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, Revised April 2019

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 colleagues 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, Revised April 2019
JEL No. C53,I20,I21

ABSTRACT

PISA is seen as the gold standard for evaluating educational outcomes worldwide. Yet, being 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 in low-stakes exams by leveraging information in computer-based assessments in PISA 2015. When skipped items are given a zero score, we show that this bias is large: most countries increase their scores significantly had their students been serious. A country can rise up to 15 places in rankings if its students took the exam seriously.

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 could 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 their is evidence that students are not taking the exam seriously so that scores and rankings

[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).

could be biased, and if so, how we can adjust for these biases to obtain a reliable snapshot of student skills?

PISA assesses the extent to which 15-year-old students have acquired key competencies (knowledge, skills and attitudes) to solve problems in a real-life context. One might argue that the attitudes and dispositions exhibited in taking PISA are also relevant to everyday life and for this reason, the scores need no adjustment. However, there is little reason to think that a student who blows off PISA, perhaps because he has been overtested, will fail in life because he does not take life seriously.

It is worth noting that though PISA is a low stakes exam for students, there is much at stake for 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. Every three years when the new PISA results come out, they are cited authoritatively in countless newspapers and policy reports. In many countries they even start to deeply influence educational practices. In 2014, more than one hundred academics around the world wrote a letter to the director of PISA to express their deep concern about the impact of PISA results[2]. They wrote:

> “As a result of PISA, countries are overhauling their education systems in the hopes of improving their rankings. Lack of progress on PISA has led to declarations of crisis and "PISA shock" in many countries, followed by calls for resignations, and far-reaching reforms according to PISA precepts. ”

One example is India. 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.

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[3], 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

[2]See the article in The Guardian, May 6, 2014 entitled "OECD and Pisa tests are damaging education worldwide-academics", Retrieved from the following link: https://www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics

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

first two clusters. We restrict attention to the Science component of the test as in 2015, all students had to take two clusters in this area. The remaining two clusters are for the reading, math and CPS (collaborative problem solving) components. As a result, some students may take only reading, while others may only take math or CPS which would reduce the sample size[4]. In the PISA exam, there is no negative marking (wrong answers are not penalized); therefore there is no reason to skip a multiple-choice question: students should guess even if they have no idea of the correct answer[5].

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 or high socioeconomic status or 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 performance. We ac-

[4]In the appendix we provide some data that suggests that the results we see in the science section will likely be correlated to and magnified in the reading and math sections as non seriousness seems more prevalent in math and reading than in science.

[5]One might argue that students do not understand that it is better to guess than to skip. However, if this was the only reason for skipping, then skipping behavior should not be related to the position of the item, which clearly is as shown below. One might also argue that as this is a computer based test, students cannot go back to answer skipped item as they might in a paper test. If students do not realize this, they may skip inadvertently. Again, since they will quickly learn that they cannot go back even if they do not know this to begin with, skipping should be less prevalent in the second cluster than the first. Again, the opposite is true.

count 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[6]. PISA documents are clear that their scores and rankings come with confidence intervals, see Figure 1.2.14 in OECD (2015a). We follow what PISA does in generating plausible values for the imputed data, then calculate the 95% confidence intervals for the imputed scores, and finally use these intervals to calculate the rank intervals for countries using the computer based assessments.

We make a number of comparisons. First, we compare the fully imputed score (FIS) to the original score (OS) where skipped items are given zero points, as is the case in most tests. One could also compare the fully imputed score to the score when skipped items at the end of the exam are treated as if they were not administered, SENA for skipped at end not administered, which is the procedure followed by PISA. The fully imputed score is what would be obtained by a country if its students took the test seriously.

There is evidence that a few countries are trying to get their students to take the exam seriously with a view to gaming the system. For example, Abu Dhabi gave mock PISA exams to prepare students for the PISA exams in 2018. Each school was sent a student report as well as a school report comparing them to other schools and the international averages. For each student, the areas that teachers need to work on were highlighted[7]. Canadian school teachers are given a handbook on how to prepare students for the PISA exam.(See Prince Edward Island (2002)) In the handbook, teachers are urged to "encourage them (students) to take the assessment seriously and strive for excellence."

We show that a country can improve its ranking by up to 15 places by encouraging its own students to take the exam seriously when we compare the FIS rankings to rankings using the OS. Of the 58 countries, 24 have rank confidence intervals that do not overlap. Using the FIS versus the SENA score, a country can improve its ranking by less, up to 7 places. Only 1 of 58 countries have rank confidence intervals that do not overlap. If all countries become serious, then the rankings change by little. But this is to be expected: if everyone tries to game the system, their efforts cancel out.

What matters for countries is not only their rankings but also their students' performance. The difference in the scores (FIS versus OS) is significant at the 5% level in 50 of 58 countries, and at the 1% level in 46 of 58 countries. The difference in the FIS and SENA is significant for only 7 countries at the 5% level and 2 countries at the 1% level. This suggests that the PISA approach of treating skipped items at the end as not administered and ignoring them

[6]This procedure uses other responses of the same agent as well as those of agents like him to do the imputation.

[7]See the article in the National, Sept 25, 2017 entitled "Abu Dhabi pupils prepare for Pisa 2018", Retrieved from the following link: https://www.thenational.ae/uae/abu-dhabi-pupils-prepare-for-pisa-2018-1.661627

goes part of the way in correcting for the effect of non seriousness. In effect, imputing data for skipped items at the end versus treating them as not administered gives roughly the same answer for the fraction correct obtained as the only difference becomes the number of questions administered. Of course, just dropping skipped items at the end is not enough to fully account for non seriousness. This is why the FIS also imputes data for skipped items in the middle of the test as well as items on which too little time was spent. In addition, countries can take advantage of PISA's approach if they are aware it. For example, they can instruct their students to spend as much time as they need on earlier questions because even if they do not have time to reach the latter questions, those questions will be dropped in score calculation.

