NBER WORKING PAPER SERIES

TAKING PISA SERIOUSLY:
HOW ACCURATE ARE LOW STAKES EXAMS?

Ş. Pelin Akyol
Kala Krishna
Jinwen Wang

Working Paper 24930
http://www.nber.org/papers/w24930

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2018

We would like to thank Joris Pinkse, Keisuke Hirano and Kim Ruhl for their comments and suggestions and Meghna Bramhachari for help in proof reading. We owe special thanks to Miyako Ikeda and Judit Pal at the OECD for answering our numerous questions about the data. Huacong Liu was instrumental in our working on this project and we thank her for all her help. We are responsible for all errors. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Taking PISA Seriously: How Accurate are Low Stakes Exams?
Ş. Pelin Akyol, Kala Krishna, and Jinwen Wang
NBER Working Paper No. 24930
August 2018
JEL No. C53,I20,I21

ABSTRACT

PISA is seen as the gold standard for evaluating educational outcomes worldwide. Yet, as it is a low-stakes exam, students may not take it seriously resulting in downward biased scores and inaccurate rankings. This paper provides a method to identify and account for non-serious behavior by leveraging information in computer-based assessments in PISA 2015. We show that this bias is large: a country can rise up to 15 places in rankings if its students took the exam seriously. We ask where the bias is coming from and show that around half of it comes from the proportion of non-serious students, while 36% comes from their ability, with the remaining coming from the extent of non-seriousness

Ş. Pelin Akyol
Bilkent University
Department of Economics
06800 Ankara / TURKEY
pelina@bilkent.edu.tr

Kala Krishna
Department of Economics
523 Kern Graduate Building
The Pennsylvania State University
University Park, PA 16802
and NBER
kmk4@psu.edu

Jinwen Wang
Pennsylvania State University
jxw490@psu.edu

1 Introduction

Standardized tests are widely used to evaluate students, to rank countries in terms of educational outcomes, and to certify achievement. If the outcome of the test matters for the student taking it, the test is regarded as a high-stakes one, otherwise it is a low-stakes test. High-stakes exams motivate effort on the part of the student. However, to the extent that students have differential access to inputs that affect outcomes on the test, the resulting rankings may provide a biased picture of achievement. For example, well-off students tend to prepare for the SATs, often going to tutoring centers that show them how to raise their scores, while poor students may be less informed and less able to do so. For this reason, if the aim is to obtain a snapshot of where students are, then a low-stakes exam may be preferable to a high-stakes one.

However, the disadvantage of low-stake exams is that students may not take them seriously, so their performance on the exam may not reflect their true ability. As a result, scores from low-stake exams may be inaccurate. Correcting for this bias can be difficult. It is less of a problem if being non-serious is totally random and can be identified, as then one can confine oneself to the serious sub-sample. However, if effort during the test is related to ability, socioeconomic status and other characteristics, it is not obvious how one might correct for such bias. For example, if high-ability students are more likely to be non-serious in low-stake tests, then test scores would considerably underestimate average ability and underestimate the gap between low-ability and high-ability students.

In this paper we ask whether there is any evidence that the low-stakes environment is distorting the results obtained, albeit in a different way than a high-stakes exam would have done. Is there evidence that students are “blowing off” the test so that test scores do not serve as a valid measure of their knowledge level? We investigate these questions by using data from the Programme for International Student Assessment (PISA) which is a worldwide study organized by the Organization for Economic Cooperation and Development (OECD) in member and non-member countries. The aim of the exam is to have a common yardstick by which to measure students' performance in mathematics, science, and reading at age 15. PISA is often seen as the gold standard for such evaluations.[1] It is a low-stakes exam as the performance on the exam has no consequences for those taking the exam. We ask if the PISA results are biased, and if so, how we can adjust for these biases to be able to obtain a

[1] Other well known low-stakes tests include Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS). PISA assesses whether students can apply what they have learned to solve “real world” problems. PIRLS and TIMSS are grade-based (4th and 8th graders) and curriculum oriented. PISA is overseen by governments under the auspices of the Organization for Economic Cooperation while PIRLS and TIMSS are run by a consortium - International Association for the Evaluation of Educational Achievement (IEA).

reliable snapshot of student skills?

We use cognitive item response data from the 2015 PISA exam to investigate the differences in the effort level of students across countries. We focus on countries using the computer-based assessment (CBA) so that we have data on response time for each item[2], in addition to the response content for each item, whether it was correct or not and the order of the item. The exam is composed of four clusters of items with a short break after the first two clusters. As there is no negative marking (wrong answers are not penalized), there is no reason to skip a question: students should just guess even if they have no idea of the correct answer.

The skipping and timing data allows us to identify non-serious students as those who skip too many questions or spend too little time on too many questions, i.e., seem to not put reasonable effort into the exam. By definition, non-serious students on average spend less time than serious students but we find that this is especially so on items which they get wrong, suggesting that their inaccuracy is due to their spending less time on them. We note a marked fall in response time and accuracy with both position within a cluster and position of the cluster and this is more pronounced for non-serious students. These patterns suggest that we are truly identifying students who are not engaged in taking the exam.

We also investigate the factors correlated to effort on the part of students. The proportion of non-serious students as well as their characteristics vary substantially across countries. One might conjecture that non-serious students are those who are disenchanted with the system either because they feel it is not working for them (as when schooling does not seem to improve future prospects) or that it is abusing them (as with an over emphasis on testing). There is some evidence consistent with the second story. Overall, we find that students are less likely to take low-stake exams seriously if they have low ability, high socioeconomic status and spend excessive time studying. If they come from countries where standardized tests are associated with high stakes, their probability of being non-serious increases with the frequency of the test. The opposite is true for countries with low-stake standardized tests. We also find, not surprisingly, that more difficult questions as well as complex multiple choice and open response questions are more likely to be skipped than simple multiple choice ones, but less likely to have too little time spent on them.

We quantify the effects of non-serious behavior on the country rankings. We account for the bias of being non-serious by imputing the scores for skipped questions and for questions on which too little time is spent using multiple imputation techniques. We show that a country can improve its ranking by up to 15 places by encouraging its students to take the exam seriously. However, if all countries get their students to take the exam seriously,

[2] One item is one question. We use the word “item” or “question” interchangeably in the paper.

their actions cancel out and the maximum change in the rankings shrinks to 5. If all other countries' students become serious, the left out country falls in the rankings by at most 10 places. This change in ranking is not driven solely by the proportion of non-serious students, but also by their ability and the extent of their non-seriousness. There are countries with a large fraction of non-serious students (such as the Dominican Republic) who do not move up at all in their ranking because their non-serious students are of low ability.

We decompose the increase in the fraction correct of questions due to the imputation for each country into its component parts. Countries vary considerably in terms of the importance of these components. Across countries, 45% of the variation comes from the proportion of non-serious students, while 36% comes from their ability, with the remaining coming from the extent of non-seriousness.

We are not the first to point out that low-stakes exams might be inaccurate because they are not taken seriously. It has been recognized in the literature that low student motivation is associated with low performance (Pintrich and De Groot (1990), Wise and DeMars (2005), Cole et al. (2008), Penk and Richter (2017), and Jalava et al. (2015)), and students may not put their best effort in low-stakes exams (Wolf and Smith (1995), Duckworth et al. (2011), see Finn (2015) for a recent review). Attali et al. (2011) show that the stakes of an exam affect performance of students differentially according to socioeconomic status, gender and race. The difference between high and low-stakes exams is larger for males, whites and higher SES students. Similarly Azmat et al. (2016) find that women perform better than men in low-stakes exams, but as the stakes increase, this performance gap disappears.

Eklöf (2010) points out that it is important to take into account students' test-taking motivation especially on exams where the stake is low for the test-taker but high for the other stakeholders. Jacob (2005) documents that when the Iowa Test of Basic Skills was low-stakes, a large proportion of students left some questions blank despite there being no penalty for guessing. After it became a high-stakes exam, the percentage of questions answered increased by 1–1.5 percentage points, and the fraction correct of those answered also rose by 4–5 percentage points. This suggests that effort plays an important role in the performance of students.

Although the literature provides ample evidence on the relationship between effort, motivation and performance, there is little work that quantifies the effect of differential effort on the cross country rankings. Zamarro et al. (2016) attempt to explain the effects of differences in students' effort on the observed differences in country scores in the 2009 PISA exam. However, as this was not a computer-based assessment, they can only use the random ordering of questions, students' responsiveness to student survey questions and the consistency of these responses to tease out effort differences.[3] They then regress the score on their measures of effort and country fixed effects and argue that their measures of effort explain 32 to 38 percent of the observed variation in test scores across countries.

Gneezy et al. (2017) is the paper most closely related to ours. In an experimental environment, they incentivize U.S. and Shanghai students to increase their effort level and explore the effects of doing so on student performance. Their experiment has less than 500 students in the U.S. and less than 300 in China. The assumption is that student response in the experiment is what it would be if they had taken the PISA exam seriously. They show that incentives increased U.S. students' effort and performance, but did not affect the Shanghai students' performance. They then carefully project their experimental results on PISA data and estimate that the increased effort of U.S. students is equivalent to improving U.S. mathematics ranking in the 2012 PISA exam from 36 to 19. However, they are unable to do this for each country as their experiment is limited to two countries.

In contrast, we develop a simple way to control for non-serious behavior in both skipping questions and expending too little attention on them so that we can look at all countries. Our approach involves imputing the answers for each question skipped as well as each question deemed not taken seriously using multiple imputation methods. We show that just the share of non-serious students is not informative. To change rankings, non-serious students need to be good students so that they improve performance significantly when they become serious. We also show that a country can improve its rank by motivating their students, and that a country's rank falls a lot if other countries motivate their students while it does not. However, if all countries motivate their students simultaneously, rankings barely change. In other words, the game is close to a zero-sum one.

Our work contributes to the literature in two ways. Firstly, it extends the findings of Gneezy et al. (2017) to all countries by using some unique features of the PISA 2015 data. Computer based assessments allow us to better see how students respond to questions in terms of time spent and response content, which allows us to correct for non-seriousness without having to do an experiment for each country.[4] It analyzes the effects on scores and ranking if non-serious students behaved like serious ones for the 58 countries and areas that participated computer-based PISA exam in 2015. As a result we can do "partial equilibrium" analysis (one country is serious at a time) or general (all countries are serious together) and analyze the effect of being the left out one (all other countries are serious). Secondly, we investigate the factors that are related to low student effort across countries, find large

[3] One of their measures of effort is the extent to which performance falls over time. Another is the extent to which questions are skipped in the survey that students have to fill out and a third is the extent of carelessness in filling the survey.

[4] Our results turn out to be quite close to those of Gneezy et al. (2017) for the US and China.

differences across countries and explore some possible reasons for these differences.

The organization of the paper is as follows: The next section gives the necessary background about PISA exams. Section 3.1 discusses how we identify non-serious students and presents the data patterns. Section 4 investigates the factors correlated with being non-serious. Section 5 presents and discusses the effects of non-seriousness on rankings of countries. Section 6 decomposes the change in the fraction correct of each country after becoming serious and Section 7 concludes.

2 The PISA Exams

The PISA exams have been given every three years since 2000. In 2015 over half a million students participated in PISA exams, representing 28 million 15-year-olds in 72 countries and economies. For the first time in 2015, PISA was conducted as a computer-based exam, however the paper-based version was also available for countries that did not have the technical infrastructure needed[5]. As a result, 58 countries and economies took PISA 2015 in computer-based-assessment mode (CBA), accounting for 86.1% of the whole sample. In this paper, we will focus on these countries as only CBA items have data on the response time and the order of the questions, which we use below.

PISA is a two-hour exam[6]. It includes four 30-minute clusters, and students have 60 minutes for the first two clusters and 60 minutes for the last two with a 5-minute break in between (OECD (2015)). Each student gets different clusters based on a random number assigned to students [7]. Each cycle of PISA emphasizes one domain. While the emphasis was on reading in PISA 2009 and mathematics in PISA 2012 exam, the 2015 exam focused on science. Therefore, each student had two consecutive science clusters in the test, and they took these clusters either in the first hour or in the second hour of testing. According to OECD (2015), time is not a binding constraint for most students. On average students completed a cluster in around 18 minutes and 75% of students completed a cluster in less than 22 minutes. The PISA exam includes three types of questions: simple multiple choices, complex multiple choice [8] and open response. Each type accounts for approximately one third of all questions.

[5] In the 2012 PISA exam, 32 countries/regions were invited to complete both a paper and a computer version of mathematics test. However, by 2015, 58 moved to a computer based assessment. Jerrim (2016) and Jerrim et al. (2018) find that taking the PISA exam in a computer-based mode affects students' performance negatively in many countries.

[6] For countries that choose to implement the assessment of financial literacy, it requires an additional 60 minutes.

[7] For more detail see PISA 2015 Technical Report Chapter 2. (OECD (2015))

[8] One complex multiple choice question includes several yes-or-no questions.

PISA rankings are important to countries. Governments look at PISA scores to see where weaknesses lie in their educational systems. What is even more important, in some ways, is the role of PISA in providing the public with an objective view of how well their government is doing in this area. When PISA was given for the first time in India it was restricted to only a few states. When the results showed India to be second from the bottom in the rankings, with the average eighth grader in India at the level of a South Korean third grader in math abilities and a second grade student from Shanghai in reading skills, there was an uproar in the Indian press. The response of the Indian Government was to ban future PISA tests in India. In contrast, China was first in the PISA rankings in 2009 and 2012. However, in 2015, when for the first time students outside Shanghai took the exam, its rankings fell considerably, leaving Singapore in the first place. This suggests both that China is doing a far better job than India in training its youth on average, and that metropolitan dwellers are at a considerable advantage in this dimension. The U.S. consistently underperforms in PISA despite spending far more per pupil than China.

PISA 2015 also asked students and school principals to fill in questionnaires. The responses to the questionnaires, combined with the assessment results, can provide a broader and more nuanced picture of student, school and system performance. The student questionnaire seeks information about students and their family backgrounds, and aspects of students' lives such as their attitudes towards learning, their habits and life in and outside of school, and their family environment. The school questionnaire provides information on aspects of schools such as institutional structure, class size, learning activities in class, type and frequency of students' assessments.[9] Table A.5 in the Appendix contains descriptive statistics for the data used below.

In the next section, we describe how we identify the students who did not "take the exam seriously", i.e., put too little effort into it.

3 Serious versus Non-serious

3.1 Identifying Non-serious Behavior

We will distinguish between *serious students* and *non-serious students* below. Later on we will define *questions that are taken seriously* and those that are *not taken seriously* as we will impute the data for the latter.

