Sci Public Interest*, *15*(3), 75–141. https://doi.org/10.1177/1529100614541236

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics*, *39*(6), 502–523. https://doi.org/10.3102/1076998614558485

Demars, C. E. (2007). Changes in Rapid-Guessing Behavior Over a Series of Assessments. *Educational Assessment*, *12*(1), 23–45. https://doi.org/10.1080/10627190709336946

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The Role of Gender in Test-Taking Motivation under Low-Stakes Conditions. *Research & Practice in Assessment*, *8*, 69–82. http://www.rpajournal.com/dev/wp-content/uploads/2013/11/A4.pdf

Kane, M. (2006). *Content-Related Validity Evidence in Test Development* (pp. 131–153). Lawrence Erlbaum Associates Publishers. https://psycnet.apa.org/record/2006-01815-007

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the Response Time Threshold Parameter to Differentiate Solution Behavior From Rapid-Guessing Behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items. *International Journal of Testing*, 1–19. https://doi.org/10.1080/15305058.2019.1706529

Nix, S., Perez-Felkner, L., & Thomas, K. (2015). Perceived mathematical ability under challenge: a longitudinal perspective on sex segregation among STEM degree fields. *Frontiers in Psychology*, *6*, 530. https://doi.org/10.3389/fpsyg.2015.00530

Perez-Felkner, L., Nix, S., & Thomas, K. (2017). Gendered pathways: How mathematics ability beliefs shape secondary and postsecondary course and degree field choices. *Frontiers in Psychology*, *8*, 386. https://doi.org/10.3389/fpsyg.2017.00386

Schnipke, D. L. (1995). *Assessing Speededness in Computer-Based Tests Using Item Response Times.* https://files.eric.ed.gov/fulltext/ED383742.pdf

Schnipke, D. L., & Scrams, D. J. (1997). Modeling Item Response Times with a Two-State Mixture Model: A New Method of Measuring Speededness. *Journal of Educational Measurement*, *34*(3), 213–232. http://www.jstor.org/stable/1435443

Setzer, J. C., Wise, S. L., den Heuvel van, & Ling, G. (2013). An Investigation of Examinee Test-Taking Effort on a Large-Scale Assessment. *Applied Measurement in Education*, *26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Smith, M. L., & Glass, G. v. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. https://doi.org/10.1037/0003-066X.32.9.752

Soland, J. (2018a). Are Achievement Gap Estimates Biased by Differential Student Test Effort? Putting an Important Policy Metric to the Test. *Teachers College Record*, *120*(12). https://www.nwea.org/resource-library/research/are-achievement-gap-estimates-biased-by-differential-student-test-effort-3

Soland, J. (2018b). The Achievement Gap or the Engagement Gap?Investigating the Sensitivity of Gaps Estimates to Test Motivation. *Applied Measurement in Education*, *31*(4), 312–323. https://doi.org/10.1080/08957347.2018.1495213

Soland, J., Jensen, N., Keys, T. D., Bi, S. Z., & Wolk, E. (2019). Are Test and Academic Disengagement Related? Implications for Measurement and Practice. *Educational Assessment*, *24*(2), 119–134. https://doi.org/10.1080/10627197.2019.1575723

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two Approaches for Identifying Low-Motivated Students in a Low-Stakes Assessment Context. *Applied Measurement in Education*, *24*(2), 162–188. https://doi.org/10.1080/08957347.2011.555217

Wise, S. L. (2006). An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test. *Applied Measurement in Education*, *19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee Noneffort and the Validity of Program Assessment Results. *Educational Assessment*, *15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Gao, L. (2017). A General Approach to Measuring Test-Taking Effort on Computer-Based Tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., & Kong, X. (2005). Response Time Effort: A New Measure of Examinee Motivation in Computer-Based Tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. https://www.nwea.org/content/uploads/2012/04/Setting-Response-Time-Thresholds-for-a-CAT-Item-Pool.pdf

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of Rapid-Guessing Behavior in Low-Stakes Testing: Implications for Test Development and Measurement Practice. *Applied Measurement in Education*, *22*(2), 185–205. https://doi.org/10.1080/08957340902754650

Wise, S. L., Soland, J., & Bo, Y. (2020). The (Non)Impact of Differential Test Taker Engagement on Aggregated Scores. *International Journal of Testing*, *20*(1), 57–77. https://doi.org/10.1080/15305058.2019.1605999

Zamarro, G., Hitt, C., & Mendez, I. (2019). When Students Don't Care: Reexamining International Differences in Achievement and Student Effort. *Journal of Human Capital*. https://doi.org/10.1086/705799

Figure 1: Change in the gender gap in math achievement as a percentage of one SD

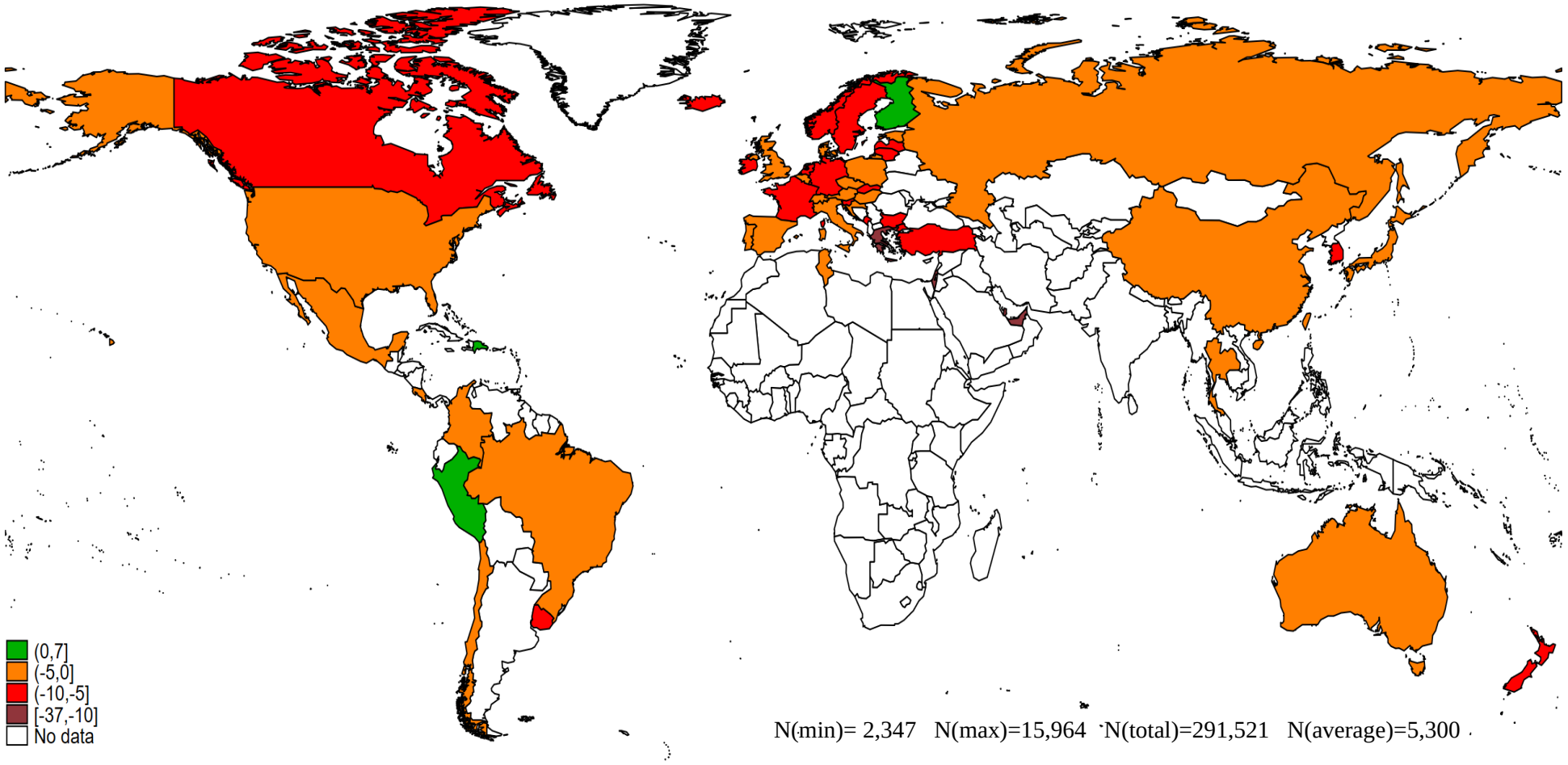


Figure 2: Change in the gender gap in science achievement as a percentage of one SD