Countries vary substantially in terms of the change in score and ranking if their students took the exam seriously. The change 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 move up very little in their ranking because their non-serious students are of low ability. There are also countries with a medium fraction of non-serious students (such as Russia) whose students' performance improved by a large extent as their non-serious students have high 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, 68% of the variation comes from the proportion of non-serious students, while 26% comes from the extent of non-seriousness, with the remaining coming from their ability.

There is a literature that uses PISA data to study the role of institutional differences such as effects of instruction time, school autonomy and tracking on students' performance (Lavy (2015), Hanushek et al. (2013), Hanushek and W ößmann (2006)) or to analyze how students' performance differs according to their background characteristics (Lounkaew (2013)). If non-serious behavior is correlated with the variable used in these studies, then their findings may be biased. This is another reason to account for non-serious behavior. 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, responses to student survey questions and the consistency of these responses to tease out effort differences.[8] 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. Borghans and Schils (2012) document the same fall in test performance over question order as does Zamarro et al. (2016), but in addition they use two other datasets and show that this decline is related to personality traits, like agreeableness, and to motivational attitudes towards learning.

Butler and Adams (2007) use self reported expenditure of effort by students and argue that because it is fairly stable across countries, and is unlikely that systematic differences in the effort expended by students invalidates the comparison across countries in PISA results. Baumert and Demmrich (2001) conduct an experiment on German students to see if different ways of increasing the stakes involved affect performance. They offer monetary incentives tied to performance, feedback on performance, or the test mattering for school grades as well as the standard PISA setup as a control. They find no significant effects on performance or self reported effort from any of these treatment arms. Our work using much richer keystroke data suggests differently. The extent of non seriousness and its consequences vary considerably across countries and it would be a mistake to project results from one country

[8] One of their measures of effort is the extent to which performance falls when the question occurs later in the exam. 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.

to other 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.[9]

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. 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 performance by motivating their students, and that a country's rank falls a lot if other countries motivate their students while it does not.

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.[10] It analyzes the effects on scores and ranking if non-serious students behaved like serious ones for the 58 countries and areas that participated in the computer-based PISA exam in 2015. As a result we can do "partial equilibrium" analysis (one country is serious at a time), which is the most relevant since most countries do not intervene actively so as to raise PISA scores, or general equilibrium analysis (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 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 back-

[9]Our estimates below also show that China seems to be less affected by non serious behavior than the US.

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

ground 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 scores and 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.[11] 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.[12] 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 (2015b)). Each student gets different clusters based on a random number assigned to students.[13] 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 (2015b), 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 [14] and open response. Each type accounts for approximately one third of all questions.

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

[11]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.

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

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

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

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.[15] Table A.8 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 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.[16] This is why open response and multiple choice questions are treated differently below. We distinguish between *who* is not serious and when a *question* is not being taken seriously. For example, if a *student* leaves too many questions blank while having the time to complete them, the *student* is seen as being non-serious. Similarly, if a question is left blank while the student has time left, the *question* is not taken seriously. Skipping open response questions (i.e. not responding to the question after spending time on them) is not seen as an indication of non-seriousness. This is because the student may have absolutely no idea what the answer is and for such questions, guessing has a zero return. However, not attempting them (i.e., not spending any time on them) when there is time left is seen as an indicator of non seriousness. That this is the right thing to do is suggested

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

[16]Note that students who skipped open response questions in the middle of the exam, even if they spent very little time on them, were not seen as non serious. They could have equally well been labeled as non-serious. However, such open questions, which are both not answered and spent too little time on, only account for 0.7% of the total questions, so we are not worried this will affect our results.

by the fact that skipping multiple choice questions is more likely later in a cluster and in later clusters, see Figure 6. This would not be the case if the reason for skipping is complete ignorance of the answer or if they skipped because they were naive and did not realize that guessing was always better than skipping.

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 follow the slightly confusing PISA terminology and label the items which are at the end of the cluster and for which the student did spend some time on as "non reached". If the student did not reach the item at all, it is labeled 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 (more than 5 minutes) to attempt an additional question.[17] In each of the criteria below we set the cutoff so that no more than 10% of the students meet it.[18]

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. Note that "not reached" is PISA terminology and a bit misleading since some time is spent on these questions.

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*. In other words, they skipped at least two questions. As this was a computer based exam students could not go back once they move on to the next question. If students are aware of this, they should just guess on multiple choice questions they are unsure of. Skipping thus suggests not caring. 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*. In other words, there is time left and there are questions they did not get to. 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

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

[18]We also did a robustness check by setting the cutoff at a diffent level, and found similar patterns. (see Table A.5 and A.4)

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. This is for both multiple choice and open response questions.

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. 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. There may have been technical issues that prevented them from taking it and in any case, there is no way for their responses to be imputed as there is no information on which to do so. Then we remove outliers for each country in terms of total response time, following Chapter 9 in the technical report (see OECD (2015b), 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 the maximum between $\big[mean - \ 2.5 * median(\big\|(r_j - median)\big\|)\big]$ and 5 seconds. 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.[19] We make sure we avoid such mislabeling as follows.

[19] 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.

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. Table 1 gives the fraction of non reached, no response, missing and too little time items for each country in columns 1 to 4 for the science component. Note that countries differ in the way their students are non serious. Brazil and Peru, for example, have the highest share of missing items, 20% and 12% respectively. On the other hand, the Dominican Republic has the highest number of non reached items at 15%. Recall that PISA treats both non reached and missing items as not administered. In contrast, Montenegro has 10% no response items (skipped in the middle of the exam) which are counted as a zero. If there is no time constraint, differential treatment of questions in the middle of the test and at the end is hard to justify. If however, time was a constraint, so that spending more time on a question increased the chances of getting it correct, then countries could improve their scores by encouraging students to focus all their time on a small fraction of questions and again this differential treatment is questionable.