It is natural to expect serious students to try and answer the questions to the best of their ability. There is no negative marking in PISA. For this reason, guessing is a dominant

[9] Some countries also have parent and teacher questionnaire.

strategy for multiple choice questions. Even if a student does not know the answer, and there is time remaining, the student is better off choosing some answer than leaving the answer blank. For open response questions, there may be no point in guessing as a continuum of answers exists. This is the first criterion for defining *who* is not serious and when a *question* is not being taken seriously. If a *student* leaves many multiple choice questions blank while having the time to complete them, the *student* is seen as being non-serious. Similarly, if a multiple choice question is left blank while the student has time left, the *question* is not taken seriously.

In the data, if a student spends some time on an item but does not answer it, this item is marked as *no response* (if this item is in the middle of the cluster) or *non reached* (if the item is in the end). In the PISA exam, students are not supposed to leave the room till the exam is over. If a student quits the test in the middle and thus does not spend any time on an item, this item is marked as *missing.*

We implement the definition of non-serious students as follows. A student is non-serious if too many items are unanswered (non reached, missing or no response) while there is ample time remaining (5 minutes) to attempt an additional question.[10] In each of the criteria below we set the cutoff so that no more than 10% of the students meet it.

Criterion 1. A student is non-serious if more than 5 minutes are left on the exam and there are K or more multiple choice questions *not reached* where K is set so that no more than 10% of the students meet this criteria. In the data $K = 1$. This criterion covers 4.2% of the students.

Criterion 2. A student is non-serious if more than 5 minutes are left and at least 2 or more multiple choice questions are marked as *no response.* This criterion covers 6.95% of students.

Criterion 3. A student is non-serious if more than 5 minutes are left on the exam and 3 or more questions (both multiple choice and open response) are *missing.* This identifies 9.33% of students as being non-serious.

Another requirement for a student to be serious should be reading each question and then formulating an answer. Students who spent so little time on a question that they could not have even read it, let alone formulated an answer, are also non-serious about that question. This criteria identifies non-serious students on the basis of their response time per item.

Response time data has been used as a measure of test-taking motivation in the education literature, see (Schnipke and Scrams (1997), Schnipke and Scrams (2002), Wise and Kong (2005)). Different methods have been applied to identify the items not taken seriously.

[10] There are roughly 60 minutes allocated for the two science clusters which have in total an average of 31 questions.

Schnipke and Scrams (1997) and Wise (2006) use methods based on the frequency distribution of the time spent on each item under the assumption that serious and non-serious students' response time distributions are different. Wise and Kong (2005) proposed a threshold selection method based on the item characteristics such as total length of item's stem and options. However, these methods do not take into account the ability of individuals. By using the same threshold for all test-takers, high-ability test-takers may mistakenly be labeled as non-serious. We identify non-serious students taking this issue into account as follows.

We first drop the 1181 students whose total time spent on the science part of the exam is 0 as there is no information in their responses. Then we remove outliers for each country in terms of response time, following Chapter 9 in the technical report (see OECD (2015), Leys et al. (2013)). Outliers are defined as those whose total response time on the science part of the exam is too large: i.e., if student i's total response time, R_i, exceeds $[mean + \ 3 * median(\|(R_i - median)\|)]$. The median and mean are of course country specific. The purpose of this step is to remove students whose total time is too large, possibly due to technical issues. This cutoff is typically larger than the total time allowed for this part of the exam. In this step, we drop 5034 students. In total, these 6215 students account for 1.39% of the sample.

Following this, we mark the item for a student in a country as a too-little-time item if the response time of item j, r_j, is less than $[mean - \ 2.5 * median(\|(r_j - median)\|)]$. The median and mean are again country specific. This method is similar to that used in setting thresholds suggested in Wise and Ma (2012).

A student spends too little time on an item either because he is randomly guessing an answer or because he easily gets the true answer. If the latter is the case, then we would be mislabeling smart students as non-serious.[11] We make sure we avoid such mislabeling as follows.

Criterion 4: A student is non-serious if he spends too little time on at least 3 answered items and the fraction correct for too-little-time items is lower than that for normal-time ones. This identifies 8.93% of students as being non-serious.

We use the union of these four criteria, and identify 25.69% of the students in the sample as non-serious students. There is considerable variation in the fraction of non-serious students across countries with Brazil and the Dominican Republic having over 50% non-serious. The fraction of non-serious students by country can be found in the last column of Table 7.

[11] This is indeed an issue as high-ability students (those with high scores) have a higher fraction correct for too-little-time items than that for normal-time ones, while the opposite is true for low-ability students.

3.2 Behavior Patterns of Serious versus non-serious Students

We identify non-serious students based on the four criteria above. In this section we first describe the patterns in the data and compare the behavior patterns of non-serious students to serious ones and see how they differ. We do so to assure ourselves that the students we are identifying as non-serious have behavior patterns that we might expect to see if they were truly not engaged in taking the exam.

A strong feature of the data across all countries is that both time spent and accuracy fall with item order and jump back up after a break. In addition, this seems to be more so for non-serious students. This pattern is consistent with student "fatigue".[12] This pattern is depicted in Figure 1 and 2 where we depict the median time spent and mean accuracy respectively per item as a function of item order. Time spent on *each* question (by all students who are faced with the question and who spend some time on it, whether or not they answer it) is standardized so it has mean zero and variance 1. If a student spends no time on an item, it is "missing" as described earlier and is dropped from this calculation. This standardization removes the impact of question characteristics, such as difficulty and question type, on time spent. The median of the standardized time is depicted for serious and non-serious students. We further decompose the non-serious student group by plotting the median time by each of the four criteria separately.

The standardized score for each question is constructed in a similar manner as follows. Each person either gets the question correct, partially correct or wrong, getting a score of 1,0.5 or 0 respectively. The standardized score for the question is then normalized with mean zero and variance 1 to account for differences between questions. We follow the PISA approach here and drop all questions that are *not reached* or are *missing* and put a score of 0 for questions marked as *no response.* For each position in a cluster, the average standardized score of the questions in that position is calculated. A lower average standardized score means the student's response is less accurate.

Time spent by serious students increases slightly within the first cluster. Then it falls sharply coming to the second cluster and remains stable in the rest of second cluster. The same pattern repeats for the third and fourth cluster. Time spent by non-serious students falls more sharply upon reaching the second and fourth clusters and continues to fall with item order within a cluster. The cost of skimping on time is accuracy since accuracy closely tracks time spent as is evident in Figure 2.

The heterogeneity among non-serious students according to the criterion used for classifi-

[12] It is worth noting that time is not a constraint in this exam. Less than 3% of students have less than 5 minutes left out of 60 minutes allocated for 2 clusters.

cation is also apparent.[13] In particular, non-serious students according to criterion 3 (missing items) spend even more time than serious ones when they answer a question. But looking at the total time spent on each cluster as in Table A.1, it becomes clear that they spend the most time of any group on the first cluster, but then spend the least time of any group on the second cluster. Moreover, this pattern is repeated in the third and fourth cluster. In other words, they are skipping most of the questions in the second and fourth clusters. Also note that as is evident in Table A.2, these students are more likely to answer correctly when they attempt a question than other non-serious students. All of this is consistent with their getting tired more quickly as the exam progresses, and getting reinvigorated during the break. Non-serious students according to criterion 2 and 4 (no response and little time) spend less time and have lower accuracy than non-serious students overall but the same pattern over item order is present.

Next, in Figure 3 we look at the time spent on correct and incorrect answers[14] by serious and non-serious students as the difficulty level (as measured by the fraction who got the question correct) rises. In contrast to Figure 1, here time spent is *conditional* on having answered the question. We argue below that the patterns here are consistent with serious students trying to figure out questions when they are not sure of the answer (even if they get them wrong) while non-serious ones (other than those with missing items) just take their best guess.

Time spent does not rise with difficulty for wrong answers for both serious and non-serious students, but does rise with difficulty for *correct* answers. Moreover, non-serious students spend about the same time as serious ones for incorrect answers but spend more time for correct answers as shown in Figure 3. Though non-serious students spend more time per question, overall, they spend less time per cluster[15] as they answer fewer questions. Figure 4 shows that students with missing items drive this result as they spend more time on all questions they attempt.

Removing these students from the non-serious group as in Figure 5 shows that non-serious students spend roughly the same time as serious ones when they get the answer correct (top panel), but spend less time when they get it wrong (bottom panel). Serious students spend roughly the same time on a question independent of whether they get it right or wrong,

[13] We did not plot time spent on the last 3 items for missing-item students because they miss these items by definition

[14] To do so we regress time spent on each item on type of question (multiple choice or open ended), position within a cluster and position of the cluster. We then remove the effect of question type, position and cluster to get the residual for each student and question. We plot the residuals for correct and incorrect answers for serious and non-serious students. We do not include individual fixed effects in the regression as we wish to see how serious and non-serious students differ in their responses.

[15] Serious students spend 19.5 minutes per cluster while non-serious ones spend 17.8 minutes per cluster.

while non-serious ones spend less time on questions they get wrong.

4 What Drives Being Non-serious?

We have seen in Section 3.2 that serious and non-serious students behave very differently. The next question is, what factors correlate with being non-serious? We explore this in two levels. First we look at the correlates of *individuals* being non-serious. After this, we look at correlates of the *question* not being taken seriously.

4.1 Who is Non-serious?

The factors[16] that correlate with a student being non-serious are explored in Table 1. Column 1 shows the results for all countries. The dependent variable is 1 if the student is non-serious. In columns 1 to 3, being non-serious is defined as meeting at least one of criterion 1, 2 or 4. In column 4, being non-serious is defined as meeting criterion 3. We make this distinction because the patterns explored in the previous section differ across these two groups. We also look at high-stake countries, ones where the standardized tests given in school are high-stakes[17], as well as low-stake countries as the patterns in the two might be different. If, for example, high-stakes exams result in exam fatigue while low-stakes ones do not, we might expect a higher probability of being non-serious in PISA in high-stake countries.

To begin with, we ask whether better students are more or less likely to be non-serious. Columns 1-3 suggest that higher math scores (a proxy for ability) are associated with a student being less likely to be non-serious, except when we use criterion 3, suggesting that students with missing items are a different breed. Students with high socioeconomic status (ESCS) and in lower grades are more likely to be non-serious. Again the sign in column 4 is reversed. This suggests that poor able students use criteron 3 when they are non serious while the rest use criterion 1,2 or 4.

Students from richer countries are more likely to be non-serious, though the shape is that of an inverted U with a turning point at about $33,000 for per capita GDP. However, this pattern is again reversed in column 4 where the pattern is U shaped with a turning point at about $38,500.

[16] The definition of the variables used and their summary statistics are in the Appendix.

[17] We calculate the stakes of standardized tests given in school as follows. In school questionnaire, school principles were asked whether the school used standardized tests for 11 different purposes. We mark the stake of each purpose to be betweens 1 to 3 and sum up the stakes for each school. Then we sort countries by their mean stakes and mark the top 36 countries as high-stake countries while the remaining 36 countries are marked as low-stake ones.

Gender matters: women are less likely to be non-serious in columns 1-3, but are more likely to be non-serious (by quitting in the middle of the exam) in column 4 suggesting that women "blow off" the exam in different ways than men. As might be expected, being anxious or ambitious is associated with being less likely to be non-serious, while being undisciplined, i.e., having a pattern of skipping class or arriving late, is associated with being non-serious.

One might speculate that students who are over-worked and over-tested, especially with high-stakes exams, have test fatigue and passively resist taking yet another test, and therefore are more likely to not take PISA seriously. There is some evidence in favor of this. First, countries with high-stakes exams do seem to make students work harder. On average students spend 1.3 hours more per week in class and 3.1 hour more on out-of-school learning in all subjects in high-stakes countries relative to low-stakes ones. Working harder seems to be associated with not taking PISA seriously. In column 1, spending more time on studies out of school is significant though time spent in school is positive but not significant.[18] Having more tests (standardized or teacher-developed) does seem to correlate positively with being non-serious though the coefficients are not significant.[19] Having more standardized tests raises the likelihood of being non-serious in high-stakes countries (column 2) but does the opposite in low-stakes ones (column 3). However, when teacher-developed tests are being given, raising the stakes seems to make students more likely to be serious, not less, suggesting that such testing may be less likely to result in test fatigue. Students from better schools, as reflected in the log of the school science score, are also less likely to be non-serious in low stakes countries, but more likely to be non-serious in high stakes countries. This makes sense if better schools push students more in high stakes countries resulting in fatigue.

Though we do not emphasize them in the next section, similar patterns in terms of individual characteristics, are observed in Tables A.3 (when we run individual fixed effects from the linear probability model on individual characteristics) and Table 3 (for the Logit model) when we consider the probability of skipping or spending too little time on a question as a function of student characteristics.

4.2 Which Questions are Not Taken Seriously?

We explore the effects of question characteristics on the probability of a question being skipped. We also do the same for the probability of too little time being spent on a question. In both cases we run a linear probability model with individual fixed effects as well as question characteristics. The results are presented in Table 2. It is easier to see the implications of the

[18] See column 1 and the row for time in class and out-of-school science learning.

[19] Again, the pattern is reversed in column 4 suggesting that students who don't take the test seriously by having missing items seem different.

interaction coefficients in a graph and Figure 6 and 7 depict these results. Figure 6 shows the probability of skipping a question and the probability of spending too little time on a question for each cluster as a function of the difficulty of the question. In all clusters, as the difficulty of the question increases, the probability of skipping increases (top panel) and the probability of spending too little time decreases (bottom panel). Students seem to try to answer if the question is easy but as it gets difficult, they simply skip it. There are also differences between clusters. Consistent with the "fatigue" hypothesis, the second and fourth clusters are where students are more likely to behave badly.

In Figure 7, we explore whether question type affects the probability of skipping or spending too little time as a function of question order. The probability of skipping rises with order, or sequence, in a cluster and jumps down at the beginning of the new cluster for all question types, consistent with "fatigue". Open response questions are most likely to be skipped as can be seen in the top panel of Figure 7. Complex and simple multiple choice questions follow the same pattern but the graph of complex multiple choice questions lies between the open response and simple multiple choice questions. This makes sense as it is easy to guess an answer for multiple choice questions so that they are less likely to be skipped.

In contrast, bad behavior in terms of spending too little time falls with the order of the question for both simple and complex multiple choice questions in the first and third cluster, but is weakly rising in the second and fourth. However, for open response questions, the probability of spending too little time always falls with order in a cluster and jumps up at the beginning of the new cluster. Students could behave non-seriously by either skipping and/or spending too little time. The above pattern suggests that for open response questions at least, as the exam proceeds, students substitute towards skipping with a reset at the end of each cluster. Hence we see a fall with sequence within a cluster and a jump up in each new cluster. The pattern is less clear for other question types.

In order to understand the effects of individual characteristics on the probability of being skipped or spending too little time, we run individual characteristics on estimated individual fixed effects from our linear probability model, see Table A.3. The results are in line with those of Table 1.