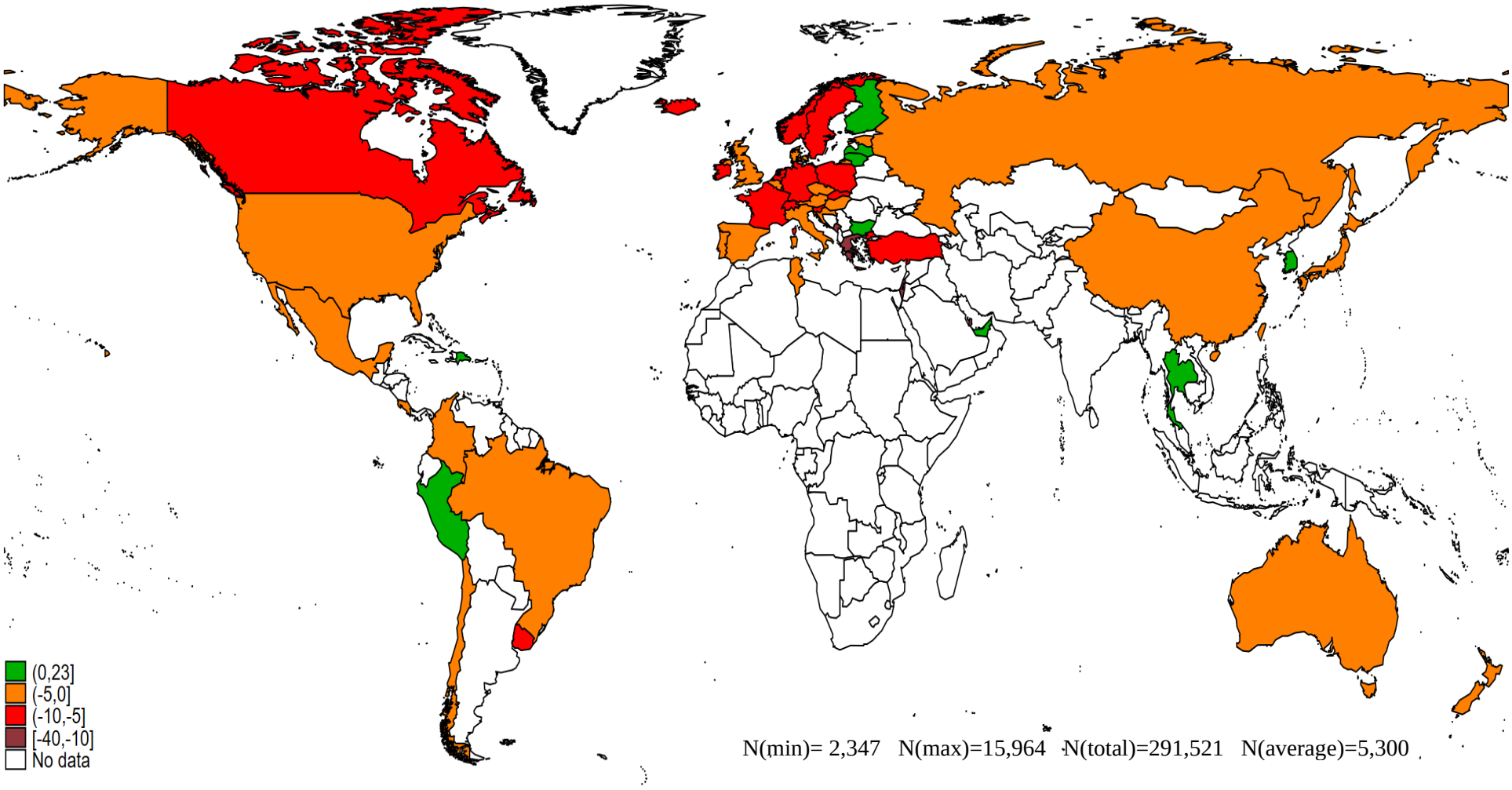


Figure 3: Change in the gender gap in reading achievement as a percentage of one SD

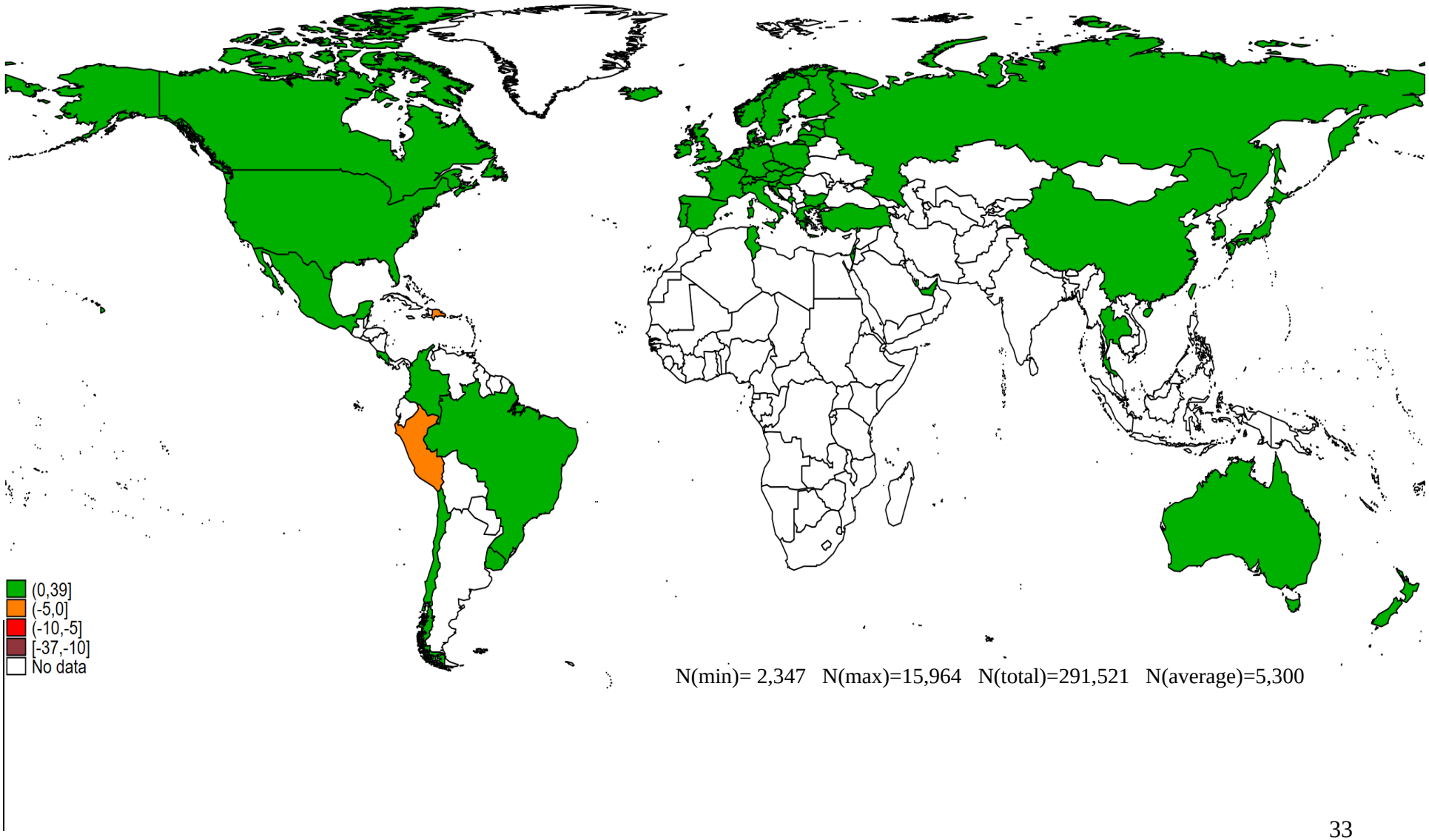
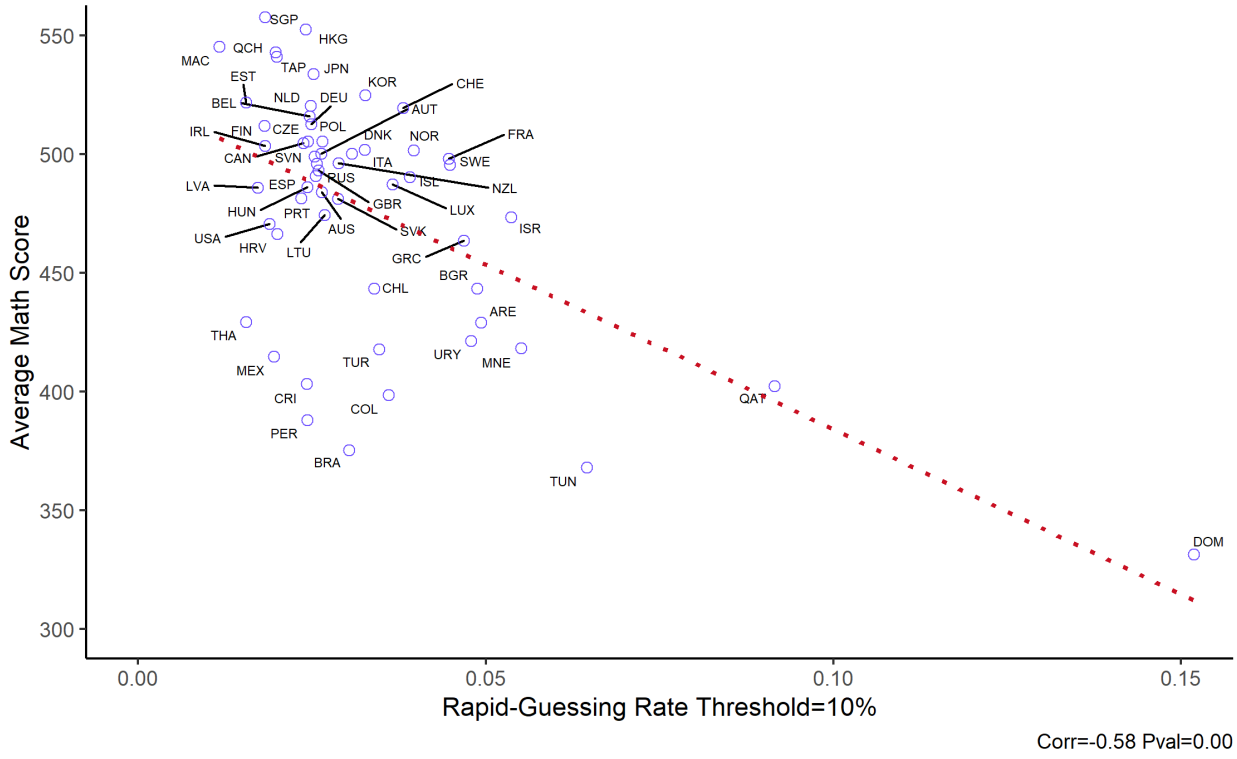