Also note that as the fraction of skipped items (no response, non reached and missing items) is larger than the fraction of too little time items, it is this margin that is going to do most of the heavy lifting in the imputation. One would have expected also, that for multiple choice questions, unless students have very high ability, the improvement from imputing too little time items would be small (they are already guessing) compared to the improvement from imputing skipped items where the lower bound is a gain of 0.25 obtained from just guessing. The fraction of no response items for the reading and math tests are a bit higher on average as shown in Table A.1 in the appendix. It is also highly correlated with the numbers in science. For example, the correlation between the fraction of no-response items for science and for reading is 0.98, showing that non seriousness is common across subjects of the test as might be expected.

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. 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.

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. Table A.2 shows time per science cluster across positions for serious and non-serious students. As it is clear from the table, students on average have more than 15 minutes left.

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". 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.[20] 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 in, for example difficulty, 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-

[20]Note that the standardization is done separately for each question, but these questions may appear in different orders. So, in each order, there are different questions.

cation is also apparent.[21] 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.2, 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 despite having plenty of time left.[22] Also note that as is evident in Table A.3, 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[23] 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[24] 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

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

[22]Note that students satisfying criterion 3 have on average 15 more minutes left.

[23]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.

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

roughly the same time on a question independent of whether they get it right or wrong, 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[25] that correlate with a student being non-serious are explored in Table 2. 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[26], 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. One might want to do these regressions country by country, but with 58 countries, this would be overkill as this is not the main object of this paper. Also note that we are not claiming any causal effects, merely pointing out some correlations in the data.

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.[27] 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 criterion 3 when they are non serious

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

[26] 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 between 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.

[27] Our results are robust to using fraction correct on items that are answered seriously as a measure of ability.

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.

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. The data reveals that 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 for all countries together, but the effect seems to be coming from high stakes countries. Time spent in school is positive but not significant for all countries together, but is significant for low stake countries.[28] Having more tests (standardized or teacher-developed) does seem to correlate positively with being non-serious overall, though the coefficients are not significant. This might be because the effects differ in high stake and low stake countries. 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).

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. In Table A.4 and Table A.5, we present correlates of each non-seriousness criterion for cutoff level of 10% (as defined in Section 3.1) and 6%, respectively. The results are consistent across different cutoff levels.

Though we do not emphasize them in the next section, similar patterns in terms of individual characteristics, are observed in Tables A.6 (when we run individual fixed effects

[28]See column 1 and the row for time on classes and out-of-school science learning.

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.

Our results here should be seen as preliminary as there is no causation implied, merely correlation. The patterns described above are suggestive and might be worth exploring in future work.

4.2 Which Questions are Not Taken Seriously?

We define a non serious question as those that were *not reached*, for which there was *no response*, were *missing*, or on which *too little time* was spent. *Non-reached and no-response* items were looked at by the student who then chose not to answer the item despite having time left. Had he taken the exam seriously, he would have answered to the best of his ability. The student did not even look at *missing* items. But he had time left. In general, students have ample time to do the exam. Not even bothering to even read the question is again an indication of non seriousness. One might argue that *no-response* items, i.e., those that were skipped in the middle of the exam, should be treated differently as this was a computer based exam and students could not go back. Assuming they knew this, their choosing to skip again indicates the question is not taken seriously. Questions on which *too little time* was spent (as explained in criterion 4 for defining non serious students) are those where the response time is below a threshold which is country specific and for which the proportion correct is lower than that for normal time items for the same person. This is to prevent us from mistakenly labeling a question as non serious when in fact the student knew the answer immediately and so spent little time on it.

We explore the effects of question characteristics on the probability of a question being skipped, i.e., being *not-reached* or *no-response*. 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. Figure 6 shows the predicted probability of skipping a question and the predicted probability of spending too little time on a question for each cluster as a function of the difficulty of the question.[29] In all clusters, as the difficulty of the question increases, the probability of skipping increases though there is a slight decrease as questions become very difficult(top panel). In the bottom panel, we see that the probability of spending too little time is roughly flat: first increasing, then decreasing and finally increasing again. Students seem to try to answer if the question is easy but as it gets difficult, they seem to give up. There are also differences between clusters.

[29]For each cluster, the predicted probability at each level of question difficulty in the figure takes the mean value of the predicted probability at that level of difficulty.

Consistent with the "fatigue" hypothesis, questions are more likely to be taken non seriously in the second and fourth clusters.

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

Non-serious behavior in terms of spending too little time weakly falls with the order within a cluster for all question types. However, there is a large jump up at the beginning of the second and fourth clusters. 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. While skipping is more likely for open response questions, spending too little time is less likely for such questions relative to 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.6. The results are in line with those of Table 2.

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, answer with too little time or answer with normal time. 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. Open response and complex multiple choice questions move students towards skipping and away from spending too little time. However, as the difficulty of the questions increase, the students become more likely to skip and spend too little time. The coefficients on individual characteristics are roughly in line with those in Table 2. The math score of the student is negatively correlated with the probability of skipping and the probability of spending too little time. Female students are less likely to skip or spend too little time. Ambitious students are less likely to skip. 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 Scores and 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 scores and 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.[30] All of these are treated as missing data. Non-reached and no-response items were looked at by the student who then chose not to answer the item despite having time left. Had he taken the exam seriously, he would have answered to the best of his ability which is exactly what the imputation does. Note that in Section 3.1 we did not include open response items in defining non reached and no response items. We did include both multiple choice and open response items in defining missing and too little time items. We did so as

[30]We only impute too little time items for students who satisfy Criterion 4.

we wanted to be conservative in terms of defining who was non serious. After all, skipping open response items could well be due to not knowing the answer and guessing being a waste of time with open response items. Since we want to estimate performance had all questions been taken seriously, in this section we always impute open response questions (whether non reached or no response) as long as they are taken non seriously. Missing items are not even looked at by students despite having time left. Not even bothering to read the question again is an indication of non seriousness, and this is why we impute the answers. We also impute too little time items but only for people who seem to be paying a price in terms of accuracy for greater speed. Again, these people are not serious.