So far we ran choice regressions as if they were independent. However, the appropriate model is a multinomial choice one as the student has three mutually exclusive and exhaustive options for each question: skip, spend too little time or do not spend too little time answering it. We used the linear probability model as it allowed us to incorporate individual fixed effects, which we could not do with Logit. With logit, we can control for individual characteristics, but as we are unlikely to have information on all possible characteristics, we might have omitted variable bias.

Table 3 presents the results of a logit regression where the baseline choice is spending normal time answering the item. In the regression, we control for the question characteristics and the individual characteristics used in the previous tables. The first and second columns show the factors affecting the probability of skipping and the probability of spending too little time, respectively. The position within a cluster is positively correlated with the probability of skipping and negatively correlated with the probability of spending too little time, consistent with students switching from spending too little time to skipping as the exam progresses. If a question is in the second, third or fourth cluster relative to being in the first cluster, it is more likely to be skipped and this likelihood is much higher in the second and fourth clusters as they are the last clusters in each science session. Questions that require more effort to answer (difficult, open response or complex multiple choice ones) move students towards skipping and away from spending too little time. The coefficients on individual characteristics are roughly in line with those in Table 1. The math score of the student is negatively correlated with the probability of skipping and the probability of spending too little time. Female students students are less likely to skip or to spend too little time. Ambitious students are less likely to skip and more likely to spend too little time. Consistent with our previous findings, students from richer countries are more likely to skip and spent too little time, though the shape is that of an inverted U with a turning point at about $43,000 for per capita GDP. We control for standardized test frequencies and teacher developed test frequencies to investigate whether there is any evidence that students are fed up with testing, and as a result do not take them seriously. We find that as the frequency of the *standardized* tests increases, students likelihood of skipping and spending too little time significantly increases which is consistent with the "fatigue" effect. However, the *teacher-developed* tests do the exact opposite. This suggests that students view them very differently.

In the next section, we investigate the effects of non-seriousness on country rankings in PISA.

5 Effect on Rankings

Clearly, having students take PISA non-seriously will tend to reduce the average country score and adversely affect its rankings. In this section, we explain how we adjust scores to account for non-seriousness. We then present results that quantify the effect of non-serious students on country rankings. We also decompose the change in score into its component parts.

To correct the potential bias of being non-serious, we use Multiple Imputation by Chained Equations (MICE) to impute scores for all non-serious questions. Recall these were questions that were not reached, for which there was no response, were missing, or on which too little time was spent. All of these are treated as missing data. Multiple imputation involves filling in all the missing data multiple times, creating multiple complete data-sets which are then averaged over for the final imputation. The missing values are imputed based on the observed values for the given individual and the relations observed in the data for other participants (Schafer and Graham (2002)). The variables used for imputation for a given individual are laid out in Table A.4. They include the individual's scores for other science questions in the test, other participants' scores for all science questions, the individual's characteristics as well as school characteristics.

Since imputation attempts to assign values for missing data based on the responses for similar individuals/questions/schools, one needs to assume that the probability of being non-serious is random after controlling for all the observables.[20] In the MICE procedure a series of regression models are run whereby each variable with missing data is modeled conditional upon the other variables in the data. This means that each variable can be modeled according to its distribution (Azur et al. (2011)). In our model, whether a question is right or wrong and school type are binary variables, therefore they are modeled using a logistic regression and all other continuous variables are modeled using linear regressions.

A feature of PISA tests is that students get different clusters of questions. Even if two students have a common cluster of questions, the position of the cluster might differ. We have seen in Section 3.2 that the position of an item has an substantial effect on student's performance on this item. Imputation of an item's score has to use the relations for other individuals who answer the same item in the same position. In the PISA test, all students are assigned a random number which determines the specific science clusters included on the test as well as their positions. So we divide all students into 72 groups so that students in each group answer the same questions in the same order[21]. Then we conduct multiple imputations within each group before pooling all imputation results together at the end.

Next we describe how to calculate country rankings based on all students' item responses. As different students take different tests, PISA imputes plausible values for a common test using item response theory (IRT). This is a rather complex procedure that is carried out for PISA by the Educational Testing Service and is a bit of a black box as the codes are not freely available. Instead we use the raw score which is just the total number correct. We

[20] If this were not so, there would be no similar individuals/items/schools to impute from.

[21] There are 36 random numbers in total which determine the specific science clusters assigned to students. Moreover, students have science clusters either in the first two sessions or in the last two sessions. Therefore in total there are 72 groups within which students answer the same questions in the same order

then standardize this score for each group that got the same test so that their performance is comparable [22]. Then we calculate weighted average score (each student has a weight based on the stratification frame used) for each country and rank all countries. The correlation of rankings based on standardized raw scores with that based on plausible values is 0.99, so we feel we are on safe ground using our simpler approach.

In order to understand the effect of being non-serious on country rankings, we impute the data for items not taken seriously according to criteria 1-4. Table 4 presents the list of countries and their ranks before and after imputing the scores of non-serious students. Column 1 shows the original standardized raw scores and column 2 shows the rank based on this original score. Column 5 shows the rank if only this particular country becomes serious. Notice that countries always move up in the ranking in this thought experiment as their score can only rise with the imputation. Column 6 is the change in rank between the second and the fifth column. All these numbers are weakly positive, but vary a great deal. Singapore and Chinese Taipei (Taiwan) do not change rank, while Portugal moves up by 15 places suggesting that they have a major problem with students being non-serious. It is also clear that countries at the top and bottom of the original rankings tend to move less than countries in the middle. This arises from the score gap between sequentially ranked countries being large at the top and bottom and smaller in the middle. For example, Singapore has a score of .74 while the next ranked country, Taiwan has a score of .58. Similarly, the Dominican Republic has a score of -1.21 while its neighbor, Tunisia, has a score of -.92. Small wonder that Singapore stays first in all the columns and the Dominican Republic stays last.

Column 3 shows the rank when we do the imputations for all countries so that these rankings capture the situation where students are serious in all countries. As is evident, some countries rise in the rankings (Japan) while others fall (Slovenia). However, overall there is very small change in the rankings. This makes sense. If one country can get its students to be serious about the exam, it can change its ranking a lot. But if everyone does so, general equilibrium effects come into play and individual efforts are negated. Column 4 shows what happens when all other countries become serious and this particular country alone does not. As expected, each county's rank gets worse when all other countries get serious. Again, some countries are less affected than others. Singapore for example is unaffected even in this case, while Ireland would fall from 18 to 26 if this were to happen.

Table 5 zooms in on column 5 of Table 4 so that only one country is getting serious at a time. The objective here is to see what is driving the change in rankings. Column 1 in Table 5 is the same as column 2 of Table 4. Column 2 and 3 of Table 5 show how the rankings vary depending on what is imputed. In column 2 only missing/no response/not

[22] This method is suggested and used by Jerrim et al. (2017).

reached, i.e., unanswered items are imputed (criteria 1-3). In column 3, only items meeting criterion 4 (too little time) are imputed. As is evident, the unanswered item criteria seem to do most of the heavy lifting. For some countries, like Norway, all the changes seem to come from unanswered items being imputed. The fact that most of the action comes from skipped questions makes sense. When students skip questions, the increase in fraction correct is the fraction correct they would have obtained had they actually tried to answer the question, which depends on their ability. When students spend too little time, the increase in the fraction correct is this same fraction less .25 as even a random guess with say four choices gives the right answer 25% of the time. As a result, skipping will tend to drive most of the increase in fraction correct.

Next, we investigate why some countries improve their ranking a lot, while others do not.

6 Proportion, Ability and Extent

When we impute the data for unanswered questions and for questions not taken seriously, the fraction of questions correctly answered will typically rise. In this section we decompose the source of this increase in the fraction correct (y) into three component parts for each country and for serious and non-serious students separately. The first part depends on the *ability* (a) of the non-serious student. The more able the student, the more likely he is to get the question right and the greater the increase in the fraction correct when we make our corrections. The second part depends on how prevalent the imputed items are, i.e., the *extent* (e) to which these items occur. If they are very prevalent, then our imputation will have a greater impact. We expect them to be more prevalent for non-serious students than for serious students so that the correction will have more of an impact for the former. The third part depends on the *proportion* (p) of non-serious students in the population: the greater the fraction of non-serious students, the greater the increase in the fraction correct.

6.1 Sources of Increases in the Fraction Correct

Let T_i be the *total* number of items in student i's test as this is individual specific. Let C_i be the number *correct* for i in the data and $\hat{C}_i$ be the number correct with the *imputed* data. Let $I_i = \hat{C}_i - C_i$ denote the *increase* in student i's number correct if he was serious about all items. A country has S serious students and NS non-serious students. The fraction correct for this country in the data is

$$FC = \frac{\sum\limits_{i=1}^{S+NS} C_i}{\sum\limits_{i=1}^{S+NS} T_i}$$

while the fraction correct after imputation is

$$\hat{FC} = \frac{\sum\limits_{i=1}^{S+NS} \hat{C}_i}{\sum\limits_{i=1}^{S+NS} T_i}$$

If all students in this country became serious on all items, the increase in the average fraction correct for this country, IFC, can be expressed as:

$$\begin{align}
IFC &= \frac{\sum\limits_{i=1}^{S+NS} I_i}{\sum\limits_{i=1}^{S+NS} T_i} \notag \\
&= \frac{\sum\limits_{i=1}^{NS} I_i}{\sum\limits_{i=1}^{S+NS} T_i} + \frac{\sum\limits_{i=NS+1}^{NS+S} I_i}{\sum\limits_{i=1}^{S+NS} T_i} \tag{1} \\
&= \left(\frac{\sum\limits_{i=1}^{NS} I_i}{\sum\limits_{i=1}^{NS} T_i}\right) \frac{\sum\limits_{i=1}^{NS} T_i}{\sum\limits_{i=1}^{S+NS} T_i} + \left(\frac{\sum\limits_{i=NS+1}^{NS+S} I_i}{\sum\limits_{i=NS+1}^{NS+S} T_i}\right) \frac{\sum\limits_{i=NS+1}^{NS+S} T_i}{\sum\limits_{i=1}^{S+NS} T_i} \tag{2} \\
&= IFC_{ns} P_{ns} + IFC_s \left(1 - P_{ns}\right) \tag{3} \\
&= Y_{ns} + Y_s \tag{4}
\end{align}$$

where IFC_{ns}, and IFC_s is the increase in fraction correct for non-serious students and serious students respectively, and P_{ns} is the proportion of non-serious students in the population. In the PISA test, students have different numbers of science items, and this is determined randomly. Thus, on average, non-serious students have the same number of total items as serious students so that P_{ns} measures the proportion of non-serious students in a country. Thus, the increase in the fraction correct is a convex combination of the increase in the fraction correct for serious and non-serious students. It is worth noting that $\frac{Y_{ns}}{IFC}$ is 0.84 so that most of the increase comes from non-serious students.

Next we will decompose IFC_{ns} (and IFC_s) into their component parts. Let NI_i be the number of non-serious items student i has.[23]

$$\begin{aligned} IFC_{ns} &= \frac{\sum_{i=1}^{NS}(I_i)}{\sum_{i=1}^{NS}T_i} \\ &= \frac{\sum_{i=1}^{NS}(I_i)}{\sum_{i=1}^{NS}NI_i}\frac{\sum_{i=1}^{NS}NI_i}{\sum_{i=1}^{NS}T_i} \\ &= A_{ns}E_{ns} \end{aligned}$$

A_{ns} is the increase in the fraction correct for non-serious items among non-serious students. As explained below, we would expect this to be increasing in non-serious students' ability. E_{ns} is the fraction of non-serious items among all items for non-serious students, which measures the degree of non-seriousness for non-serious students.

Thus,

$$Y_{ns} = A_{ns}E_{ns}P_{ns}.$$

The values of Y, A, E and P for each country are provided in Table 7. Dividing both sides by the geometric mean gives

$$\begin{aligned} \frac{Y_{ns}}{\bar{Y}_{ns}} &= \left(\frac{A_{ns}}{\bar{A}_{ns}}\right)\left(\frac{E_{ns}}{\bar{E}_{ns}}\right)\left(\frac{P_{ns}}{\bar{P}_{ns}}\right) \\ y_{ns} &= a_{ns}e_{ns}p_{ns}. \end{aligned} \tag{5}$$

We de-mean to make sure the regressions below start from the origin. Take the logarithm on both sides of (5) gives:

$$\ln(y_{ns}) = \ln a_{ns} + \ln e_{ns} + \ln p_{ns} \tag{6}$$

If we want to know how much of the variation in $\ln Y_{ns}$ comes from each of the three components, we can use a simple trick. Suppose we run the regression of $\ln a_{ns}$, $\ln e_{ns}$, $\ln p_{ns}$

[23] Recall that non-serious items include non-reached, no-response and missing items, and items with too little time if a student spends too little time on at least three items and the fraction correct for little-time items is lower than that for normal-time ones.

separately on $\ln y_{ns}$, that is,

$$\begin{aligned}\ln a_{ns} &= \alpha_1 \ln y_{ns} + \epsilon_a \\ \ln e_{ns} &= \beta_1 \ln y_{ns} + \epsilon_d \\ \ln p_{ns} &= \gamma_1 \ln y_{ns} + \epsilon_{p.}\end{aligned}$$

[24] Let the OLS estimates be denoted by $\hat{\alpha}_1, \hat{\beta}_1, \hat{\gamma}_1$ and note that

$$\begin{aligned}\hat{\alpha}_1 \ln(y_{ns}) + \hat{\beta}_1 \ln(y_{ns}) + \hat{\gamma}_1 \ln(y_{ns}) &= \left(\hat{\alpha}_1 + \hat{\beta}_1 + \hat{\gamma}_1\right) \ln(y_{ns}) \\ &= \ln a_{ns} + \ln e_{ns} + \ln p_{ns} \\ &= \ln y_{ns}\end{aligned}$$

so that $\hat{\alpha_1} + \hat{\beta_1} + \hat{\gamma_1} = 1$ and we can use the coefficients $\hat{\alpha_1}$, $\hat{\beta_1}$, $\hat{\gamma_1}$ to measure the contribution of non-serious students' ability, extent of non-seriousness and proportion to a country's increase in fraction correct by non-serious students.

We can decompose the increase in the fraction correct coming from serious students (what we call partially serious and fully serious) in an analogous manner. Details are in the Appendix.

6.2 Results of the Decomposition

Table 6 summarizes the decomposition results of y_{ns} and y_s. Column 1 shows that for non-serious students, proportion accounts for 45% of the increase in fraction correct while ability accounts for about 36%, and least important is the extent of non-seriousness which accounts for only 19% of the variation. Column 2 shows that ability accounts for 36% of the variation for serious students, while proportion accounts for 27% and extent accounts for 37%.