Figure 4: Relationship rapid-guessing and test performance

(a) Math



(b) Science

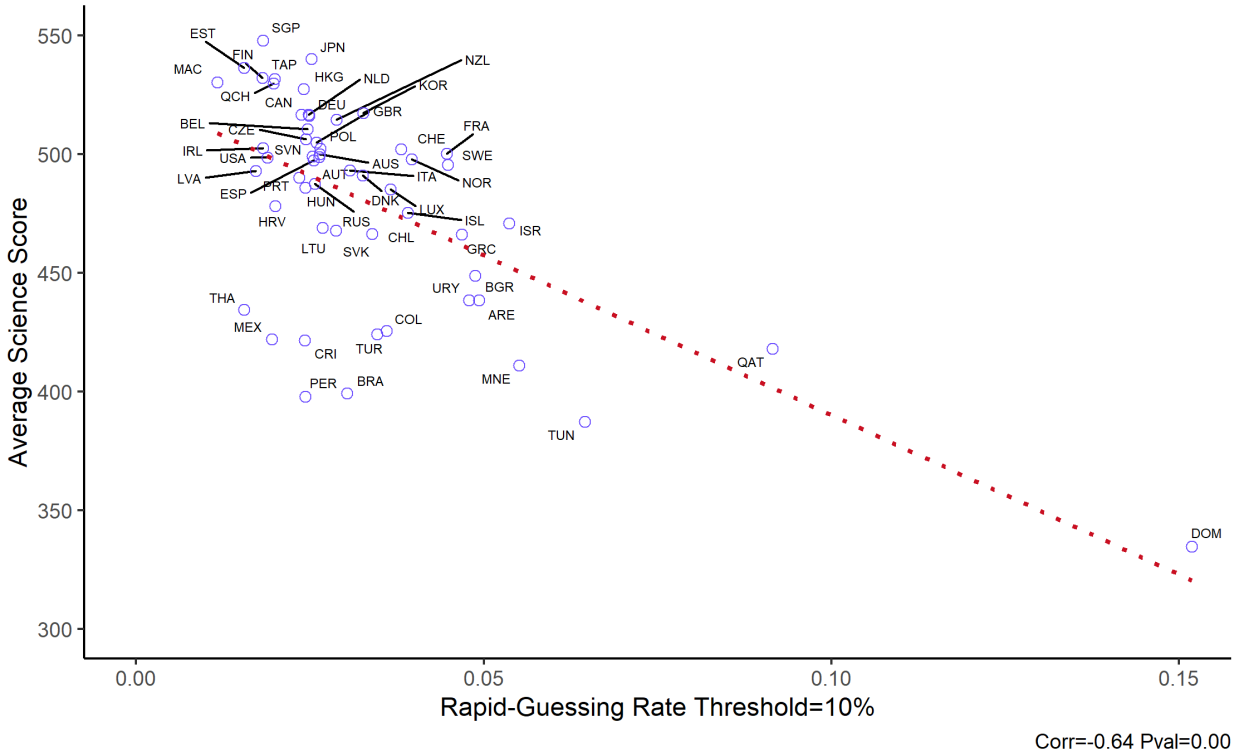


Figure 4: Relationship rapid-guessing and test performance (cont.)

(c) Reading

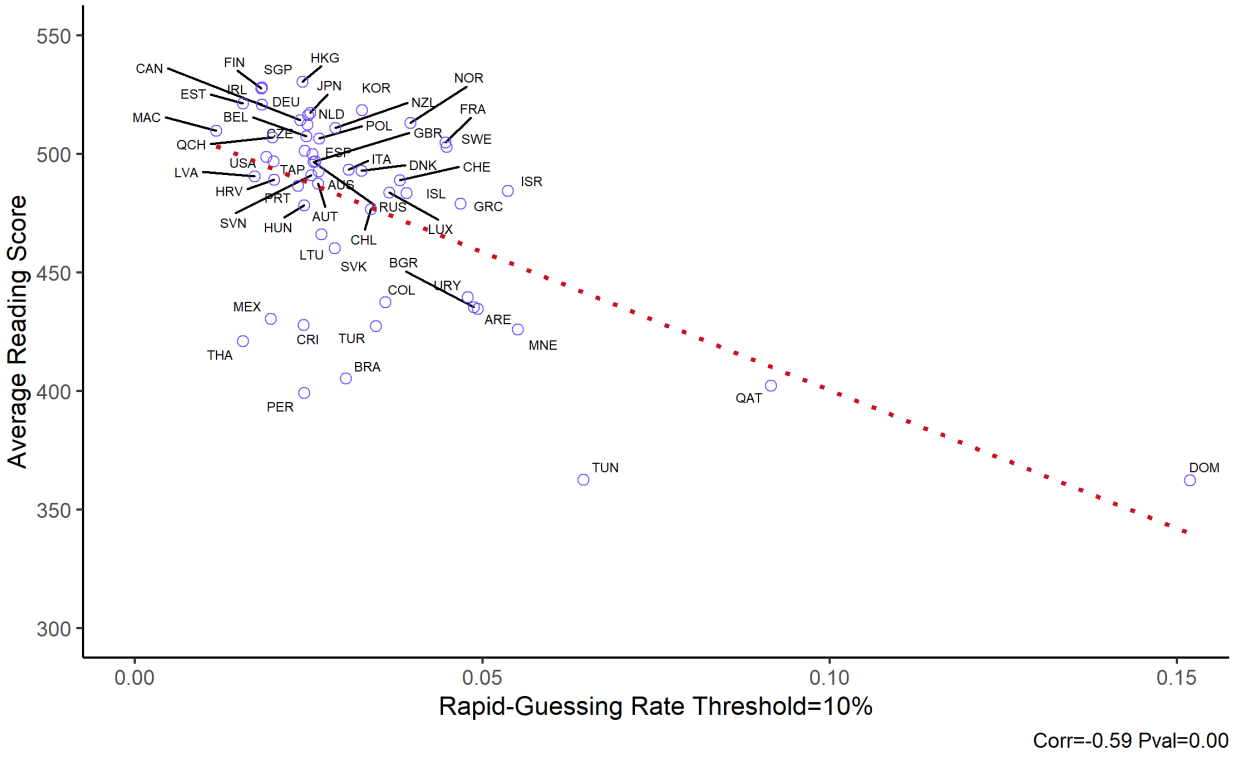
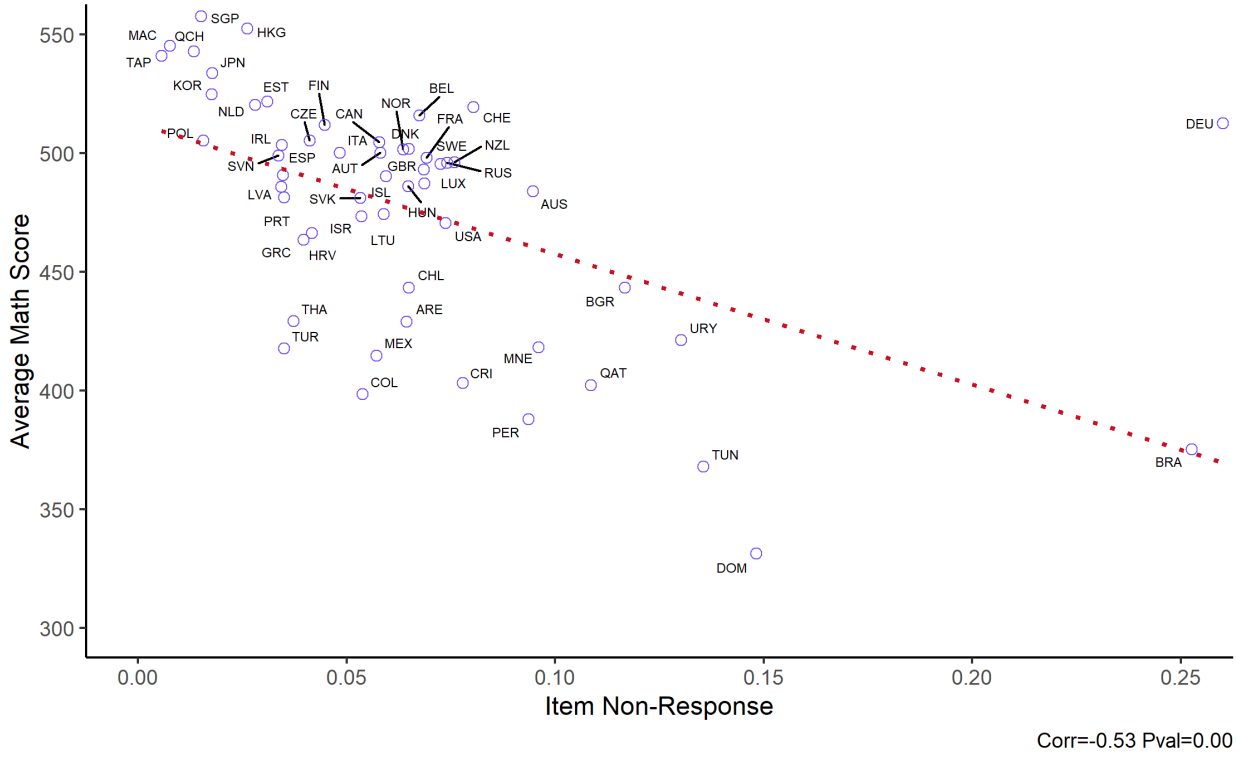