Multiple imputation involves filling in all the missing data multiple times, creating multiple complete datasets 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.7. They include the individual's scores for other science questions in the test, other participants' scores for all science questions, the individual's characteristics, school characteristics and country fixed effects. The same individual and school characteristics are used by PISA in generating their plausible values. We also use whether the student is non serious or not. If non serious students are more alike in their responses than serious ones, it makes sense to include this variable in the imputation.

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.[31] 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[32], therefore they are modeled using a logistic regression and all other continuous variables are modeled using linear regressions.

One might be concerned that if students spend more time on a question they had skipped/spent too little time on, their behavior may change on the questions that they actually had answered. There are at least two possible channels here. First, they may have less time to spend on other questions. Second, they may be more fatigued after answering/spending more time. Since students have almost two more minutes they can use for

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

[32]In the imputation, we categorize partial credit answers as wrong answers for simplicity. On average students have only 8% of questions in their exams which allow partial credit.

each non-serious item in addition to the time they had already spent, the time constraint is unlikely to be binding, so the first channel seems irrelevant. As far as the second channel goes, our imputation attempts to assign values based on the responses of similar individuals who have the similar observable characteristics and take the same questions in the same order and so should incorporate this potential "fatigue" effect.

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 get the same questions in the same order[33]. Then we conduct multiple imputations within each group. By doing multiple imputations we get the probability of a student answering a given question correctly. From this, we can generate the distribution of total number correct which follows a Poisson binomial distribution. *Ten values* are drawn from this distribution which is unique for each student. Students with no imputations made have the same value for all ten draws.

Next we describe how to calculate student scores and country rankings based on all students' item responses, i.e., in all 72 groups. As different students take different tests, PISA imputes plausible values for a common test using a population model that combines item response theory (IRT) and a latent regression model, see chapter 9 of (OECD (2015b)) for details. 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 of trying to replicate this we do the following. For example, suppose that we calculate fraction correct with skipped items at the end being counted as incorrect using the raw data and assume that this fraction correct follows a normal distribution. We then standardize this score for each group that got the same test (with OECD countries having a mean of 500 and a standard deviation of 100) so that their performance is comparable. Since students are assigned to the 72 groups randomly, we can say that the same kinds of students took each test on average. Standardizing as above controls for different booklets having different levels of difficulty. Since our focus is on country averages/rankings, it is not necessary to control for the difficulty of each question within a booklet as done by PISA, once we have controlled

[33]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.

for the difficulty of each booklet.

The next step is to standardize the imputed score so that it is both comparable across booklets and comparable with the original score. If we just followed what we did for the original score, we would get a score which was comparable across booklets but which could not be compared to the original score distribution as both would be scaled to have a mean of 500 and a standard error of 100 for OECD countries.

Here we use a similar approach as PISA's in Chapter 12 of (OECD (2015b)). Going from fraction correct of the original data to the normalized data involves an adjustment to the mean and the variance since the distributions are assumed to be normal. For example, if the original data, X, had mean μ and variance σ^2, the normalized variable, Y, would be given by

$$Y = AX + B$$

where $A = \frac{100}{\sigma}$ and $B = 500 - \frac{100\mu}{\sigma}$. These 72 pairs of adjustment factors for the mean and variance are then applied to the imputed fraction correct to get the normalized imputed scores which are comparable across both booklets and allow the original score to be compared to the imputed ones. We do this for each of the *ten draws* and thus get *ten imputed scores* for each student. Since PISA also generates ten plausible values, we follow their approach to calculate the mean and standard deviation for each country for the original or imputed versions of the normalized scores[34]. Note that our scores and those in the PISA 2015 report are not comparable directly as they use scores in 2006 for the Science part as the base while we do not.

Table 4 contains the heart of the analysis. In order to understand the effect of being non-serious on country scores, we compare the scores (always normalized as above) after we impute the data for items not taken seriously according to criteria 1-4 to the scores under the status quo. One status quo takes the normal practice of assigning zero to all skipped items. These scores are shown in the first column. The PISA way of doing (treating skipped items at the end as not administered) is used as the status quo in the second column[35]. In the third column, the fully imputed score is shown. The fourth column gives the imputed score when skipped items at the end are ignored. Standard errors are below each score.

The fifth, sixth and seventh columns give the t-statistic for the significance of the difference in column 1 and 3, columns 2 and 3, and 3 and 4 respectively. Comparing columns 1

[34]See page 148 of OECD (2015b), chapter 9.

[35]To quote PISA (page 149 of OECD (2015a))
"Omitted responses prior to a valid response are treated as incorrect responses; whereas, omitted responses at the end of each of the two one-hour test sessions in both PBA and CBA are treated as not reached/not administered."

and 3 we compare the score when all items count to the imputed score. As seen in column 4, these are significantly different for 50 out of 58 countries at the 5% level and for 46 of them also at the 1% level. This means that if a country could make its students take the exam seriously, it could do much better. Comparing columns 2 and 3, we see that using the PISA approach as the status quo brings these numbers closer together. A smaller fraction are significantly different from one another - only 7 differ at the 5% level of significance and 2 at the 1% level. Thus, the PISA way of treating skipped items at the end as not administered goes part way toward accounting for non seriousness. Finally, comparing columns 3 and 4, we see that imputing all the items and imputing only the no response (skipped in the middle of the exam) and too little time items give results that are essentially the same as none differ significantly.