Figure 8 plots the scatter plot and regression lines above for non-serious students. The countries with high y_{ns} tend to be those who would gain a lot from their students taking the exam seriously. Where does the gain come from? As is evident from the figure, Brazil stands to gain the most. This gain is driven by the large proportion of non-serious students and the high extent of non- seriousness. However, the contribution of ability is relatively small: even if the exam had been taken seriously, the performance would not have improved much as non-serious students in Brazil are of low ability. In contrast, both Russia and Portugal who also have high y_{ns} have the contribution of ability being high since their non-serious students

[24] These three regression lines add up to the 45^0 line.

are quite able. The Dominican Republic, which ranked at the bottom of all of our countries and which did not change its rankings in any of our counterfactuals, has y_{ns} at about the median. This is despite their extremely high proportion of non-serious students. The reason is that the ability of these students is very low. The U.S., Netherlands and Turkey are remarkable because their non-serious students' ability, extent and proportion roughly track their gains, though Turkey and the Netherlands gain little while the U.S. gains more.

7 Conclusion

The PISA exam which is seen as the gold standard for evaluating how countries are faring in terms of their education system is a low-stakes exam. As such, there is little incentive for students to take the exam seriously. It is well understood that this feature limits the accuracy of the results *and* biases the resulting rankings. However, there is (i) limited understanding of the factors that drive students to be non-serious, (ii) no attempt to quantify the score gains across a host of countries from students taking the exam seriously and the consequent effects on rankings, and (iii) no decomposition of score gains into their constituent parts.

This paper contributes in all three of these dimensions. With respect to the first contribution, we find, amongst other things, that the fraction of non-serious students varies enormously by country (from 13.6% in Korea to 67% in Brazil) and that low ability and high socio-economic status students tend to be more likely to be non-serious. Exam fatigue also seems to be consistent with the patterns we find: students who face numerous high-stakes exams and who spend long hours studying in and out of school tend to be non-serious about the PISA exam.

With respect to changes in score and rankings we find that even in countries with many non-serious students, the increase in the fraction correct and hence score is limited. For example in Brazil, where 67% of the students are non-serious, the increase in the fraction correct is only 3.6%. This is due to the ability of non-serious students being low, so even if they had tried, they would not have done much better! Nevertheless, the change in rankings can be quite substantial and is not the largest for those countries with the most non-serious students. For example, Portugal with 27% of students being non-serious, would gain the most in terms of rank change, (15 places out of 58 countries) if only its students became serious while Brazil gains only 1 place. If all other countries' students also became serious, Portugal would rise by 5 places in the rankings while Brazil would still rise by only one place. The difference in Brazil and Portugal comes partly from Portugal's non-serious students being of relatively high ability so they provide more leverage in terms of change in the fraction correct. In addition, as Brazil is close to last in the rankings, there is a wide gap between it and its neighbor which makes a rank change hard. Portugal, in contrast, is in the middle of the distribution and has a small gap in terms of its score and that of its neighbors'. These examples highlight the importance of doing the counterfactuals seriously rather than just looking at the fraction of non-serious students.

Finally, we decompose the source of the increase in fraction correct into the part that comes from the proportion, from ability and from extent (intensity). The U.S. would gain 1.58% in score and 5 places in rank (from 27th to 22nd) if its students alone took the exam seriously. Using a standard decomposition, we show that the contribution of the three components varies widely across countries.[25] We use a simple regression on this decomposition which shows that across all countries, roughly 45% comes from the proportion component, 36% comes from the ability component and 19% comes from the extent component.

This paper thus has a simple bottom line. Using PISA scores and rankings as done currently paints a distorted picture of where countries stand in both absolute and relative terms. Simple adjustments like those proposed here help provide a better picture.

References

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

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

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

Cole, J. S., Bergin, D. A., and Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, 33(4):609–624.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., and Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, 108(19):7716–7720.

[25] Table 7 shows that the increase in score coming from non-serious students (y_{ns}) is 1.29%. y_{ns} is the product of three components, with the ability, extent and proportion components being .2791, .2031, .2277 respectively. Note their product is .0129 or 1.29%.

Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, 17(4):345–356.

Finn, B. (2015). Measuring motivation in low-stakes assessments. *ETS Research Report Series*, 2015(2):1–17.

Gneezy, U., List, J. A., Livingston, J. A., Sadoff, S., Qin, X., and Xu, Y. (2017). Measuring success in education: the role of effort on the test itself. Technical report, National Bureau of Economic Research.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the chicago public schools. *Journal of public Economics*, 89(5-6):761–796.

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

Jerrim, J. (2016). Pisa 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, 23(4):495–518.

Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., and Shure, N. (2017). What happens when econometrics and psychometrics collide? an example using the pisa data. *Economics of Education Review*, 61:51–58.

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., and McKeown, C. (2018). Pisa 2015: how big is the 'mode effect'and what has been done about it? *Oxford Review of Education*, 44(4):476–493.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

OECD (2015). Pisa 2015 technical report. Technical report, OECD.

Penk, C. and Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, 29(1):55–79.

Pintrich, P. R. and De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology*, 82(1):33.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177.

Schnipke, D. L. and 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.

Schnipke, D. L. and Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. *Computer-based testing: Building the foundation for future assessments*, pages 237–266.

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.

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

Wise, S. L. and Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2):163–183.

Wise, S. L. and Ma, L. (2012). Setting response time thresholds for a cat item pool: The normative threshold method. In *annual meeting of the National Council on Measurement in Education, Vancouver, Canada.*

Wolf, L. F. and Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, 8(3):227–242.

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

Figure 1: Standardized Time for Serious and Non-serious Students

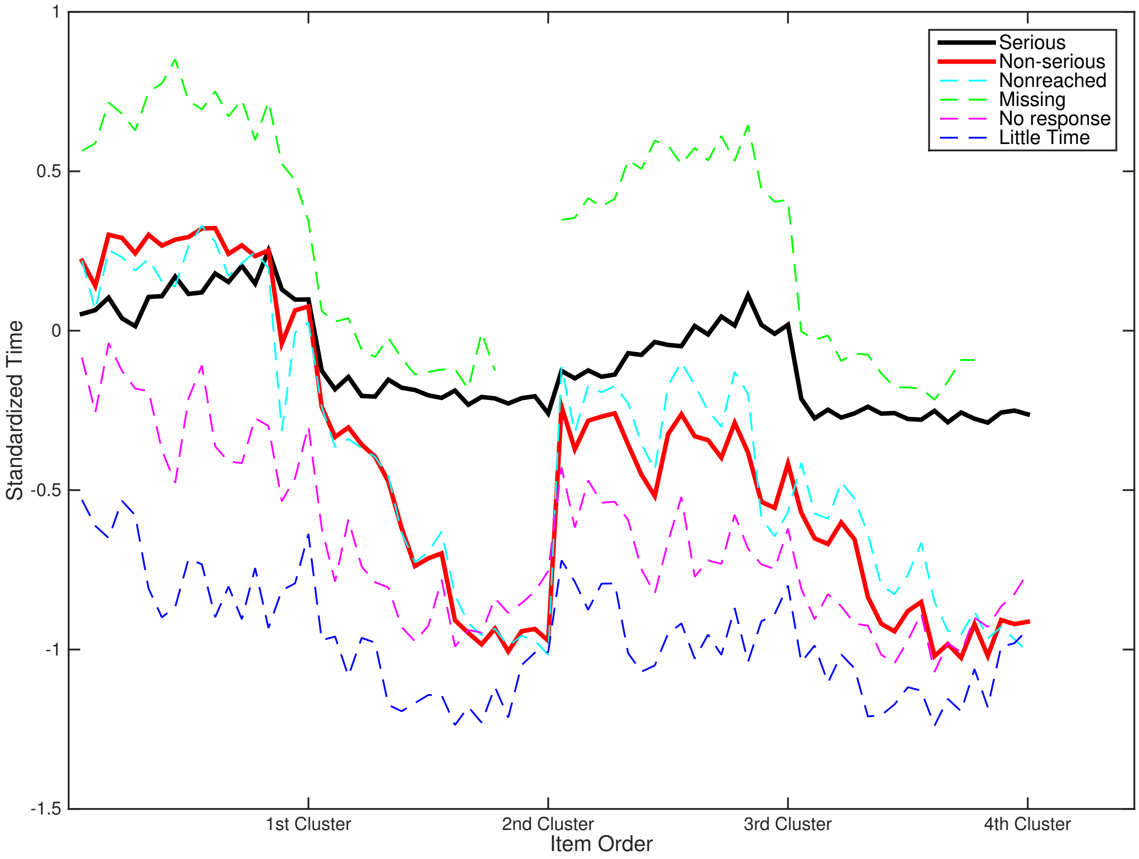


Figure 2: Standardized Score for Serious and Non-serious Students

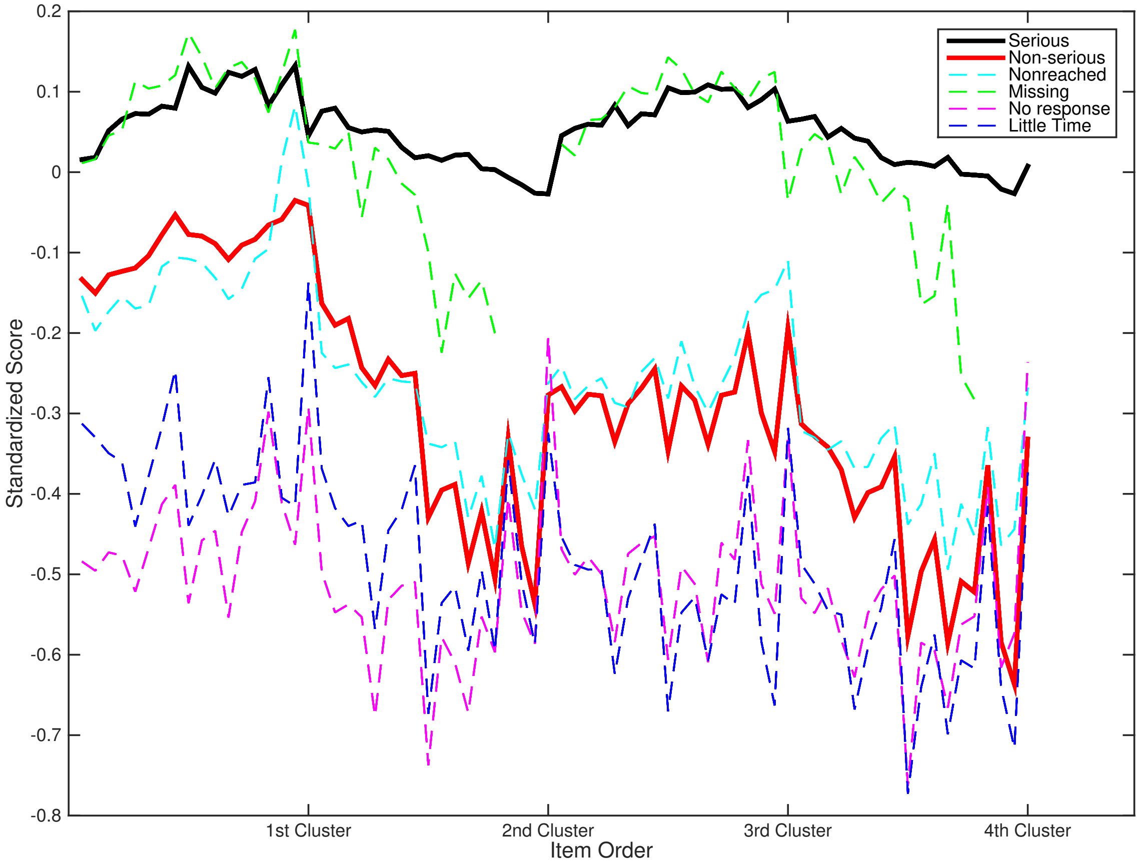
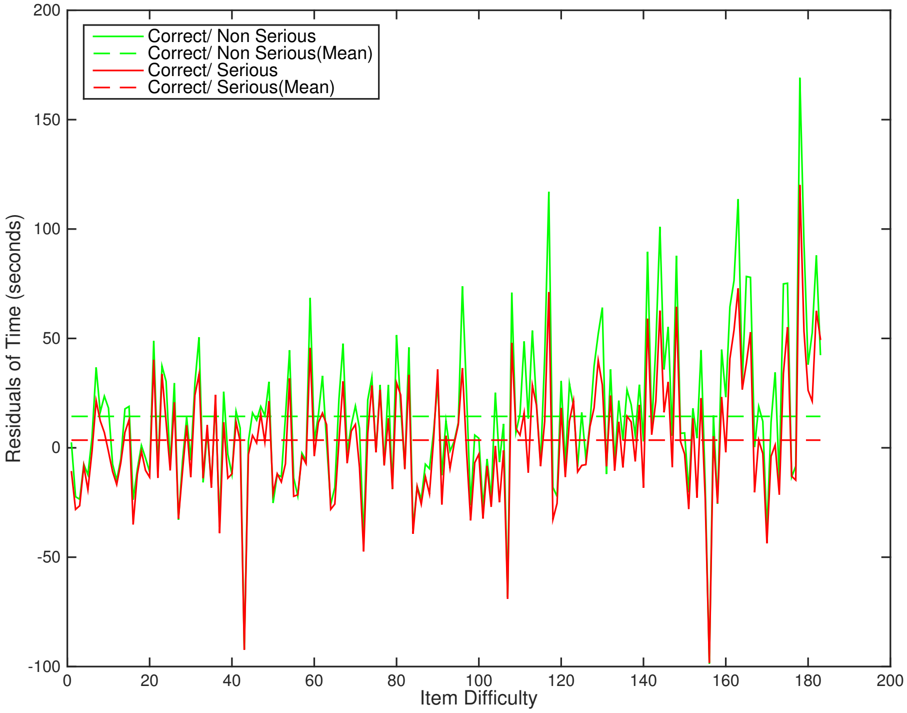


Figure 3: Time for Correct and Incorrect Answers for Serious and Non-serious Students