Figure 5: Relationship between item non-response and test performance

(a) Math



(b) Science

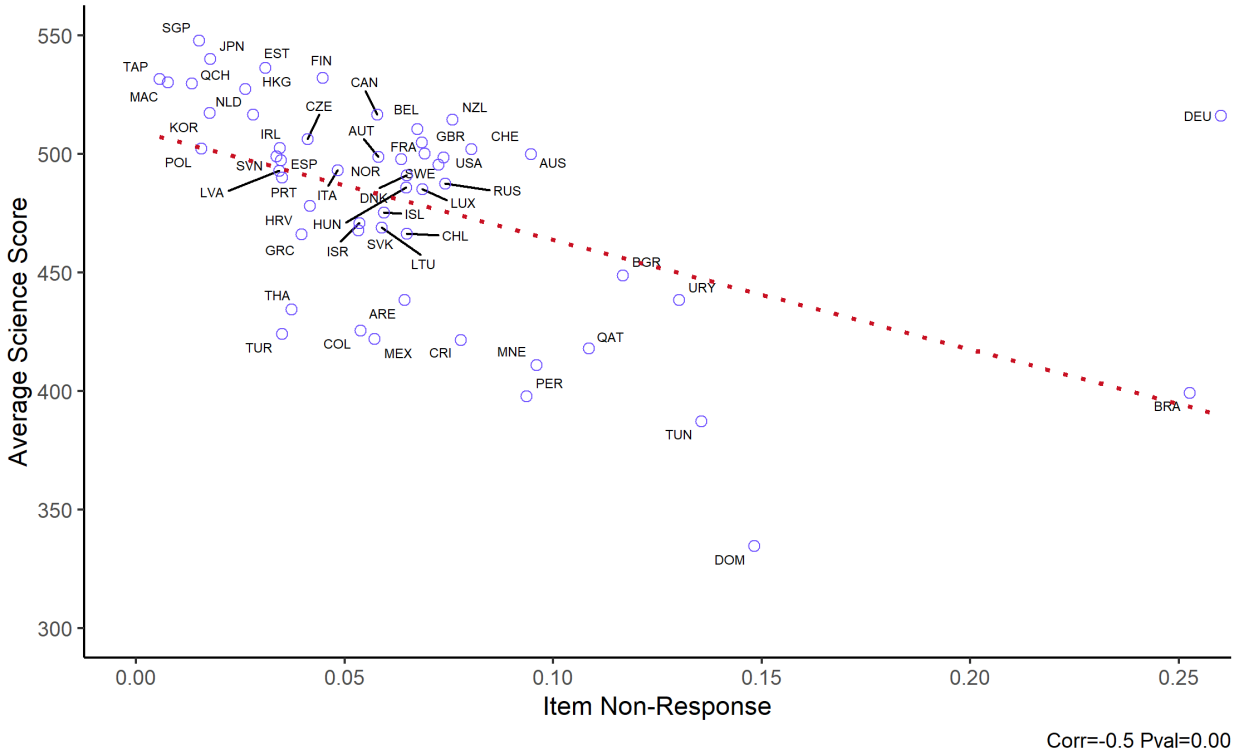
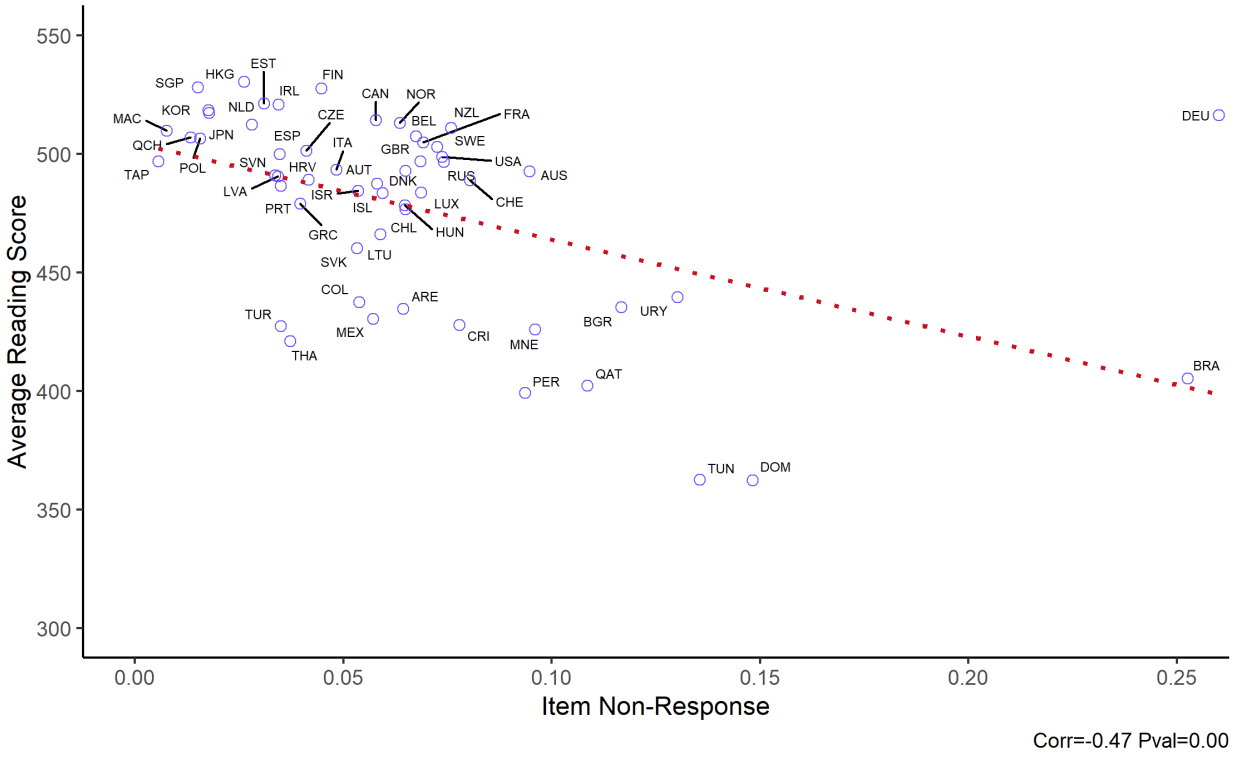


Figure 5: Relationship item non-response and test performance (cont.)

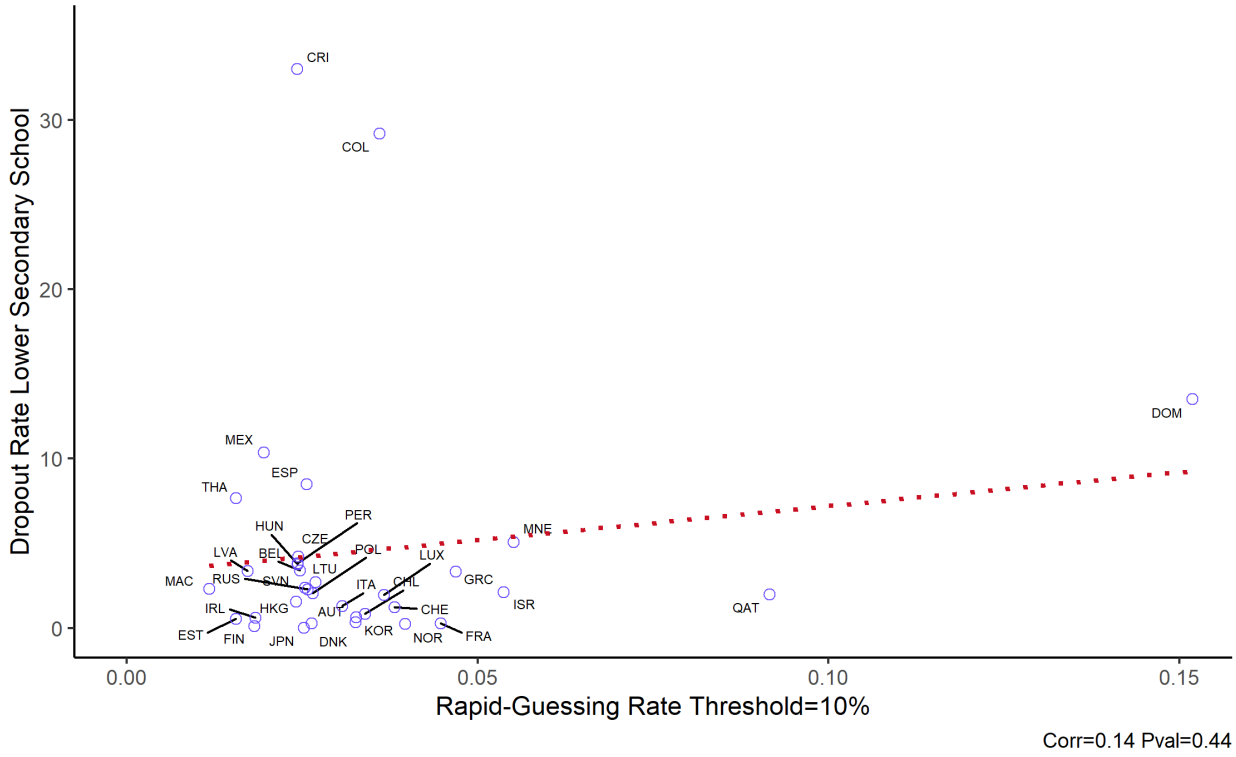
(c) Reading



Corr=-0.47 Pval=0.00

Figure 6: Relationship between effort measures and dropout rate in lower secondary school