Table 5 presents the list of countries and their ranks before and after imputation. Column 1 shows the rank based on the original scores, i.e., column 1 in Table 4. Column 2 shows the rank based on the imputed scores, i.e., column 3 in Table 4. This corresponds to every country becoming serious. Column 3 shows the rank if only this country is serious. Column 4 shown the rank if all other countries become serious and this country does not. Below each rank is the corresponding rank interval at the 95% confidence levels.

Comparing columns 1 and 3, we see that 54 of 58 countries differ in the two columns. Among them 24 countries have significantly different ranks as the intervals do not overlap. Notice that countries always move up in the ranking in this thought experiment as their score can only rise with the imputation. This change captures the extent to which a single country could strategically raise its rankings by somehow getting its own students to take the exam seriously.

Similarly, while the rank in columns 1 and 4 (all other countries become serious) differ for 55 countries, only 26 of them are significantly different. If other countries become serious, while you do not, your ranking can only fall. Again, some countries are less affected than others. Singapore for example is unaffected even in this case, while Ireland would fall from 18 to 31 if this were to happen.

Finally, the rank between columns 1 and 2 (everyone becomes serious) differ for 36 countries, but only 3 of these are significantly different. In other words, if all countries become serious, there is little significant change in the rankings. As is evident, some countries rise in the rankings (Japan) while others fall (Slovenia). However, overall there is a far smaller 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.

Looking at some interesting individual countries, we see that Singapore and Chinese

Taipei (Taiwan) do not change rank between columns 1 and 3 , 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 564.9 in column 1 of Table 4 while the next ranked country, Taiwan, has a score of 547.8. Similarly, the Dominican Republic which is last has a score of 365.4 while Tunisia, which is second last, has a score of 395.6.[36] Small wonder that Singapore stays first in all the columns and the Dominican Republic stays last.

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

6 Proportion, Ability and Extent

When we impute the data 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

[36] These numbers differ slightly from the numbers in the original working paper posted as we used sampling weights for each student in this version and not in the earlier one. The ranks do not change across the versions.

for this country in the data is

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

while the fraction correct after imputation is

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

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:

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

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 linear combination of the increase in the fraction correct for serious and non-serious students. It is worth noting that $\frac{Y_{ns}}{IFC}$ is 0.74 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.[37]

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

A_{ns} is the average 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 average of 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 $\ln a_{ns}$, $\ln e_{ns}$, $\ln p_{ns}$ separately on $\ln y_{ns}$, that is,

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

[37]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. Here we also include open response items which are non-reached or no-response.

[38] 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.[39] Column 1 shows that for non-serious students, proportion accounts for 68% of the increase in fraction correct while the extent of non-seriousness accounts for about 26%, and least important is ability which accounts for only 6% of the variation. Column 2 shows the similar results for partially-serious students. Proportion accounts for 64% of the variation for serious students, while extent accounts for 32% and ability accounts for 4%.

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. The same story applies to Dominican Republic. In contrast, both Russia and Portugal who also have high y_{ns} have the contribution of ability being high since their non-serious students are quite able. Both Netherlands and Turkey gain very little because the proportion of their non-serious students are very low, so are these students' ability and extent of non-seriousness. US's non-serious students ability, extent and proportion roughly track their gains as all these values are at a median level among all countries.

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

[39] Imputed number correct is calculated by taking the mean of ten draws of number correct.

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 socioeconomic status students tend to be more likely to be non-serious. Exam fatigue is also 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.

We show that scores and rankings change substantially when non-seriousness of the students is taken into account. The comparison between fully imputed score (FIS) and the original score (OS) shows that most of the countries increased their scores significantly were a country to make its students take the exam seriously. For example, Brazil's score increases by 29 points and its fraction correct increases by 6.72%. This change leads to a rise of 5 places in the rankings from 56 to 51. We also show that 24 out of 58 countries increase their rank significantly, i.e., rank confidence intervals of OS and FIS do not overlap. A country can improve by up to 15 places if its students are encouraged to take the exam seriously, but if all countries become serious, then the change in the rankings would be small. The PISA approach partially accounts for non-seriousness by treating skipped items at the end as not administered.[40] However, such an approach is subject to manipulation: a country can game the system by instructing its students to spend as much time as they need on earlier questions and to quit the latter questions if they do not have time or feel tired.

Finally, we decompose the source of the increase in fraction correct into the part that comes from the proportion, ability, and extent (intensity). Using a standard decomposition, we show that the contribution of the three components varies widely across countries. For example, the Dominican Republic has a large increase in fraction correct because it has a high proportion of non-serious students who take a large fraction of questions non seriously. However, the contribution of ability is relatively small as its non-serious students

[40]However, they do not account for skipped items in the middle and too little time items.

are of low-ability. The Russian Federation has a similar gain in fraction correct despite its proportion of non-serious students being much lower. The reason is that their non-serious students have much higher ability. We also show that across all countries, roughly 68% of the increase in fraction correct comes from the proportion component, 26% comes from the extent component and 6% comes from the ability 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.

Baumert, J. and Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, 16(3):441.

Borghans, L. and Schils, T. (2012). The leaning tower of pisa: decomposing achievement test scores into cognitive and noncognitive components. *Unpublished manuscript*.

Butler, J. and Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, 8(3):279–304.

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.

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.

Hanushek, E. A., Link, S., and Woessmann, L. (2013). Does school autonomy make sense everywhere? panel estimates from pisa. *Journal of Development Economics*, 104:212–232.

Hanushek, E. A. and W ößmann, L. (2006). Does educational tracking affect performance and inequality? differences-in-differences evidence across countries. *The Economic Journal*, 116(510):C63–C76.