(a) Time for Correct Answers



(b) Time for Incorrect Answers

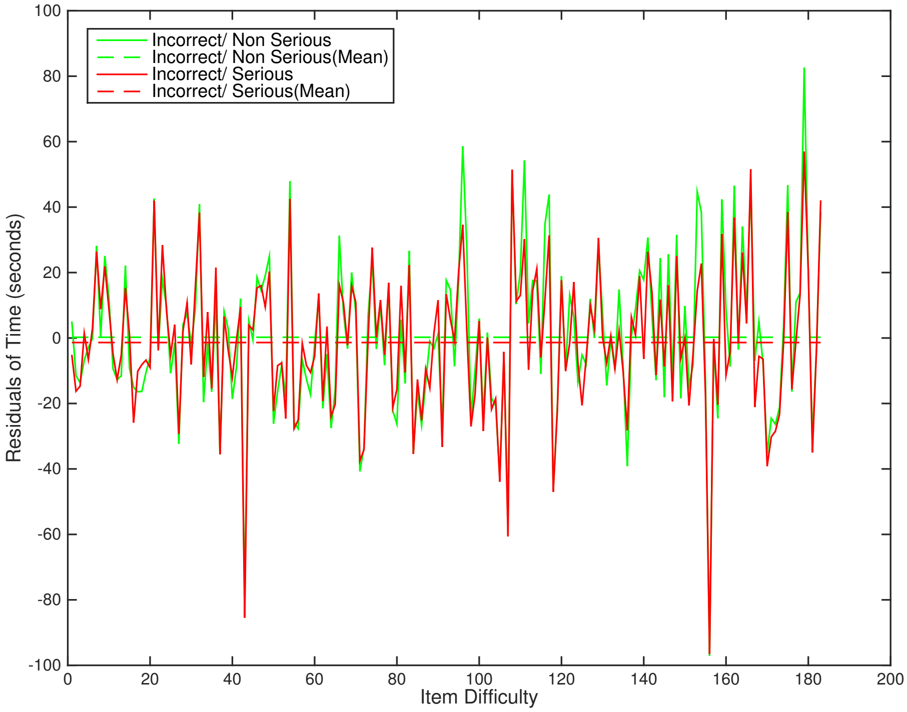
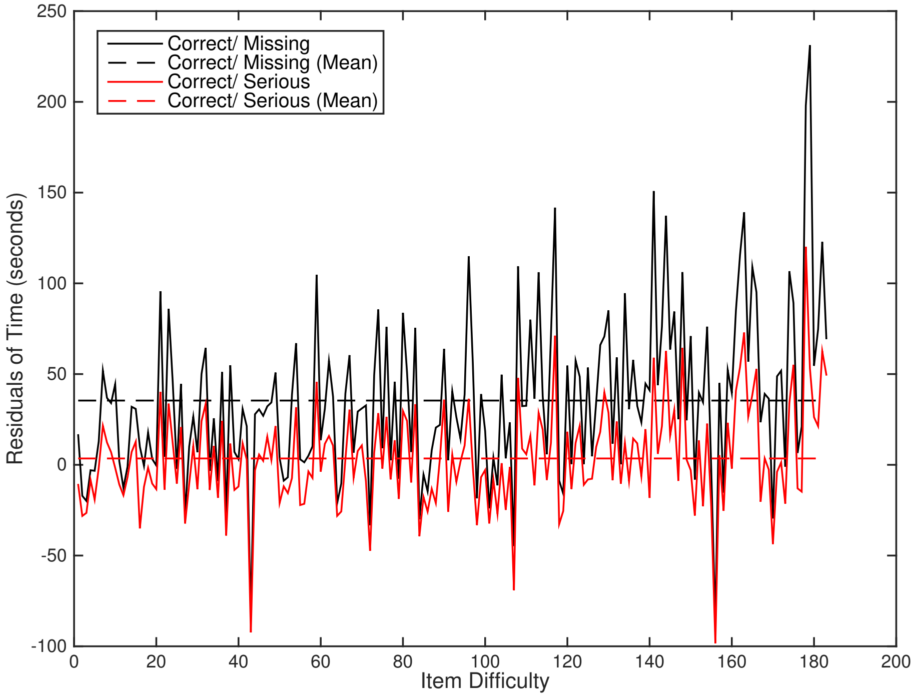


Figure 4: Time for Correct and Incorrect Answers for Serious and Missing-item Students

(a) Time for Correct Answers



(b) Time for Incorrect Answers

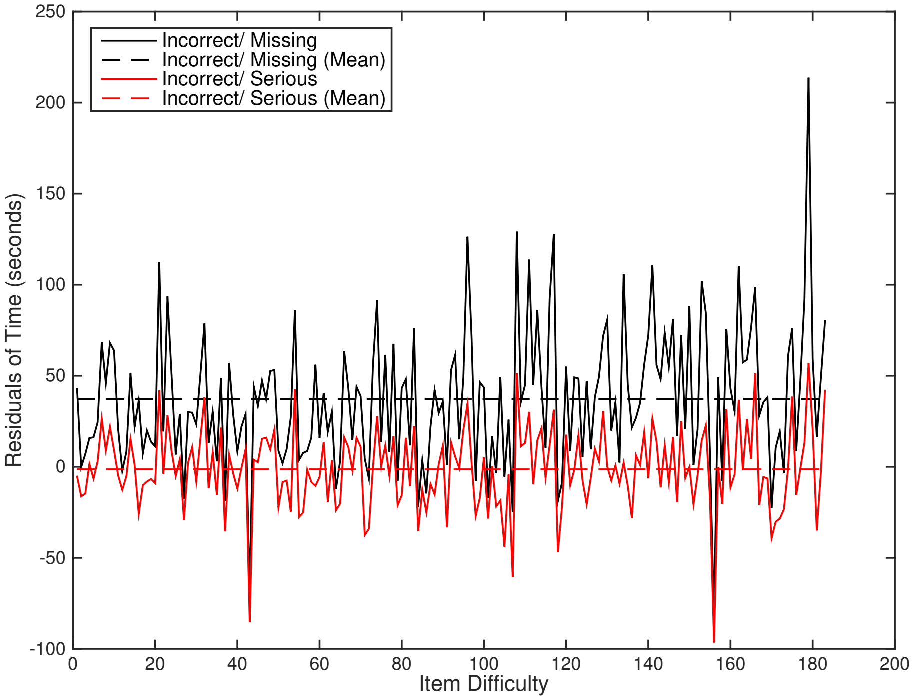
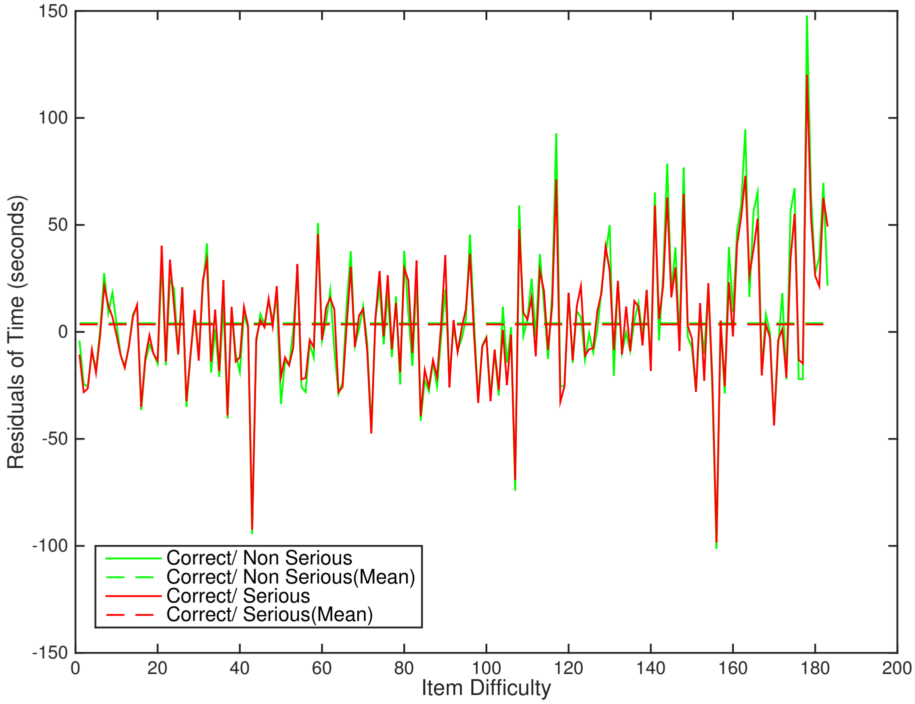


Figure 5: Time for Correct and Incorrect Answers for Serious and Non-serious Students After Removing Missing-item Students

(a) Time for Correct Answers



(b) Time for Incorrect Answers

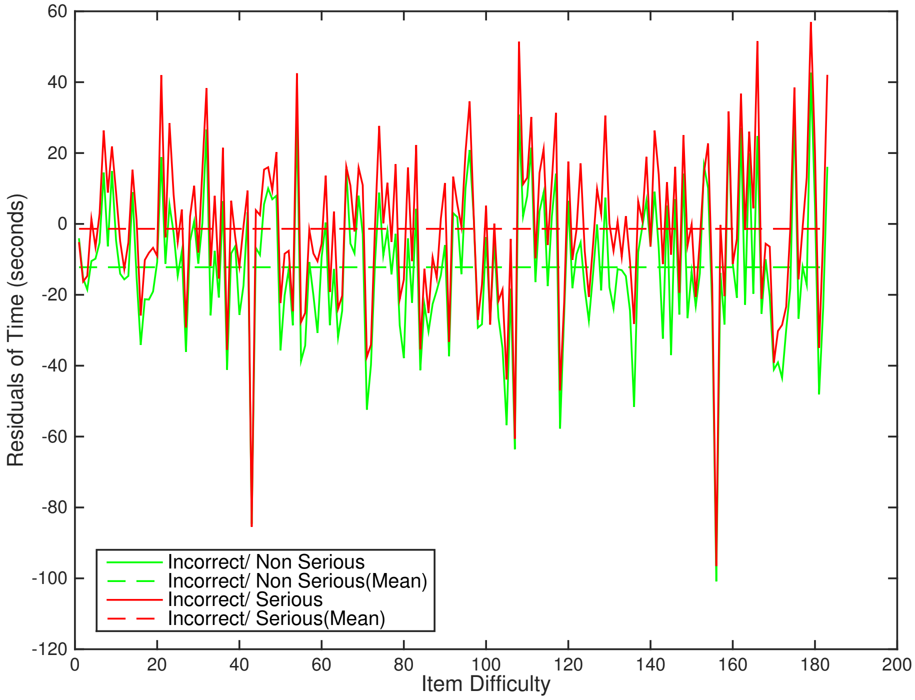


Table 1: Factors Related to Being Non-Serious

	Being non-serious (Criterion 1,2,4)			Criterion 3
	All countries	High stake countries	Low stake countries	All countries
Log (math score)	-0.3294***	-0.3383***	-0.3472***	0.0565***
	(0.0122)	(0.0163)	(0.0157)	(0.0092)
ESCS	0.0074***	0.0036	0.0195***	-0.0062***
	(0.0019)	(0.0027)	(0.0021)	(0.0016)
ESCS^2	0.0004	-0.0000	0.0027**	0.0041***
	(0.0009)	(0.0012)	(0.0012)	(0.0008)
Grade	-0.0087***	-0.0078***	0.0026	0.0103***
	(0.0021)	(0.0028)	(0.0032)	(0.0019)
Female	-0.0149***	-0.0198***	-0.0074**	0.0210***
	(0.0029)	(0.0039)	(0.0037)	(0.0026)
Anxiety	-0.0052**	-0.0037	-0.0090***	0.0111***
	(0.0023)	(0.0031)	(0.0029)	(0.0020)
Ambition	-0.0054**	-0.0042	-0.0008	-0.0090***
	(0.0025)	(0.0035)	(0.0033)	(0.0022)
Skipping class/Arriving late	0.0032***	0.0031**	0.0042***	-0.0002
	(0.0009)	(0.0013)	(0.0012)	(0.0008)
Log per capita GDP	1.4846***	0.9744***	1.8385***	-4.5828***
	(0.1159)	(0.1387)	(0.1777)	(0.1051)
(Log per capita GDP)^2	-0.0714***	-0.0473***	-0.0856***	0.2167***
	(0.0057)	(0.0068)	(0.0087)	(0.0051)
Out-of-school learning (hrs/week)	0.0005***	0.0005***	0.0001	-0.0005***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Time on classes	0.0002	-0.0001	0.0005***	-0.0014***
	(0.0002)	(0.0002)	(0.0002)	(0.0001)
Log (school average science score)	-0.0216	0.0768***	-0.2592***	-0.0399***
	(0.0167)	(0.0227)	(0.0209)	(0.0137)
Standardized test frequency	0.0022	0.0044*	-0.0080***	0.0016
	(0.0018)	(0.0024)	(0.0027)	(0.0017)
Teacher-developed tests frequency	0.0008	-0.0022	0.0034*	0.0075***
	(0.0013)	(0.0018)	(0.0018)	(0.0012)
Stakes of Standardized tests	0.0001	0.0000	0.0005	-0.0002
	(0.0002)	(0.0004)	(0.0003)	(0.0002)
Stakes of teacher-developed tests	-0.0012***	-0.0017***	0.0000	0.0008***
	(0.0003)	(0.0004)	(0.0005)	(0.0003)
Observations	283,674	128,668	155,006	283,674
R-squared	0.033	0.031	0.046	0.084

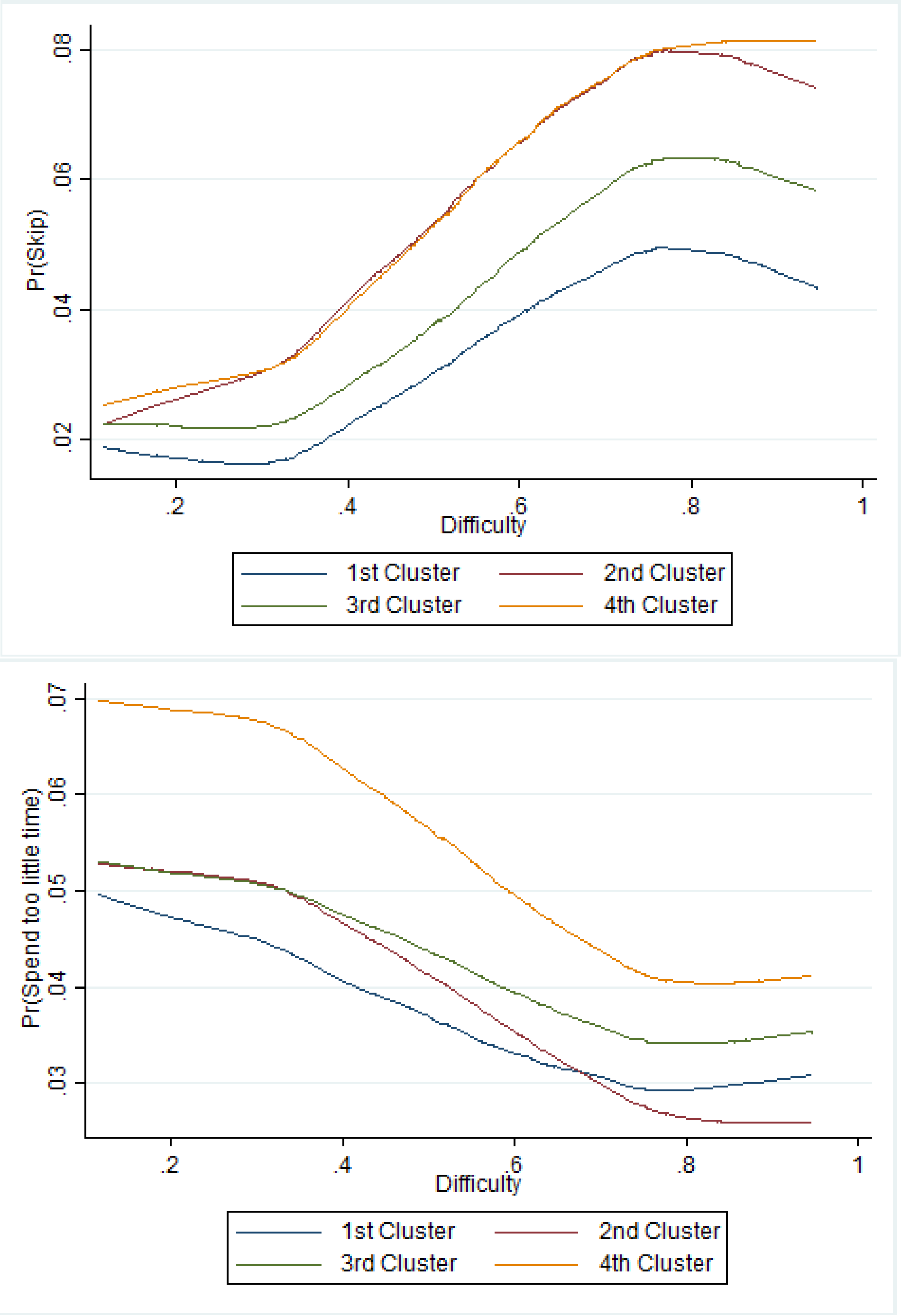
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Robust standard errors in parentheses

Note: In column 1-3 being non-serious does not include students meeting criteria 3. The latter group is analyzed separately in column 4. The number of observations is less than the number of students because students with missing variables are dropped.