(a) Rapid-guessing



(b) Item non-response

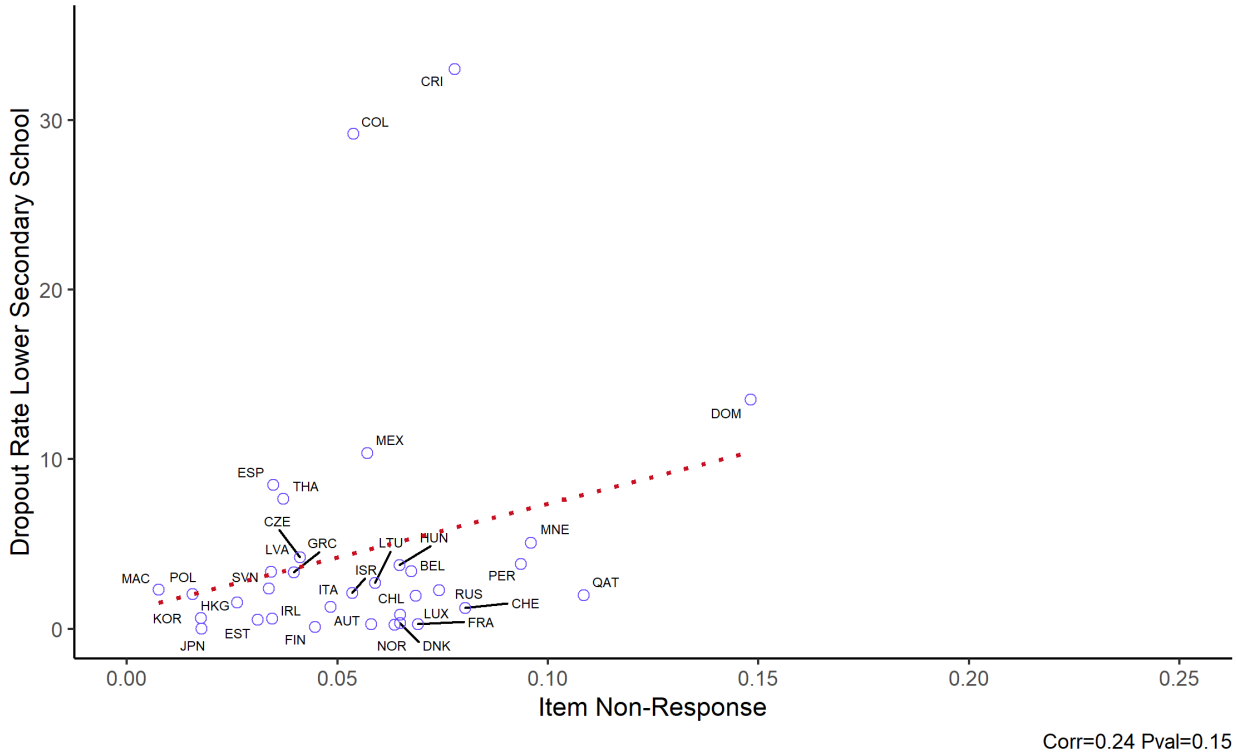
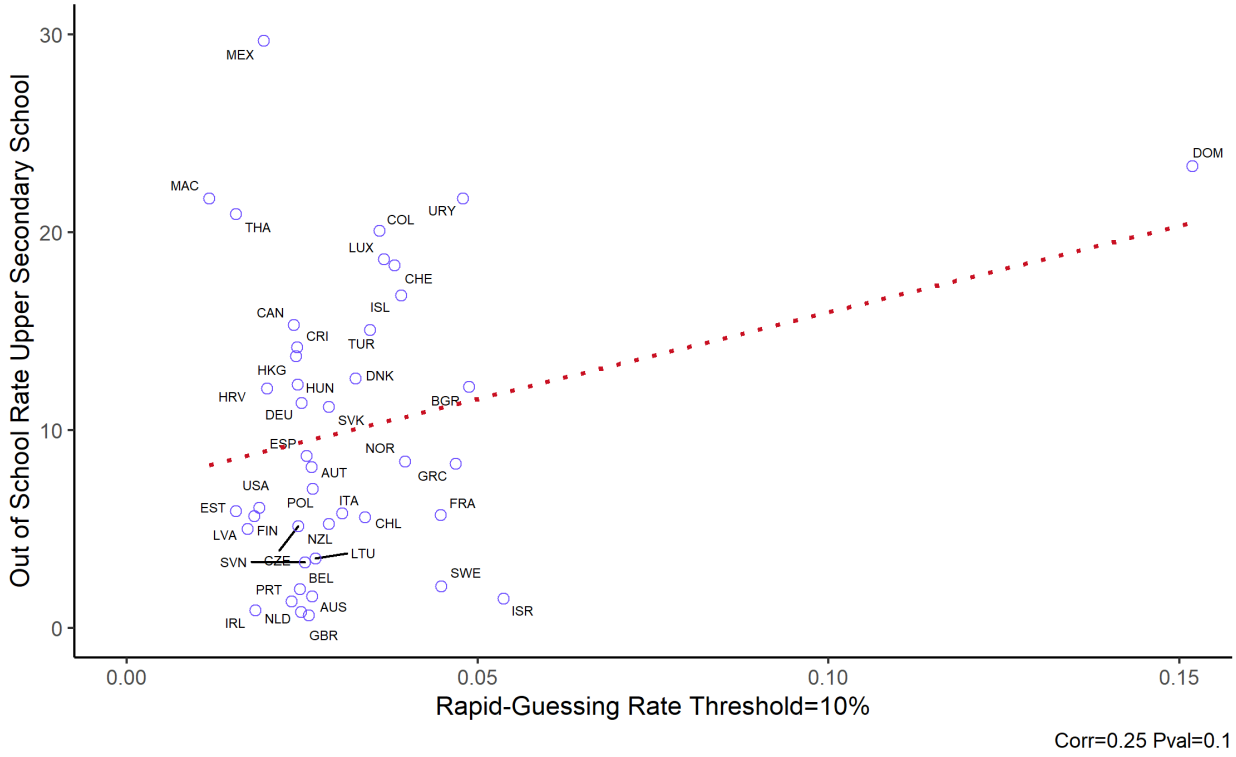


Figure 7: Relationship between effort measures and out-of-school rate (upper secondary)

(a) Rapid-guessing



(b) Item non-response

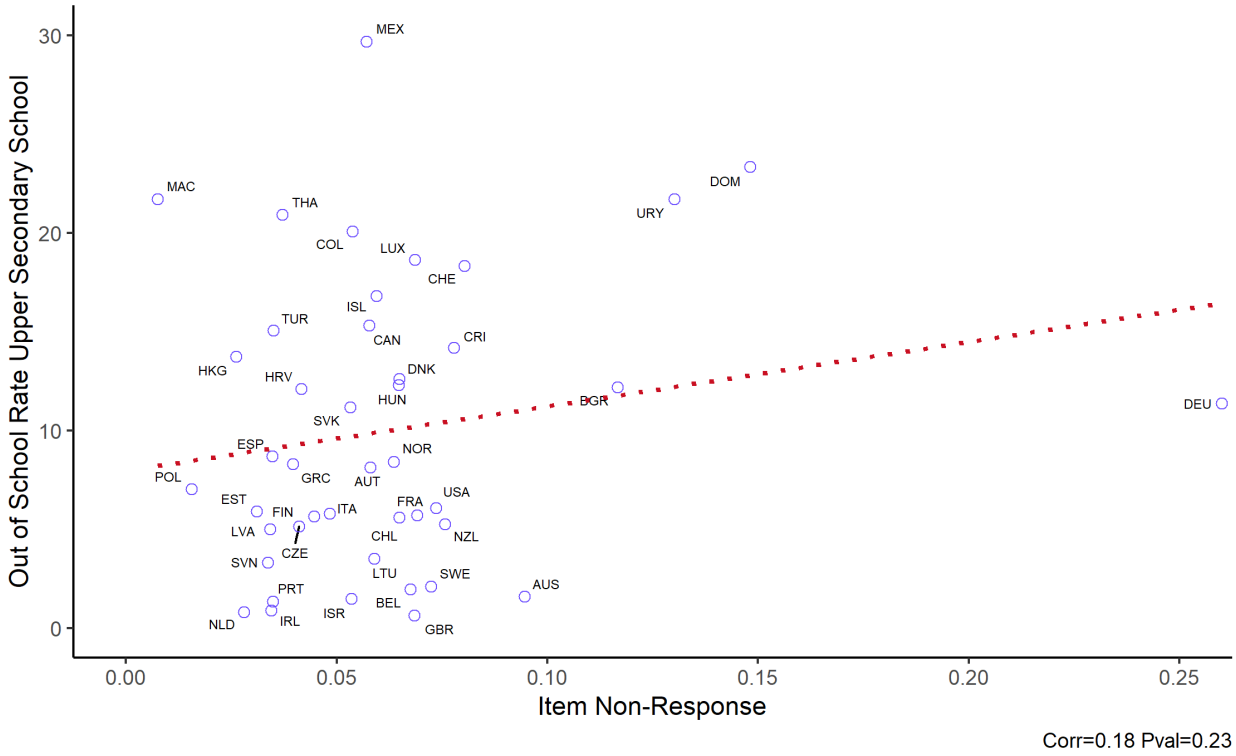
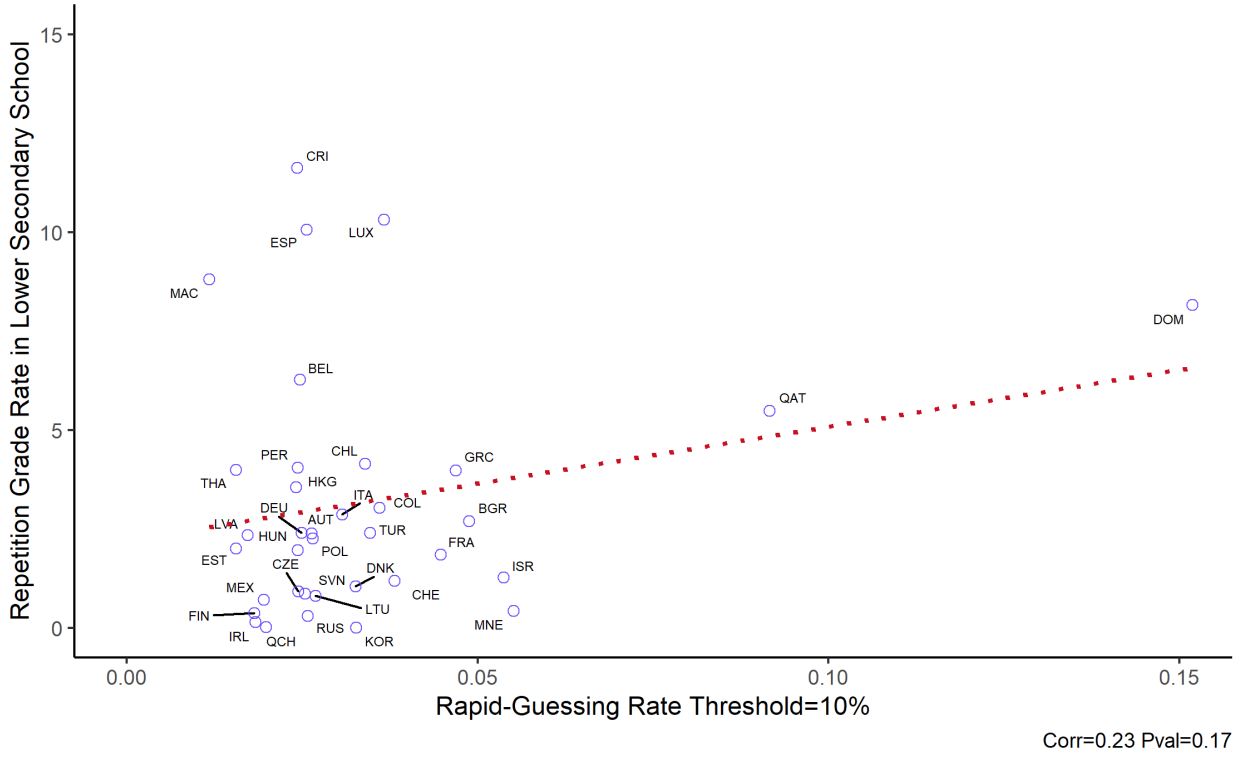