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., 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.

Lavy, V. (2015). Do differences in schools' instruction time explain international achievement gaps? evidence from developed and developing countries. *The Economic Journal*, 125(588):F397–F424.

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.

Lounkaew, K. (2013). Explaining urban–rural differences in educational achievement in thailand: Evidence from pisa literacy data. *Economics of Education Review*, 37:213–225.

OECD (2015a). Pisa 2015 results(volumn 1): Excellence and equity in education. Technical report, OECD.

OECD (2015b). 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.

Prince Edward Island (2002). Preparing students for pisa, mathematical literacy, teacher's handbook. Technical report, Prince Edward Island.

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

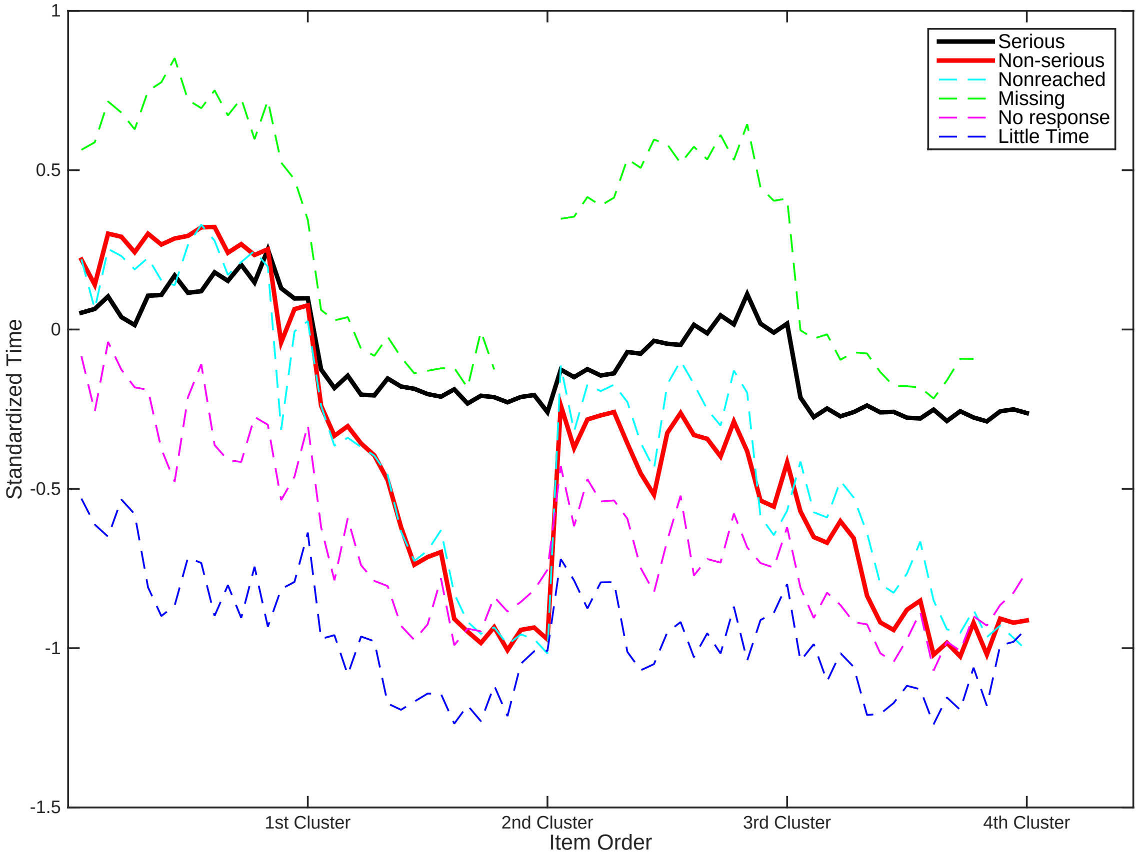


Table 1: Fraction of Non-serious items

Country	Fraction of Non-reached items (%)	Fraction of No-response items (%)	Fraction of Missing items (%)	Fraction of Too-little-time items (%)
Singapore	0.62	1.30	0.57	2.15
Chinese Taipei	0.58	1.98	0.19	1.74
Estonia	0.92	1.83	0.86	1.94
Japan	0.97	2.78	1.18	1.65
Finland	0.75	2.13	0.72	1.67
Hong Kong	0.65	1.60	0.68	2.05
USA (Massachusetts)	0.45	1.18	1.83	2.06
Canada	1.02	2.09	0.86	1.72
Macao	0.31	0.98	2.21	2.14
Slovenia	1.11	3.27	0.32	1.77
B-S-J-G (China)	0.87	2.02	0.75	2.02
Netherlands	0.71	1.61	0.03	2.24
Korea	1.06	2.51	0.04	1.87
United Kingdom	1.39	3.31	0.52	1.60
Germany	1.38	3.43	1.51	1.57
Australia	1.37	3.20	2.32	1.25
New Zealand	1.46	3.38	3.14	1.36
Ireland	1.05	2.10	0.60	1.95
Poland	1.14	3.02	0.85	2.13
Denmark	1.57	3.30	1.61	1.63
Switzerland	1.50	3.47	1.58	1.82
USA (North Carolina)	0.43	1.22	1.82	1.84
Belgium	1.35	3.06	2.42	1.62
Austria	1.34	4.00	0.70	1.41
Norway	1.75	3.59	1.57	1.53
Czech Republic	1.25	3.84	1.10	1.56
United States	0.61	1.44	2.49	1.75
Spain (Regions)	1.21	2.88	1.96	1.83
France	2.19	4.75	1.67	1.37
Spain	1.21	2.91	2.53	1.85
Portugal	1.37	3.40	3.99	0.97
Latvia	0.82	2.25	1.08	1.67
Sweden	2.06	4.76	3.37	1.19
Italy	1.70	4.08	1.37	1.40
Lithuania	1.41	3.77	0.80	1.26
Luxembourg	1.57	4.27	2.45	1.58
Hungary	1.18	3.89	1.74	1.52
Croatia	1.28	4.35	1.06	1.32
Russian Federation	1.37	3.47	4.96	1.26
Iceland	1.67	3.75	1.90	1.44
Slovak Republic	1.31	4.20	2.04	1.26
Israel	1.96	4.37	3.74	1.78
Greece	1.73	3.95	0.96	1.67
Bulgaria	2.15	6.14	2.91	1.08
Chile	2.26	4.05	3.37	1.44
United Arab Emirates	1.68	3.11	0.57	1.42
Turkey	1.28	4.26	0.14	1.57
Uruguay	2.87	6.44	4.83	0.61
Qatar	3.73	4.95	0.26	2.02
Thailand	0.35	1.89	4.22	0.70
Costa Rica	1.27	3.22	5.89	1.16
Colombia	2.32	2.78	3.86	1.20
Montenegro	2.94	9.54	3.61	0.73
Mexico	1.09	1.98	7.76	1.16
Peru	1.07	3.46	12.44	1.01
Brazil	1.91	5.57	20.40	0.17
Tunisia	5.11	7.20	6.68	0.45
Dominican Republic	14.97	7.94	1.22	0.62
Overall	**1.62**	**3.48**	**3.04**	**1.46**