Table 2: Factors Affecting Pr(Skip) and Pr(Spend too little time)

	Skip	Spend too little time
Sequence within cluster	-0.0021***	-0.0018***
	(0.0001)	(0.0001)
2nd Cluster	-0.0114***	-0.0145***
	(0.0012)	(0.0015)
3rd Cluster	-1.3824***	-0.0064***
	(0.0010)	(0.0012)
4th Cluster	-1.3898***	0.0059***
	(0.0012)	(0.0014)
2nd Cluster * Sequence	0.0028***	0.0027***
	(0.0001)	(0.0001)
3rd Cluster * Sequence	0.0001**	0.0015***
	(0.0001)	(0.0001)
4th Cluster * Sequence	0.0020***	0.0025***
	(0.0001)	(0.0001)
Difficulty	-0.0232***	-0.0069***
	(0.0018)	(0.0022)
Sequence * Difficulty	0.0035***	-0.0007***
	(0.0002)	(0.0002)
Complex multiple choice	0.0042***	-0.0055***
	(0.0005)	(0.0009)
Open response	0.0218***	-0.0121***
	(0.0009)	(0.0009)
Complex MC * Sequence	0.0007***	0.0003***
	(0.0000)	(0.0001)
Open response * Sequence	0.0037***	-0.0008***
	(0.0001)	(0.0001)
2nd Cluster * Difficulty	0.0149***	0.0014
	(0.0020)	(0.0021)
3rd Cluster * Difficulty	0.0073***	0.0059***
	(0.0019)	(0.0021)
4th Cluster * Difficulty	0.0125***	0.0032
	(0.0020)	(0.0022)
2nd Cluster * Complex MC	-0.0014**	-0.0044***
	(0.0006)	(0.0010)
3rd Cluster * Complex MC	0.0022***	-0.0033***
	(0.0006)	(0.0009)
4th Cluster * Complex MC	0.0022***	-0.0074***
	(0.0006)	(0.0010)
2nd Cluster * Open response	0.0166***	-0.0108***
	(0.0009)	(0.0010)
3rd Cluster * Open response	0.0091***	-0.0059***
	(0.0010)	(0.0010)
4th Cluster * Open response	0.0303***	-0.0198***
	(0.0012)	(0.0011)
Observations	13,268,385	13,053,985
Number of students	439,069	438,988
Individual FE	YES	YES
R-squared	0.0349	0.00641

Figure 6: Pr(skip) and Pr(spend too little time) w.r.t. cluster and difficulty



Note: In the figure, lowess-smoothed lines are presented.

Figure 7: Pr(skip) and Pr(spend too little time) w.r.t. sequence and the type of the question

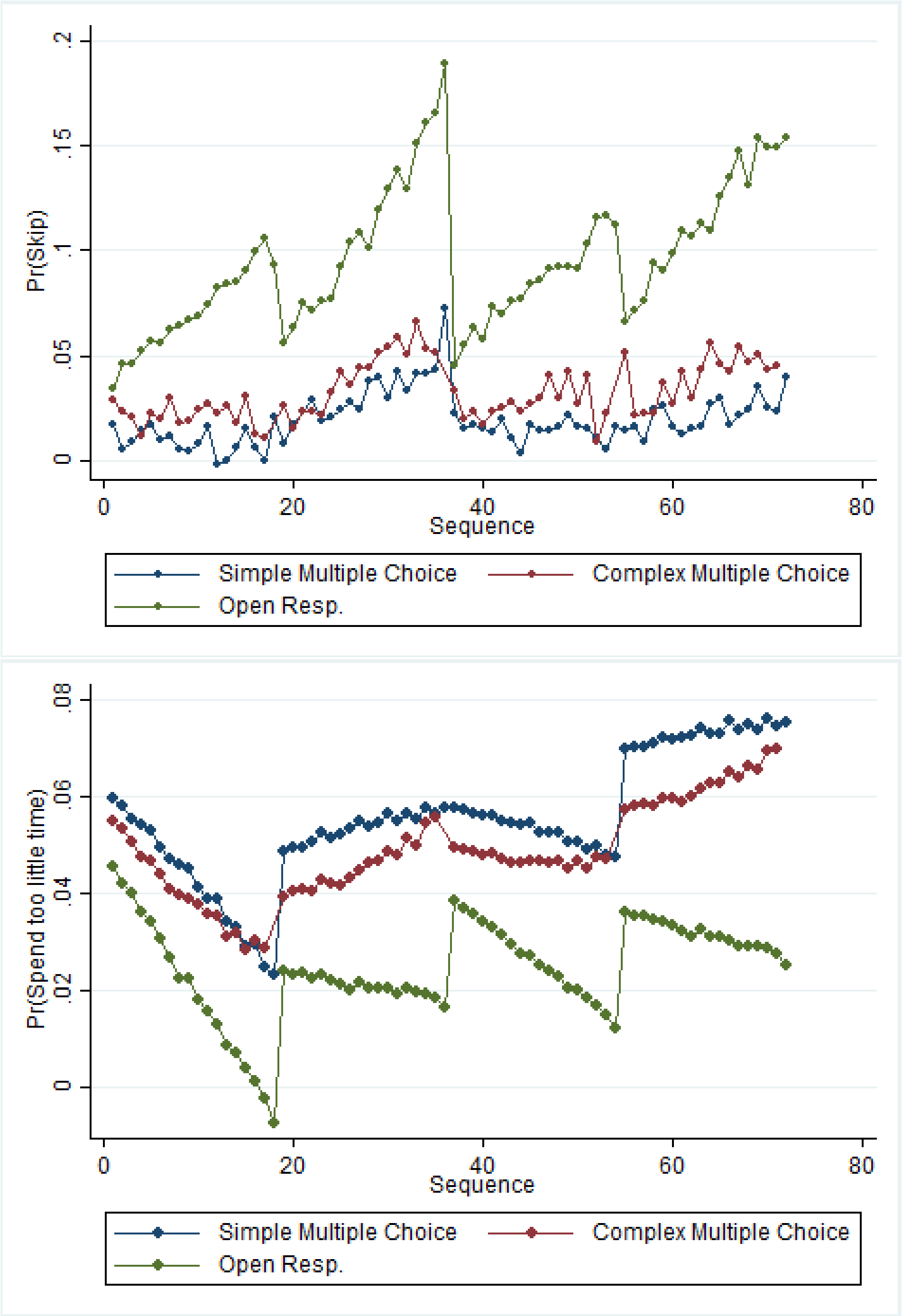


Table 3: Factors Affecting Pr(Skip) and Pr(Spend too little time) (Logit results)

	Skip	Spend too little time
Sequence	0.0663***	-0.0174***
	(0.0004)	(0.0003)
Difficulty	0.5753***	-0.1820***
	(0.0123)	(0.0106)
2nd Cluster	0.5960***	-0.0890***
	(0.0052)	(0.0048)
3rd Cluster	0.2685***	-0.0551***
	(0.0055)	(0.0048)
4th Cluster	0.6731***	0.1509***
	(0.0051)	(0.0046)
Complex Multiple Choice	0.3527***	-0.1733***
	(0.0064)	(0.0039)
Open Response	1.6417***	-0.7883***
	(0.0061)	(0.0051)
Log (math score)	-3.0174***	-1.1199***
	(0.0119)	(0.0121)
Log per capita GDP	6.2976***	5.7633***
	(0.1014)	(0.1033)
(Log per capita GDP)^2	-0.2947***	-0.2592***
	(0.0048)	(0.0049)
ESCS	0.0173***	0.0008
	(0.0022)	(0.0020)
ESCS^2	-0.0272***	-0.0026**
	(0.0011)	(0.0012)
Grade	-0.0396***	-0.0483***
	(0.0023)	(0.0024)
Female	-0.0161***	-0.2020***
	(0.0036)	(0.0034)
Anxiety	0.0026	-0.0267***
	(0.0027)	(0.0026)
Ambition	-0.1091***	0.0294***
	(0.0030)	(0.0029)
Skipping class/Arriving late	0.0489***	0.0337***
	(0.0009)	(0.0010)
Out-of-school learning (hrs/week)	-0.0007***	0.0028***
	(0.0001)	(0.0001)
Time on science classes	-0.0177***	0.0116***
	(0.0007)	(0.0006)
Standardized test frequency	0.0286***	0.0202***
	(0.0023)	(0.0023)
Teacher-developed tests frequency	-0.0117***	-0.0160***
	(0.0017)	(0.0017)
Stake of Standardized test	-0.0059***	0.0010***
	(0.0003)	(0.0003)
Stake of Teacher-developed tests	-0.0099***	-0.0006
	(0.0004)	(0.0004)
Log (school average science score)	-1.6053***	0.3700***
	(0.0176)	(0.0180)
Observations	9,349,185	9,349,185

Table 4: Countries' Rank before and after Imputation

Country	Original score	Original rank	All countries serious	All other countries serious	Only one country serious	Change in rank (Column 2-Column 5)
Singapore	0.74	1	1	1	1	0
Chinese Taipei	0.58	2	4	6	2	0
Estonia	0.56	3	3	7	2	1
Japan	0.56	4	2	7	2	2
Finland	0.54	5	5	8	2	3
Hong Kong	0.53	6	6	9	2	4
USA (Massachusetts)	0.50	7	7	9	3	4
Canada	0.48	8	8	9	5	3
Macao	0.45	9	9	10	6	3
Slovenia	0.40	10	11	14	10	0
B-S-J-G (China)	0.39	11	10	15	9	2
Netherlands	0.36	12	16	17	12	0
Korea	0.36	13	14	17	10	3
United Kingdom	0.34	14	15	18	11	3
Germany	0.33	15	12	21	10	5
Australia	0.28	16	17	26	12	4
New Zealand	0.28	17	13	26	10	7
Ireland	0.28	18	21	26	15	3
Poland	0.26	19	20	26	15	4
Denmark	0.26	20	19	28	15	5
Switzerland	0.25	21	18	29	14	7
USA (North Carolina)	0.24	22	23	32	16	6
Belgium	0.22	23	22	32	16	7
Austria	0.22	24	25	32	16	8
Norway	0.21	25	24	32	16	9
Czech Republic	0.18	26	27	32	20	6
United States	0.17	27	31	33	22	5
Spain (Regions)	0.17	28	28	33	20	8
France	0.15	29	29	33	21	8
Spain	0.15	30	32	33	22	8
Portugal	0.13	31	26	34	16	15
Latvia	0.12	32	33	35	27	5
Sweden	0.10	33	30	37	22	11
Italy	0.05	34	34	39	31	3
Lithuania	0.03	35	39	40	34	1
Luxembourg	0.03	36	36	40	33	3
Hungary	0.02	37	37	40	33	4
Croatia	0.01	38	38	40	34	4
Russian Federation	-0.03	39	35	41	32	7
Iceland	-0.05	40	40	42	35	5
Slovak Republic	-0.12	41	42	42	40	1
Israel	-0.12	42	41	42	39	3
Greece	-0.21	43	43	44	43	0
Bulgaria	-0.29	44	44	46	43	1
Chile	-0.31	45	45	46	44	1
United Arab Emirates	-0.33	46	46	46	44	2
Turkey	-0.41	47	48	48	47	0
Uruguay	-0.45	48	47	48	47	1
Qatar	-0.51	49	49	50	49	0
Thailand	-0.54	50	50	51	49	1
Costa Rica	-0.59	51	51	54	50	1
Colombia	-0.60	52	52	54	51	1
Montenegro	-0.64	53	53	54	51	2
Mexico	-0.65	54	54	54	51	3
Peru	-0.83	55	56	56	55	0
Brazil	-0.87	56	55	57	55	1
Tunisia	-0.92	57	57	57	56	1
Dominican Republic	-1.21	58	58	58	58	0