Figure 8: Relationship between effort measures and rate of grade repetition (lower secondary)

(a) Rapid-guessing



(b) Item non-response

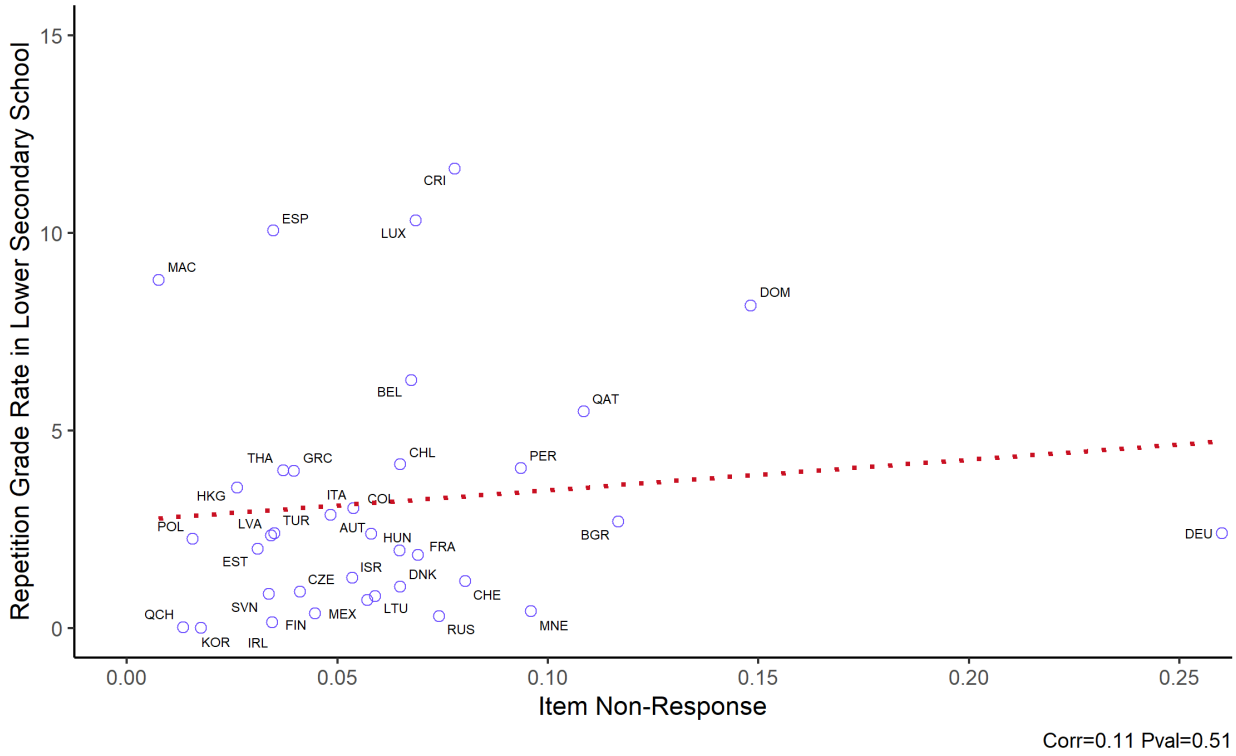


Figure 9: Accuracy rates for rapid-guessing and solution behavior – 10 percent threshold

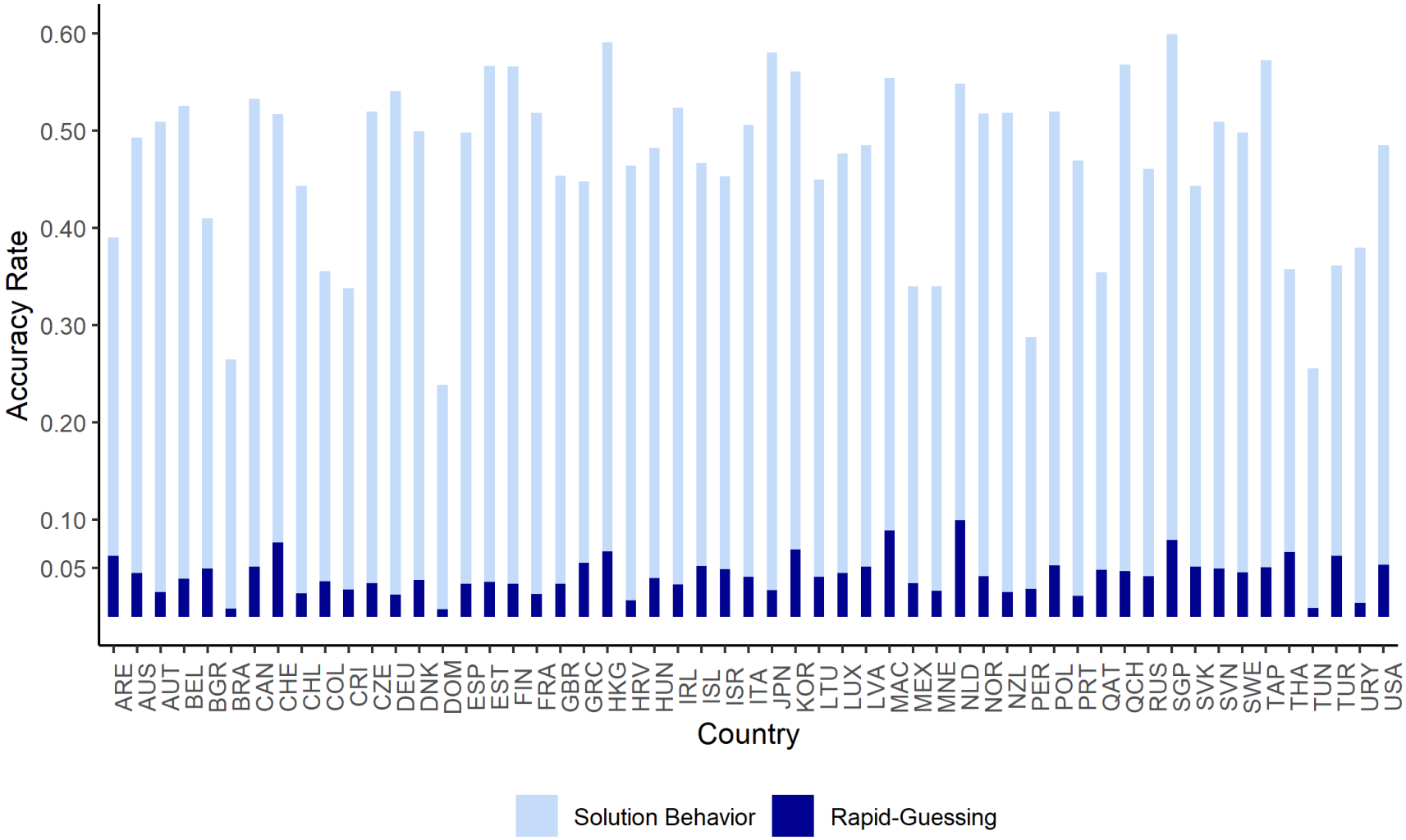


Table 1: Country names and abbreviations in PISA 2015

Abbreviation	Country Name	Abbreviation	Country Name
SGP	Singapore	ESP	Spain
JPN	Japan	LVA	Latvia
EST	Estonia	RUS	Russia
TAP	Chinese Taipei	LUX	Luxembourg
FIN	Finland	ITA	Italy
MAC	Macao	HUN	Hungary
CAN	Canada	LTU	Lithuania
HKG	Hong Kong	HRV	Croatia
QCH	B-S-J-G (China)	ISL	Iceland
KOR	Korea	ISR	Israel
NZL	New Zealand	SVK	Slovak Republic
SVN	Slovenia	GRC	Greece
AUS	Australia	CHL	Chile
GBR	United Kingdom	BGR	Bulgaria
DEU	Germany	ARE	Arab Emirates
NLD	Netherlands	URY	Uruguay
CHE	Switzerland	TUR	Turkey
IRL	Ireland	THA	Thailand
BEL	Belgium	CRI	Costa Rica
DNK	Denmark	QAT	Qatar
POL	Poland	COL	Colombia
PRT	Portugal	MEX	Mexico
NOR	Norway	MNE	Montenegro
USA	United States	BRA	Brazil
AUT	Austria	PER	Peru
FRA	France	TUN	Tunisia
SWE	Sweden	DOM	Dominican Republic
CZE	Czech Republic		

Table 2: Summary statistics of the variables of interest

Variable		Mean	Std. Dev.	Min	Max
Rapid guessing % - test	Overall	3.4	7.7	0.0	100.0
	Between		2.1	1.2	15.2
	Within		7.4	-11.8	101.6
Item non-response % - survey	Overall	7.3	17.0	0.0	97.9
	Between		4.9	0.6	26.0
	Within		16.0	-18.7	104.5
Total time - test (min)	Overall	78.9	20.0	0.0	120.0
	Between		6.3	60.1	98.9
	Within		19.0	-9.5	135.6
Math score	Overall	471.2	97.9	113.4	826.3
	Between		50.8	331.7	557.7
	Within		82.5	87.1	807.6
Reading score	Overall	474.2	99.1	54.3	812.0
	Between		41.9	360.6	530.6
	Within		89.3	15.3	822.1
Science score	Overall	476.3	99.6	133.4	831.3
	Between		45.1	335.2	548.0
	Within		88.2	131.7	816.5
Observations	Overall student sample			N = 291,521	
	Between countries			n = 55	
	Within-country average sample			Tbar = 5,300.38	

Note: excludes observations with total time above 120 min

Table 3: Descriptive gender differences on the variables of interest

Average	Boys	Girls	Difference
Rapid-guessing % - test	3.9	2.9	1.1***
Item non-response % - survey	8.1	6.4	1.7***
Total time - test (min)	76.4	81.3	-4.9***
Math score	475.8	466.6	9.2***
Reading score	461.6	486.7	-25.1***
Science score	478.4	474.2	4.1***
Total observations	145,394	146,127	

Note: excludes observations with total time above 120 min; *** $p<0.01$.