Note: In this table non-reached items include non-reached open response items and no-response items including no-response open response items.

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

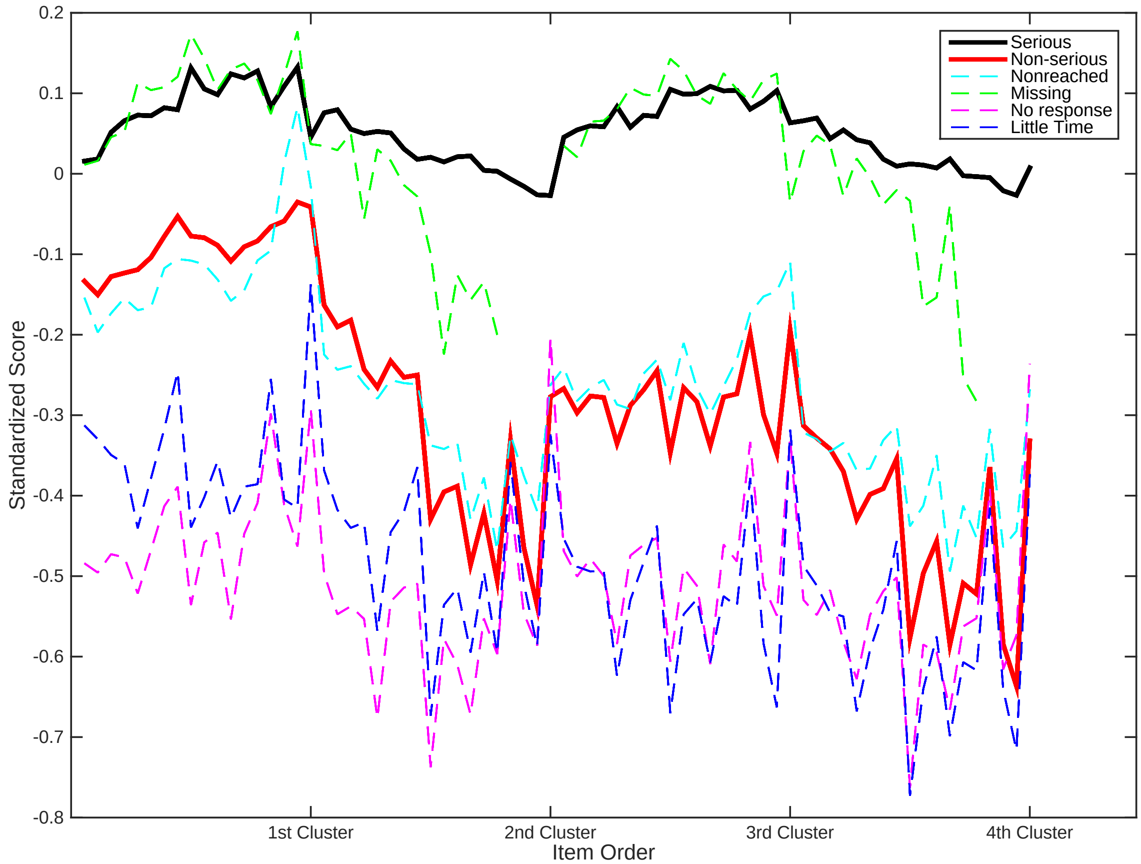
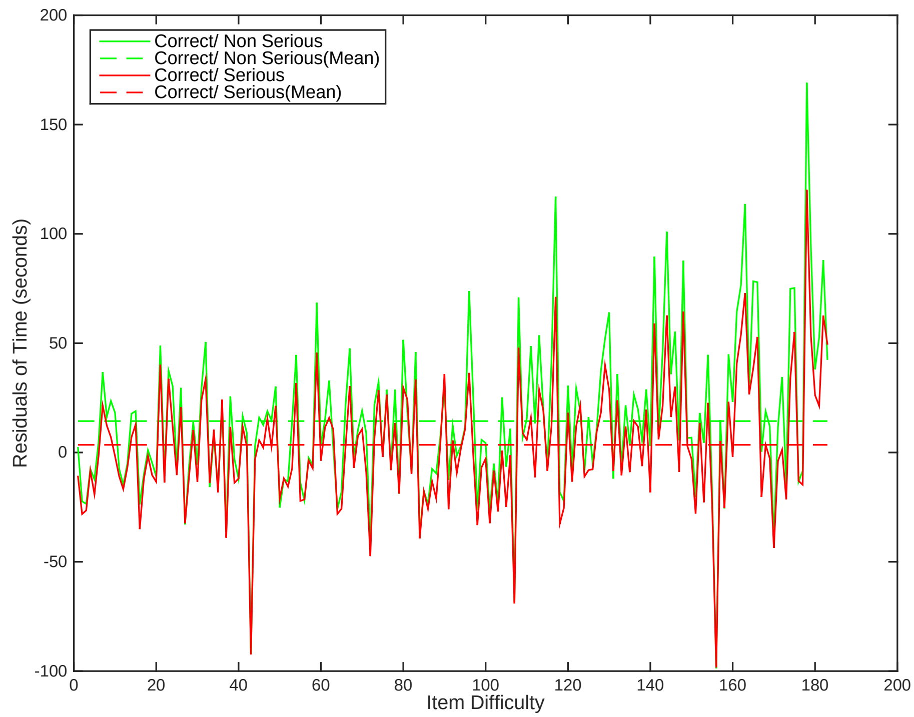