Table 5: Countries' Rank after Different Imputations

Country	Original number correct	Only impute missing items	Only impute little-time items	Impute both missing and little-time items
Singapore	1	1	1	1
Chinese Taipei	2	2	2	2
Estonia	3	2	2	2
Japan	4	2	3	2
Finland	5	3	5	2
Hong Kong	6	5	5	2
USA (Massachusetts)	7	6	7	3
Canada	8	7	8	5
Macao	9	9	9	6
Slovenia	10	10	10	10
B-S-J-G (China)	11	10	10	9
Netherlands	12	12	12	12
Korea	13	12	12	10
United Kingdom	14	12	14	11
Germany	15	11	15	10
Australia	16	14	16	12
New Zealand	17	12	16	10
Ireland	18	16	16	15
Poland	19	16	17	15
Denmark	20	16	19	15
Switzerland	21	16	19	14
USA (North Carolina)	22	16	22	16
Belgium	23	16	23	16
Austria	24	19	23	16
Norway	25	17	25	16
Czech Republic	26	23	26	20
United States	27	25	27	22
Spain (Regions)	28	23	27	20
France	29	23	29	21
Spain	30	25	29	22
Portugal	31	19	31	16
Latvia	32	30	31	27
Sweden	33	24	33	22
Italy	34	33	34	31
Lithuania	35	34	35	34
Luxembourg	36	34	35	33
Hungary	37	34	35	33
Croatia	38	34	38	34
Russian Federation	39	33	39	32
Iceland	40	38	40	35
Slovak Republic	41	40	41	40
Israel	42	40	41	39
Greece	43	43	43	43
Bulgaria	44	44	44	43
Chile	45	44	45	44
United Arab Emirates	46	45	46	44
Turkey	47	47	47	47
Uruguay	48	47	48	47
Qatar	49	49	49	49
Thailand	50	49	50	49
Costa Rica	51	50	51	50
Colombia	52	51	52	51
Montenegro	53	51	53	51
Mexico	54	51	54	51
Peru	55	55	55	55
Brazil	56	55	56	55
Tunisia	57	56	57	56
Dominican Republic	58	58	58	58

Figure 8: y_{ns} Versus its Components for Non-Serious Students

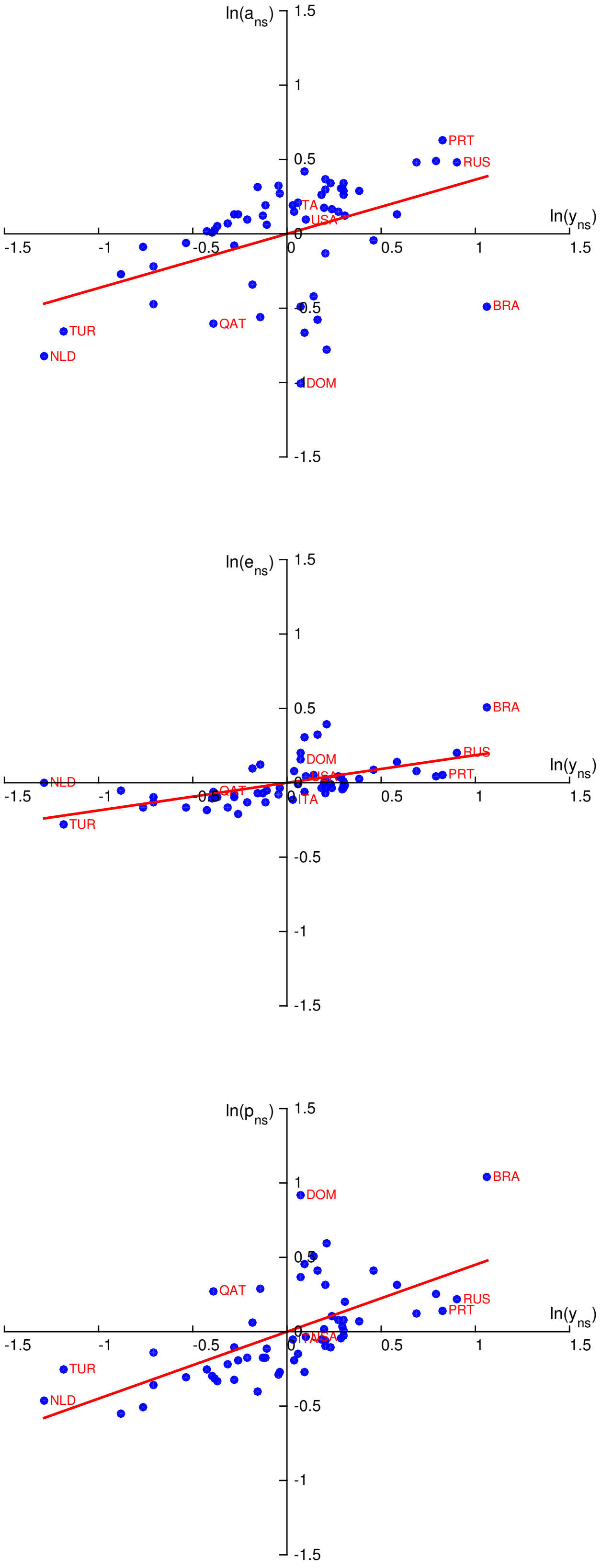


Table 6: Contribution of Factors to y_{ns} and y_s

		Dependent Variable: De-meaned Y	
		Non-Serious Students	Partial Serious Students
Coefficients for	De-meaned A	0.364	0.360
		(0.096)	(0.165)
	De-meaned E	0.186	0.371
		(0.033)	(0.097)
	De-meaned P	0.450	0.269
		(0.070)	(0.085)

Table 7: Decomposed Factors for Non-Serious Students

Country	$IFC(\%)$	$Y_{ns}(\%)$	A_{ns}	E_{ns}	P_{ns}
Brazil	3.56	3.39	0.157	0.324	0.670
Russian Federation	3.26	2.90	0.412	0.239	0.294
Portugal	3.02	2.68	0.476	0.206	0.273
Sweden	3.01	2.58	0.415	0.204	0.306
New Zealand	2.60	2.32	0.411	0.211	0.268
Israel	2.36	2.10	0.291	0.224	0.322
Uruguay	2.18	1.86	0.244	0.213	0.357
Belgium	2.01	1.71	0.340	0.200	0.252
France	1.92	1.58	0.329	0.188	0.255
Bulgaria	1.88	1.60	0.289	0.191	0.290
Australia	1.85	1.56	0.345	0.201	0.225
Peru	1.85	1.45	0.117	0.290	0.429
Iceland	1.84	1.59	0.339	0.197	0.238
Switzerland	1.84	1.57	0.345	0.187	0.244
Norway	1.84	1.58	0.358	0.192	0.230
Spain	1.82	1.54	0.296	0.204	0.256
Luxembourg	1.78	1.49	0.301	0.188	0.262
Macao	1.74	1.44	0.343	0.189	0.223
Tunisia	1.73	1.28	0.130	0.264	0.373
Spain (Region)	1.69	1.43	0.304	0.195	0.241
Germany	1.69	1.44	0.368	0.182	0.215
Denmark	1.69	1.48	0.357	0.194	0.213
Chile	1.66	1.44	0.224	0.200	0.322
Japan	1.65	1.29	0.388	0.184	0.181
Mexico	1.62	1.38	0.143	0.269	0.357
Slovak Repubic	1.61	1.41	0.331	0.189	0.225
United State	1.58	1.29	0.279	0.203	0.228
USA (North Carolina)	1.55	1.24	0.315	0.193	0.204
Montenegro	1.53	1.35	0.168	0.206	0.392
USA (Massachusetts)	1.51	1.21	0.296	0.211	0.195
Italy	1.46	1.20	0.308	0.174	0.225
Costa Rica	1.45	1.26	0.156	0.238	0.340
Dominican Republic	1.44	1.26	0.093	0.228	0.592
Canada	1.36	1.13	0.332	0.188	0.180
Estonia	1.35	1.12	0.353	0.180	0.177
Czech Republ	1.29	1.04	0.309	0.171	0.198
B-S-J-G (China)	1.26	1.03	0.288	0.181	0.198
Hungary	1.24	1.05	0.270	0.186	0.210
Finland	1.20	1.00	0.350	0.182	0.158
Colombia	1.20	1.02	0.146	0.221	0.316
Hong Kong	1.17	0.88	0.289	0.179	0.170
Thailand	1.17	0.97	0.181	0.214	0.251
Greece	1.15	0.88	0.235	0.176	0.214
Poland	1.13	0.95	0.281	0.171	0.198
Croatia	1.12	0.90	0.290	0.159	0.196
Ireland	1.05	0.85	0.273	0.165	0.190
Singapore	1.03	0.81	0.269	0.178	0.170
Austria	1.00	0.77	0.258	0.163	0.183
United Kingdom	0.96	0.79	0.257	0.175	0.175
Latvia	0.95	0.80	0.260	0.178	0.173
Qatar	0.91	0.79	0.139	0.184	0.311
Slovenia	0.83	0.69	0.240	0.165	0.174
Lithuania	0.75	0.58	0.204	0.170	0.166
Chinese Taipei	0.72	0.54	0.233	0.165	0.142
United Arab	0.72	0.58	0.159	0.177	0.205
Korea	0.60	0.48	0.194	0.184	0.136
Turkey	0.50	0.36	0.132	0.147	0.184
Netherlands	0.39	0.32	0.111	0.195	0.149

A Appendix

This appendix delves into more detail on a number of peripheral facts and issues. First, we discuss in more detail the behavior patterns of serious and non-serious students in terms of time spent and accuracy of response as a function of question position. Second, we look at the factors at the individual level that drive skipping behavior versus spending too little time separately using a linear probability model. We do so as the patterns seem very different. Third, we discuss the exact variables we use in the imputation procedure we rely on in our counterfactuals and fourth we present the summary statistics for the variables used in the paper. Finally, we explain some details behind the decomposition for partially serious students and present the results for them.

A.1 Time Spent, Accuracy and Position

Table A.1 shows time per science cluster across positions for serious and non-serious students. Note that time spent on the cluster falls with the position of the cluster and then jumps back up after the break at the end of cluster 2 and this is more so for non-serious students. As expected, serious students tend to spend more time than non-serious ones on each cluster. There is substantial heterogeneity between non-serious students according to the criterion used. Students with no-response or too-little-time items, not surprisingly, spend less time per cluster than serious students regardless of cluster position. However, the opposite holds for those with non-reached or missing items but only for the first and third clusters. For the second and fourth clusters their time spent is 30-40% less than that of serious students. It is also worth noting that for these students time is still not a constraint: on average they have more than 15 minutes left. This suggests that "fatigue" sets in faster for non-serious students.

Table A.1: Time Per Science Cluster (Minutes)

	Position 1	Position 2	Position 3	Position 4
Serious Students	22.25	17.93	20.20	17.55
Non-Serious Students (Union of 4 criteria)	27.65	12.10	19.70	11.82
Criterion 1 only (Nonreached items)	28.58	12.13	19.34	10.93
Criterion 2 only (No-response items)	20.75	11.20	15.64	10.71
Criterion 3 only (Missing items)	33.46	10.66	31.88	12.01
Criterion 4 only (Little-time items)	18.94	13.32	14.87	11.47

The upper part of Table A.2 shows proportion correct for all items (not just answered ones) across positions. Serious students have higher proportion correct than each category of non-serious students. Accuracy falls in the second cluster compared to the first one, and this is more so for non-serious students, reminiscent of the patterns for time spent. However, non-serious students will have a lower proportion correct on all items by definition as they skip many items. If we want to know what their accuracy is we should divide by the number of answered questions as done in the lower part of Table A.2. The numbers show that even with this correction non-serious students have lower accuracy than serious ones. In addition, the degree to which accuracy falls across clusters is now similar (around 2%) for both serious and non-serious students. This is consistent with non-serious students' performance experiencing a substantial drop in the second cluster primarily because they skip more items there.

Table A.2: Proportion Correct in Science Clusters

	Proportion correct for all items (%)			
	Position 1	Position 2	Position 3	Position 4
Serious Students	49.20	47.05	49.07	46.07
Non-Serious Students (Union of 4 criteria)	39.46	24.56	34.16	24.15
Criterion 1 only (Nonreached items)	33.81	19.74	27.46	17.85
Criterion 2 only (No-response items)	23.21	18.26	22.24	18.04
Criterion 3 only (Missing items)	43.17	18.23	41.96	18.27
Criterion 4 only (Little-time items)	42.83	36.98	36.46	31.49
	Proportion correct for answered items (%)			
	Position 1	Position 2	Position 3	Position 4
Serious Students	50.44	49.18	50.43	48.04
Non-Serious Students (Union of 4 criteria)	43.30	39.94	38.67	34.01
Criterion 1 only (Nonreached items)	40.17	37.19	36.41	31.83
Criterion 2 only (No-response items)	29.20	27.05	28.29	25.52
Criterion 3 only (Missing items)	46.59	44.94	45.52	41.87
Criterion 4 only (Little-time items)	44.91	41.53	39.22	35.05

A.2 Drivers of Skipping and Spending Too Little Time

Here we present the results of a linear probability model that looks at how individual characteristics affect skipping and spending too little time. Table A.3 suggests that better students (higher math score and grades) are less likely to both skip and spend too little time. Students with high socioeconomic status are less likely to spend too little time. Gender matters: women are less likely to spend too little time. Being anxious is positively associated with skipping but negatively with spending too little time, but being ambitious has the opposite pattern. Being undisciplined, i.e., having a pattern of skipping class or arriving late, is positively associated with spending too little time. Students from better schools, as reflected in the log of the school science score, are also less likely to skip but more likely to spend too little time.

Is there evidence of "fatigue"? Spending more time on studies both in and out of class, having more standardized tests with higher stakes does seem to correlate positively with spending too little time on the test. However, teacher developed tests have the opposite sign: both the stakes and frequency of these correlate negatively with spending too little time.

A.3 Variables Used in Imputation

PISA data has a rich array of information from the student and school questionnaires in the survey. In the imputation we use variables constructed from these surveys by PISA. We choose the variables that seem relevant. A list of the variables used is contained in Table A.4. Binary variables are clearly identified. All others are continuous indices. Details of these are available in the PISA technical report, (OECD (2015)), Chapter 16. The imputation also uses the individual's scores for all other items and other students' scores for all items as in the standard MICE imputations.

A.4 Descriptive Statistics

Table A.5 gives the descriptive statistics for the key variables used in the paper. Scores in the component parts of the exam (reading, math and science) are scaled so that 500 is the mean and the standard deviation is 100 for all OECD countries together. Clearly, OECD countries do better than average as the mean math and science scores overall are 464 and 474 respectively. Students are in the 10th grade roughly, and half are female. The variable "anxiety" is an index we constructed by taking questions that asked about this subject (where the ranking was from a "1" to a "4" in terms of strength of the viewpoint where 1 strongly disagree and 4 is strongly agree) and taking a simple average of the response. The median is 2.8 suggesting a fair degree of anxiety on the part of students. Similarly for "ambition" where the median response is 3.2[26]. The variable skipping class/arriving late uses the response for the three questions in ST062 about skipping, its intensity and arriving late and adds them up. A 1 is never in the last two weeks, a 2 is 1 or 2 times and a 3 is 3 or 4 times, and a 4 is 5 or more times. On average, such behavior exists but is not endemic.