Table 4: Average estimated coefficients of the role of student effort on PISA test scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Math score			Science score			Reading score		
Item non-response survey	-0.18***		-0.12***	-0.20***		-0.13***	-0.23***		-0.16***
	(0.002)		(0.002)	(0.002)		(0.002)	(0.002)		(0.002)
Rapid-guessing test		-0.29***	-0.26***		-0.32***	-0.29***		-0.34***	-0.30***
		(0.002)	(0.002)		(0.002)	(0.002)		(0.002)	(0.002)
Constant	0.05	0.06	0.05	0.03	0.04	0.03	0.03	0.04	0.03
	(0.057)	(0.054)	(0.050)	(0.051)	(0.046)	(0.043)	(0.048)	(0.044)	(0.042)
Observations	296,832	291,521	291,521	296,832	291,521	291,521	296,832	291,521	291,521
Number of countries	55	55	55	55	55	55	55	55	55
R-squared within model	0.04	0.11	0.12	0.05	0.13	0.14	0.06	0.13	0.16
R-squared overall model	0.07	0.12	0.16	0.07	0.14	0.18	0.08	0.15	0.19
R-squared between model	0.28	0.35	0.44	0.25	0.41	0.48	0.23	0.36	0.43
Min student sample size	2,368	2,347	2,347	2,368	2,347	2,347	2,368	2,347	2,347
Max student sample size	16,224	15,964	15,964	16,224	15,964	15,964	16,224	15,964	15,964
Average student sample size	5,397	5,300	5,300	5,397	5,300	5,300	5,397	5,300	5,300

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All coefficients are standardized

Table 5: Effort-adjusted and unadjusted math scores and percentage change in the gap

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Qatar	408.1	396.7	91.7	431.7	453.5	11.4	-21.9	33.3	-36.3
United Arab Emirates	432.2	424.2	89.3	444.9	455.8	8.0	-10.9	19.0	-21.2
Greece	458.7	463.6	81.1	473.8	487.3	-4.9	-13.5	8.6	-10.6
Israel	469.7	478.0	95.7	488.3	506.3	-8.4	-18.1	9.7	-10.1
Montenegro	416.8	418.9	78.8	439.3	449.1	-2.0	-9.8	7.8	-9.9
France	496.6	500.2	88.1	513.3	525.0	-3.6	-11.7	8.1	-9.2
Bulgaria	443.7	442.8	90.0	465.5	472.1	0.9	-6.6	7.5	-8.4
Korea	528.3	521.9	92.7	537.2	538.2	6.4	-1.0	7.5	-8.1
Turkey	413.5	422.0	73.3	425.3	439.8	-8.5	-14.4	5.9	-8.1
Sweden	496.7	494.4	83.2	514.5	518.9	2.2	-4.5	6.7	-8.0
Lithuania	475.5	473.4	81.1	486.0	490.2	2.1	-4.2	6.3	-7.8
Iceland	489.8	490.4	85.3	504.7	512.0	-0.7	-7.2	6.6	-7.7
Hong Kong	552.0	553.3	82.6	559.5	566.6	-1.3	-7.1	5.9	-7.1
Slovenia	496.7	501.0	79.8	505.1	515.0	-4.3	-9.9	5.5	-7.0
Uruguay	413.9	429.0	82.0	437.7	458.4	-15.0	-20.7	5.7	-6.9
Latvia	485.8	486.2	70.3	492.2	497.1	-0.3	-4.9	4.6	-6.5
Norway	501.7	501.7	79.4	517.8	522.8	0.0	-4.9	4.9	-6.2
Germany	502.3	522.2	80.8	527.8	552.7	-19.9	-24.9	5.0	-6.1
Luxembourg	480.5	494.4	88.3	495.8	515.1	-13.9	-19.3	5.4	-6.1
Canada	500.2	509.1	78.5	510.7	523.9	-8.9	-13.2	4.2	-5.4
Slovak Republic	477.1	485.2	87.3	488.8	501.4	-8.0	-12.7	4.6	-5.3
Netherlands	519.0	521.7	80.9	527.7	534.6	-2.7	-6.9	4.3	-5.3
Ireland	494.9	512.0	74.3	501.9	522.8	-17.1	-20.9	3.8	-5.1
New Zealand	491.6	501.1	86.2	505.2	519.1	-9.5	-13.9	4.4	-5.1
Switzerland	513.1	525.2	88.5	530.2	546.6	-12.0	-16.4	4.3	-4.9
Poland	499.2	511.4	81.0	507.7	523.8	-12.2	-16.1	3.9	-4.8
United Kingdom	487.9	498.3	81.0	500.2	514.3	-10.3	-14.1	3.7	-4.6
Spain	481.6	500.1	77.4	491.4	513.5	-18.5	-22.1	3.6	-4.6
Austria	487.2	512.598	87.5	498.8	528.1	-25.4	-29.3	4.0	-4.5

(continued on next page)

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Chinese Taipei	538.8	543.3	96.9	544.3	552.6	-4.4	-8.3	3.9	-4.0
Croatia	460.5	472.8	82.5	469.0	484.5	-12.3	-15.5	3.2	-3.8
Estonia	518.5	525.2	74.8	524.8	534.4	-6.8	-9.6	2.8	-3.7
Thailand	429.6	427.6	84.2	436.3	437.5	1.9	-1.1	3.1	-3.7
Japan	526.2	541.6	82.0	535.0	553.1	-15.4	-18.2	2.8	-3.4
Czech Republic	501.7	509.1	85.4	511.9	522.2	-7.4	-10.3	2.9	-3.3
Denmark	496.3	508.1	77.5	511.3	525.6	-11.8	-14.2	2.4	-3.2
Hungary	481.7	490.2	85.6	493.9	504.8	-8.5	-10.9	2.4	-2.8
Chile	434.3	452.7	83.0	449.9	470.5	-18.3	-20.6	2.3	-2.7
Australia	482.0	486.7	87.9	497.0	504.1	-4.7	-7.1	2.3	-2.7
Colombia	392.0	405.7	70.2	407.7	423.2	-13.7	-15.5	1.8	-2.6
Russian Federation	490.9	499.1	75.1	504.5	514.5	-8.2	-10.0	1.8	-2.4
United States	465.6	476.0	82.4	476.7	489.1	-10.4	-12.4	1.9	-2.3
B-S-J-G (China)	541.2	545.8	95.6	548.0	554.8	-4.6	-6.8	2.2	-2.3
Portugal	475.4	487.0	89.9	485.3	498.9	-11.6	-13.6	2.0	-2.2
Mexico	411.6	419.1	66.6	421.8	430.8	-7.5	-8.9	1.4	-2.1
Belgium	506.3	525.3	88.2	519.1	539.8	-19.0	-20.7	1.7	-1.9
Costa Rica	394.8	411.8	62.3	408.6	426.7	-17.0	-18.1	1.1	-1.8
Tunisia	363.0	370.1	73.4	394.9	403.1	-7.2	-8.2	1.0	-1.4
Italy	489.8	510.9	84.0	503.6	525.7	-21.1	-22.1	1.0	-1.2
Brazil	367.8	383.5	77.6	397.0	413.0	-15.6	-16.0	0.4	-0.5
Dominican Republic	333.3	329.9	61.7	397.7	393.7	3.5	4.0	0.5	0.8
Peru	382.9	393.2	75.5	399.2	407.9	-10.3	-8.7	1.6	2.1
Macao	548.0	542.4	73.2	551.8	548.0	5.7	3.8	1.9	2.6
Singapore	559.6	555.9	90.0	565.5	565.0	3.7	0.5	3.2	3.6
Finland	515.7	508.7	75.3	522.9	520.6	7.0	2.3	4.7	6.2

Note: excludes outliers in total time above 120 minutes

Table 6: Effort-adjusted and unadjusted science scores and percentage change in the gap

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Qatar	429.4	406.4	94.4	455.9	470.4	22.9	-14.5	37.5	-39.7
Greece	465.9	462.1	85.8	482.9	488.8	3.8	-5.9	9.6	-11.2
Montenegro	412.5	408.4	80.5	437.7	442.4	4.1	-4.7	8.7	-10.8
Israel	468.1	474.3	101.4	489.1	506.2	-6.1	-17.1	10.9	-10.8
France	501.3	499.7	96.7	520.1	527.6	1.6	-7.5	9.1	-9.4
Turkey	426.4	422.2	72.9	439.7	442.2	4.2	-2.5	6.7	-9.1
Hong Kong	528.5	526.6	76.1	536.9	541.6	1.8	-4.8	6.6	-8.7
Iceland	475.6	474.8	86.7	492.4	499.0	0.7	-6.6	7.4	-8.5
Sweden	497.3	493.6	98.0	517.3	521.1	3.6	-3.9	7.5	-7.7
Uruguay	433.8	443.1	83.7	460.5	476.1	-9.3	-15.7	6.4	-7.6
Slovenia	501.9	496.6	89.6	511.4	512.3	5.3	-0.9	6.2	-7.0
Luxembourg	480.0	490.5	96.9	497.2	513.7	-10.5	-16.5	6.0	-6.2
Norway	495.5	500.3	93.0	513.6	523.9	-4.7	-10.3	5.6	-6.0
Germany	508.4	523.2	92.7	536.8	557.2	-14.9	-20.4	5.5	-6.0
Slovak Republic	467.1	468.5	92.3	480.2	486.8	-1.4	-6.6	5.2	-5.7
Canada	515.8	517.6	87.0	527.6	534.2	-1.8	-6.6	4.7	-5.5
Switzerland	498.9	505.0	94.7	518.1	529.0	-6.0	-10.9	4.9	-5.1
Netherlands	514.8	518.5	94.2	524.6	533.0	-3.6	-8.4	4.8	-5.1
Poland	498.8	505.6	86.3	508.4	519.6	-6.9	-11.2	4.4	-5.1
Ireland	496.8	508.5	85.2	504.7	520.7	-11.6	-15.9	4.3	-5.0
New Zealand	511.5	517.8	100.6	526.7	538.0	-6.3	-11.3	5.0	-4.9
Spain	492.0	502.6	82.8	503.1	517.6	-10.5	-14.6	4.0	-4.9
Austria	490.6	506.8	92.8	503.5	524.2	-16.2	-20.7	4.4	-4.8
Chinese Taipei	529.8	533.8	95.9	536.0	544.3	-3.9	-8.3	4.4	-4.6
United Kingdom	502.9	506.7	93.2	516.6	524.6	-3.8	-8.0	4.2	-4.5
Croatia	475.6	481.0	85.5	485.2	494.1	-5.4	-8.9	3.5	-4.1
Estonia	533.8	539.1	85.5	541.0	549.4	-5.3	-8.4	3.1	-3.7
Singapore	546.7	549.2	100.6	553.3	559.5	-2.5	-6.2	3.7	-3.6
Japan	532.5	548.0	89.1	542.5	561.1	-15.5	-18.6	3.1	-3.5