The median time spent learning out of school is 16 hours per week, while time spent learning in school is 27 hours per week. Students spend more than 40 hours a week on school related work. The standard deviations are roughly 15 and 11 suggesting that a fair number of students are spending well over 60 to 70 hours a week on such work. Standardized test frequency and teacher developed test frequency is the response to question SC034. A response of 1 means there were no such tests and a response of 5 means the tests were given more than monthly. The median value is 2 or the frequency was 1-2 times a year. The variable "Stakes of standardized (teacher developed) tests comes from the answers to SC035. The question is composed of 11 yes/no sub-questions (where a yes is a 1 and a 0 is a no) regarding the purpose of these tests. We label each purpose as low, medium or high stakes for the students giving them a weight of 1, 2 and 3 respectively. Of the 11 sub-questions, 5 are low, 3 are medium and 3 are high stakes. We then add these weighted responses up to get our index. As the maximum value the index could have taken is 20, the median of 10 and 13 suggest the stakes are high, especially of teacher developed tests.

A.5 Decomposition for Partially Serious Students

We call fully serious students those who neither skip items nor spend too little time on any item. These fully serious students, together with what we call partially-serious students, make up what we have termed serious students. For fully serious students, the number correct will be the same before and after imputation by definition. The increase in fraction correct for serious students (Y_s) therefore only comes from imputations for partially serious students who did skip a few items or spent too little time on a small enough number of items so that they were not classified as non-serious. There are PS partially serious students. Next

[26] We used the 5 questions in ST118 for the anxiety variable and the 5 questions in ST119 for the ambition variable.

we will decompose Y_s into its component parts.

$$\begin{aligned}
Y_s &= \frac{\sum\limits_{i=NS+1}^{NS+S} I_i}{\sum\limits_{i=1}^{S+NS} T_i} \\
&= \frac{\sum\limits_{i=NS+1}^{NS+PS} (I_i)}{\sum\limits_{i=NS+1}^{NS+PS} NI_i} \frac{\sum\limits_{i=NS+1}^{NS+PS} NI_i}{\sum\limits_{i=NS+1}^{NS+PS} T_i} \frac{\sum\limits_{i=NS+1}^{NS+PS} T_i}{\sum\limits_{i=1}^{NS+S} T_i} \\
&= A_{ps}E_{ps}P_{ps}
\end{aligned}$$

A_{ps} is the increase in the fraction correct for non-serious items among partially serious students. E_{ps} is the fraction of non-serious items among all items for partially serious students, which measures the degree of non-seriousness. P_{ps} approximately measures the proportion of partially serious students in a country as partially serious students on average have the same number of total items as other students. The values of Y_{ps}, A_{ps}, E_{ps} and P_{ps} for each country are provided in Table A.6.

Similar to the decomposition for non-serious students, we divide both sides by the geometric mean and get

$$y_{ps} = \frac{Y_{ps}}{\bar{Y}_{ps}} = \left(\frac{A_{ps}}{\bar{A}_{ps}}\right)\left(\frac{E_{ps}}{\bar{E}_{ps}}\right)\left(\frac{P_{ps}}{\bar{P}_{ps}}\right) = a_{ps}e_{ps}p_{ps} \tag{7}$$

Take the logarithm on both sides of (7) gives:

$$\ln(y_{ps}) = \ln a_{ps} + \ln e_{ps} + \ln p_{ps} \tag{8}$$

Next we run the regression of $\ln a_{ps}$, $\ln e_{ps}$, $\ln p_{ps}$ separately on $\ln y_{ps}$, that is,

$$\begin{aligned}
\ln a_{ps} &= \alpha_2 \ln y_{ps} + \epsilon_a \\
\ln e_{ps} &= \beta_2 \ln y_{ps} + \epsilon_d \\
\ln p_{ps} &= \gamma_2 \ln y_{ps} + \epsilon_{p.}
\end{aligned}$$

Let the OLS estimates be denoted by $\hat{\alpha_2}, \hat{\beta_2}, \hat{\gamma_2}$. Similarly we can show that $\hat{\alpha_2} + \hat{\beta_2} + \hat{\gamma_2} = 1$ and the coefficients $\hat{\alpha_2}$, $\hat{\beta_2}$, $\hat{\gamma_2}$ measure the contribution of partially serious students' ability, extent of non-seriousness and proportion to a country's increase in fraction correct. Figure A.1 plots the scatter plot and regression lines above for partially serious students.

Table A.3: Factors affecting Pr(Skip) and Pr(Spend too little time) (Individual Characteristics)

	Skip	Spend too little time
Log (math score)	-0.0729***	-0.0448***
	(0.0218)	(0.0028)
Log per capita GDP	0.5978**	0.5047***
	(0.2338)	(0.0228)
(Log per capita GDP)^2	-0.0296**	-0.0241***
	(0.0115)	(0.0011)
ESCS	0.0015	-0.0013***
	(0.0039)	(0.0004)
ESCS^2	0.0008	-0.0003*
	(0.0018)	(0.0002)
Grade	-0.0035	-0.0033***
	(0.0039)	(0.0004)
Female	0.0055	-0.0073***
	(0.0060)	(0.0006)
Anxiety	0.0166***	-0.0016***
	(0.0047)	(0.0005)
Ambition	-0.0124**	0.0029***
	(0.0052)	(0.0005)
Skipping class/Arriving late	0.0015	0.0009***
	(0.0018)	(0.0002)
Out-of-school learning (hrs/week)	-0.0000	0.0001***
	(0.0001)	(0.0000)
Time on science classes	-0.0001	0.0006***
	(0.0011)	(0.0001)
Standardized test frequency	0.0027	0.0012***
	(0.0039)	(0.0004)
Teacher-developed tests frequency	0.0025	-0.0011***
	(0.0026)	(0.0003)
Stake of Standardized test	-0.0004	-0.0000
	(0.0005)	(0.0001)
Stake of Teacher-developed tests	-0.0006	-0.0003***
	(0.0006)	(0.0001)
Log (school average science score)	-0.1036***	0.0262***
	(0.0324)	(0.0035)
Observations	299,577	299,532
R-Squared	0.00236	0.0247

Table A.4: Variables Used in Imputation

Variable	Description
FEMALE	Female=1, male=0
GRADE	Grade compared to modal grade of 15-year-old students in country
ESCS	Index of economic, social and cultural status
BELONG	Sense of belonging to school
unfairteacher	Teacher fairness
TWINS	Total learning time (minutes per week)
OUTHOURS	Out-of-school study per week
COOPERATE	Enjoy cooperation
JOYSCIE	Enjoyment of science
INTBRSCI	Interest in broad science topics
DISCLISCI	Disciplinary climate in science classes
TEACHSUP	Teacher support in science classes
SCIEACT	Science activities
ANXTEST	Test anxiety
MOTIVAT	Achieving motivation
EMOSUPS	Parents emotional support
DURECEC	Duration in early childhood education and care
REPEAT	Ever repeated a grade=1, otherwise 0
TIMESCIE	Total time spent on science clusters in PISA exam
NONSERIOUS	Being non-serious in PISA exam=1, otherwise 0
CLISIZE	Class size
EDUSHORT	Shortage of educational material
STAFFSHORT	Shortage of educational stuff
PROATCE	Proportion of all teachers fully certified
CREACTIV	Creative extra-curricular activities
PROSTMAS	Proportion of science teachers with ISCED level 5A and a major in science
STRATIO	Student teacher ratio
PUBLIC	Public school=1, otherwise 0
sch_scie	School average PISA science score
log_pdgp	Log of per capita GDP in the country

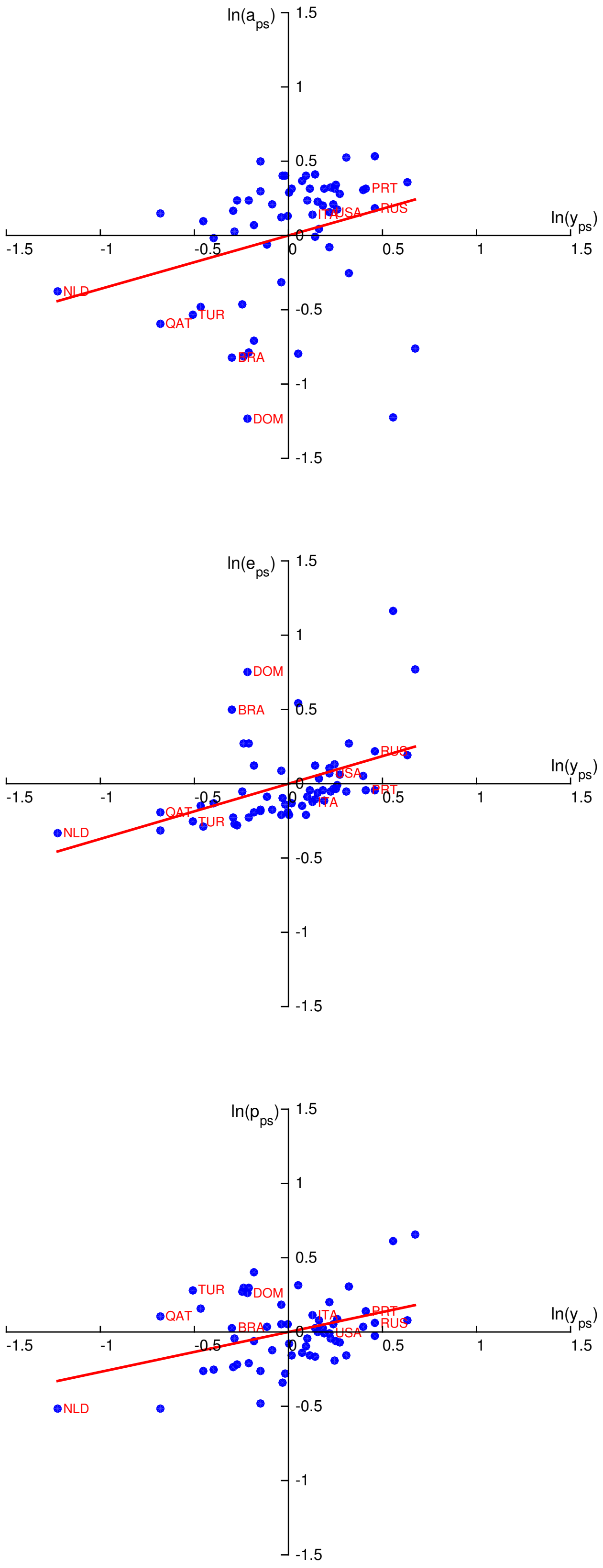
Table A.5: Summary Statistics

	mean	sd	median	min	max
Math score	464.46	97.90	463.19	108.15	826.34
ESCS	-0.42	1.15	-0.36	-7.26	4.18
Grade	9.77	0.78	10	7	13
Female	0.50	0.50	0	0	1
Anxiety	2.71	0.67	2.8	1	4
Ambition	3.13	0.60	3.2	1	4
Skipping class/Arriving late	4.32	1.68	4	3	12
Out-of-school learning(hours per week)	19.58	14.69	16	0	70
Time on classes (hours per week)	28.25	11.11	27	0	70
Standardized test frequency	2.07	0.85	2	1	5
Teacher-developed tests frequency	3.96	1.05	4	1	5
Stakes of standardized tests	9.11	7.01	10	0	20
Stakes of teacher-developed tests	12.12	5.78	13	0	20
School average science score	473.62	71.86	478.58	214.86	717.17

Table A.6: Decomposed Factors for Partially Serious Students

Country	$Yps(\%)$	Aps	Eps	Pps
Tunisia	0.449	0.214	0.096	0.219
Sweden	0.430	0.651	0.054	0.123
Peru	0.399	0.135	0.141	0.210
Japan	0.363	0.778	0.042	0.110
Russian Federation	0.363	0.548	0.055	0.120
Portugal	0.346	0.626	0.042	0.131
France	0.342	0.620	0.047	0.118
Uruguay	0.316	0.354	0.058	0.154
USA (North Carolina)	0.313	0.773	0.042	0.097
Macao	0.300	0.604	0.047	0.106
Luxembourg	0.296	0.541	0.044	0.124
Belgium	0.294	0.643	0.043	0.107
USA (Massachusetts)	0.293	0.624	0.050	0.094
Australia	0.290	0.565	0.043	0.120
Hong Kong	0.287	0.633	0.042	0.108
Bulgaria	0.286	0.421	0.049	0.139
United States	0.285	0.532	0.047	0.113
New Zealand	0.277	0.628	0.039	0.112
Spain	0.275	0.556	0.042	0.117
Greece	0.269	0.476	0.046	0.123
Spain (Region)	0.268	0.572	0.041	0.113
Switzerland	0.265	0.691	0.040	0.096
Israel	0.264	0.451	0.050	0.117
Italy	0.260	0.525	0.039	0.127
Norway	0.257	0.624	0.042	0.097
Iceland	0.254	0.575	0.040	0.109
Germany	0.251	0.680	0.036	0.103
Czech Republic	0.247	0.660	0.038	0.098
Mexico	0.242	0.206	0.076	0.155
Canada	0.234	0.623	0.039	0.097
Austria	0.230	0.611	0.036	0.105
B-S-J-G (China)	0.228	0.521	0.037	0.120
Estonia	0.226	0.683	0.038	0.086
Singapore	0.223	0.685	0.040	0.081
Chile	0.220	0.334	0.048	0.137
Croatia	0.220	0.516	0.036	0.119
Denmark	0.210	0.564	0.037	0.100
Slovak Repubic	0.205	0.429	0.040	0.118
Ireland	0.197	0.614	0.037	0.088
Finland	0.197	0.754	0.037	0.070
Thailand	0.191	0.225	0.050	0.171
Hungary	0.191	0.487	0.037	0.107
Poland	0.186	0.577	0.035	0.092
Costa Rica	0.185	0.208	0.058	0.154
Dominican Republic	0.184	0.133	0.094	0.148
Colombia	0.180	0.203	0.058	0.153
Montenegro	0.179	0.286	0.042	0.150
Chinese Taipei	0.175	0.578	0.033	0.091
Lithuania	0.172	0.469	0.034	0.109
United Kingdom	0.170	0.537	0.035	0.090
Brazil	0.170	0.201	0.072	0.117
Latvia	0.154	0.450	0.039	0.088
Slovenia	0.145	0.501	0.033	0.088
United Arab	0.143	0.283	0.038	0.133
Turkey	0.137	0.267	0.034	0.151
Korea	0.116	0.530	0.032	0.068
Qatar	0.116	0.252	0.036	0.126
Netherlands	0.067	0.313	0.032	0.068

Figure A.1: y_{ps} Versus its Components for Partially Serious Students


ln(a_ps)
ln(y_ps)
PRT
RUS
ITA
USA
NLD
TUR
QAT
BRA
DOM
ln(e_ps)
ln(y_ps)
DOM
BRA
RUS
USA
PRT
ITA
QAT
TUR
NLD
ln(p_ps)
ln(y_ps)
TUR
DOM
QAT
BRA
ITA
PRT
RUS
USA
NLD