(continued on next page)

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Czech Republic	501.6	511.0	92.9	513.1	525.7	-9.4	-12.6	3.2	-3.5
Denmark	487.2	495.5	89.9	504.1	515.2	-8.3	-11.0	2.7	-3.0
Hungary	483.4	488.3	89.4	497.1	504.7	-4.9	-7.6	2.7	-3.0
Chile	458.7	474.3	84.6	476.2	494.4	-15.6	-18.2	2.5	-3.0
Colombia	419.2	432.6	75.7	436.9	452.2	-13.3	-15.3	2.0	-2.6
Australia	499.9	501.0	100.6	516.7	520.5	-1.1	-3.8	2.6	-2.6
B-S-J-G (China)	526.3	534.1	94.9	534.0	544.2	-7.8	-10.2	2.4	-2.6
Portugal	484.0	495.7	89.3	495.0	509.0	-11.8	-14.0	2.3	-2.5
Russian Federation	484.0	490.0	79.3	499.3	507.3	-6.0	-8.0	2.0	-2.5
Mexico	418.0	427.3	66.0	429.5	440.4	-9.4	-10.9	1.6	-2.4
United States	493.6	504.0	93.9	506.0	518.6	-10.4	-12.6	2.2	-2.3
Belgium	503.1	517.7	94.2	517.5	534.0	-14.6	-16.5	1.9	-2.0
Costa Rica	413.0	430.0	66.4	428.4	446.6	-17.0	-18.2	1.3	-1.9
Tunisia	384.1	387.6	58.8	420.0	424.6	-3.5	-4.6	1.1	-1.9
Italy	484.4	501.7	85.2	500.0	518.4	-17.3	-18.4	1.1	-1.3
Brazil	396.9	401.9	81.3	429.4	434.9	-5.0	-5.4	0.4	-0.5
Dominican Republic	334.1	336.3	66.9	406.7	408.2	-2.1	-1.5	0.6	0.9
Peru	392.5	403.8	72.1	410.7	420.2	-11.3	-9.5	1.7	2.4
Macao	533.4	527.1	77.3	537.6	533.4	6.3	4.2	2.1	2.7
Thailand	437.5	429.5	84.9	445.1	440.5	8.0	4.6	3.5	4.1
Finland	542.1	522.8	91.0	550.3	536.2	19.4	14.1	5.3	5.8
Latvia	498.1	488.2	77.1	505.2	500.4	9.9	4.8	5.1	6.6
Lithuania	472.9	465.4	87.6	484.6	484.3	7.5	0.3	7.1	8.1
Bulgaria	455.9	441.7	97.4	480.4	474.7	14.2	5.7	8.5	8.7
Korea	522.0	513.0	90.7	532.0	531.4	9.1	0.6	8.4	9.3
United Arab Emirates	450.9	424.3	96.5	465.2	459.9	26.6	5.3	21.3	22.1

Note: excludes outliers in total time above 120 minutes

Table 7: Effort-adjusted and unadjusted reading scores and percentage change in the gap

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
Peru	403.5	396.0	84.2	424.0	414.4	7.5	9.6	2.1	-2.5
Dominican Republic	377.0	347.6	78.5	454.6	424.7	29.4	29.9	0.5	-0.7
Brazil	415.7	393.9	89.2	453.2	432.0	21.8	21.3	0.5	0.6
Italy	501.1	485.6	83.9	518.0	503.8	15.5	14.2	1.3	1.6
Costa Rica	436.1	419.1	74.8	453.2	437.5	17.0	15.7	1.3	1.7
Tunisia	371.5	347.6	72.9	410.7	388.2	23.9	22.5	1.4	1.9
Belgium	513.7	501.3	90.9	529.5	519.3	12.4	10.2	2.2	2.4
Mexico	438.9	423.6	70.6	451.6	438.1	15.3	13.5	1.8	2.5
United States	508.6	489.9	92.8	522.5	506.2	18.7	16.3	2.4	2.6
B-S-J-G (China)	517.8	498.9	97.9	526.0	509.6	19.0	16.4	2.6	2.6
Colombia	443.6	431.6	82.6	462.6	452.8	12.0	9.8	2.3	2.7
Portugal	493.9	479.2	87.2	505.9	493.6	14.7	12.2	2.4	2.8
Russian Federation	509.0	482.8	80.8	525.8	501.9	26.2	23.9	2.3	2.8
Macao	525.3	494.5	76.3	529.7	501.2	30.8	28.5	2.2	2.9
Australia	509.3	477.3	98.3	528.1	499.0	32.0	29.0	2.9	3.0
Hungary	490.1	466.5	89.2	505.3	484.5	23.6	20.7	2.9	3.2
Chile	482.5	471.1	83.5	501.6	493.0	11.4	8.6	2.8	3.4
Czech Republic	514.3	488.6	95.2	526.7	504.6	25.7	22.1	3.5	3.7
Denmark	502.4	484.0	83.0	520.8	505.5	18.4	15.3	3.1	3.7
Japan	523.1	512.3	84.9	533.6	526.1	10.8	7.5	3.3	3.9
Singapore	540.0	517.2	93.1	547.1	528.1	22.9	19.1	3.8	4.1
Estonia	535.0	507.8	82.1	542.8	519.0	27.2	23.7	3.5	4.3
Thailand	433.3	404.2	82.7	441.7	416.3	29.0	25.3	3.7	4.5
Croatia	502.4	475.2	85.2	513.0	489.6	27.3	23.4	3.9	4.5
Chinese Taipei	509.7	484.5	87.7	516.2	495.5	25.3	20.7	4.6	5.2
United Kingdom	506.1	488.1	87.2	521.3	507.9	18.0	13.4	4.6	5.3
Austria	498.9	476.4	94.4	513.0	495.5	22.5	17.5	5.0	5.3
Spain	507.9	491.5	79.8	519.9	507.7	16.3	12.1	4.2	5.3
Poland	521.4	492.2	83.3	531.6	507.0	29.2	24.6	4.6	5.5

(continued on next page)

	Effort-unadjusted score			Effort-adjusted score					
Country	**Girls**	**Boys**	**SD**	**Girls**	**Boys**	**Unadjusted gap**	**Adjusted gap**	**Absolute difference**	**Change in gap as % of 1 SD**
New Zealand	526.7	495.1	99.1	543.5	517.4	31.7	26.1	5.6	5.6
Netherlands	523.7	500.6	90.0	534.2	516.3	23.0	17.9	5.1	5.7
Ireland	526.6	515.4	80.9	535.2	528.6	11.2	6.6	4.6	5.7
Switzerland	501.9	477.1	90.7	522.9	503.4	24.7	19.5	5.2	5.7
Slovak Republic	478.0	444.1	93.9	492.2	464.0	33.9	28.2	5.7	6.0
Canada	526.8	502.0	84.9	539.7	520.3	24.8	19.5	5.3	6.2
Luxembourg	493.0	474.3	100.5	511.8	499.5	18.7	12.2	6.5	6.4
Norway	532.5	494.5	92.5	552.1	520.2	38.0	31.9	6.1	6.6
Finland	551.8	504.9	87.0	560.8	519.7	46.9	41.1	5.7	6.6
Germany	525.3	507.6	89.5	558.3	547.0	17.6	11.3	6.4	7.1
Latvia	511.0	470.8	76.9	518.7	484.2	40.2	34.5	5.6	7.3
Uruguay	450.4	427.7	91.8	479.9	464.2	22.7	15.7	7.0	7.6
Slovenia	513.6	471.9	85.2	523.8	488.8	41.7	35.0	6.7	7.9
Bulgaria	459.3	413.2	107.5	486.3	449.3	46.1	36.9	9.2	8.6
Iceland	502.8	462.7	92.9	521.0	488.9	40.1	32.1	8.0	8.6
Sweden	522.1	483.4	94.4	543.8	513.3	38.7	30.5	8.2	8.7
Lithuania	485.3	447.8	88.8	498.2	468.4	37.6	29.8	7.8	8.7
Hong Kong	545.3	516.2	79.0	554.4	532.2	29.1	22.2	7.0	8.8
France	521.8	488.7	104.4	542.1	518.9	33.1	23.2	9.9	9.4
Turkey	440.2	414.6	74.6	454.5	436.0	25.6	18.5	7.1	9.6
Korea	539.8	499.2	90.3	550.4	518.7	40.6	31.7	8.9	9.8
Israel	494.6	471.0	105.4	517.1	505.1	23.6	12.0	11.6	11.1
Montenegro	442.0	410.1	86.8	469.3	447.1	31.9	22.3	9.7	11.1
Greece	493.5	461.3	88.4	511.5	489.8	32.1	21.7	10.5	11.8
United Arab Emirates	459.8	408.3	101.0	475.2	446.7	51.5	28.5	23.0	22.7
Qatar	428.9	375.9	104.8	457.3	444.9	52.9	12.4	40.5	38.6

Note: excludes outliers in total time above 120 minutes