Figure 3: Time for Correct and Incorrect Answers for Serious and Non-serious Students

(a) Time for Correct Answers



(b) Time for Incorrect Answers

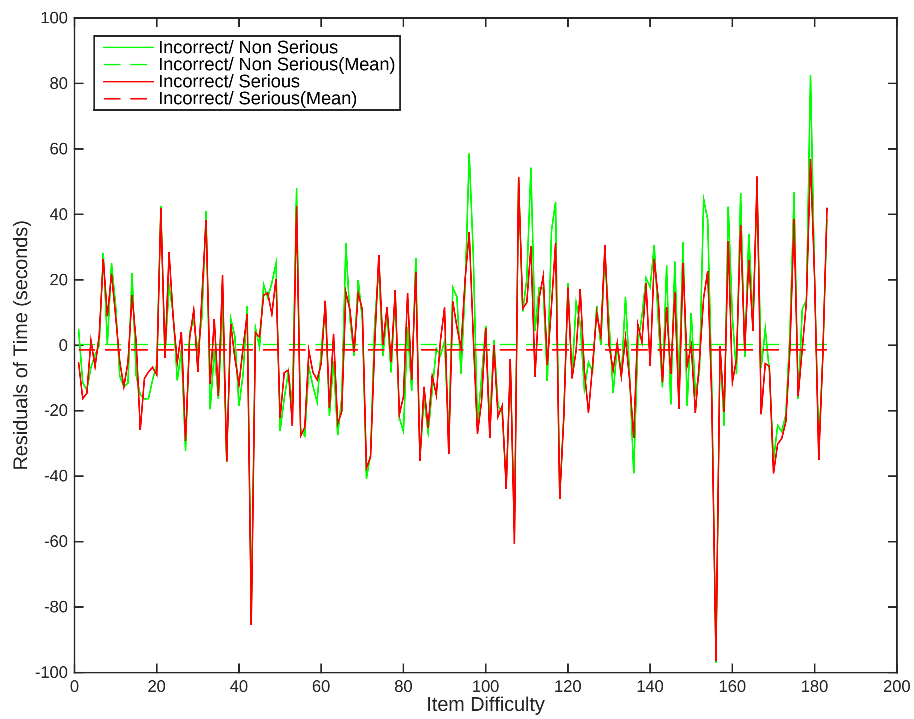
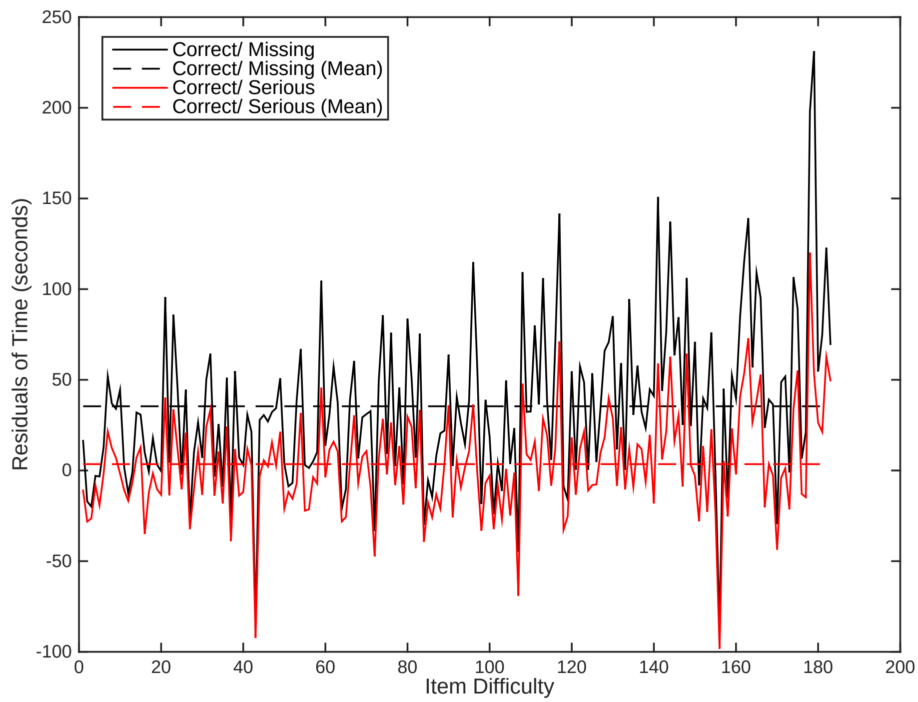


Figure 4: Time for Correct and Incorrect Answers for Serious and Missing-item Students

(a) Time for Correct Answers



(b) Time for Incorrect Answers

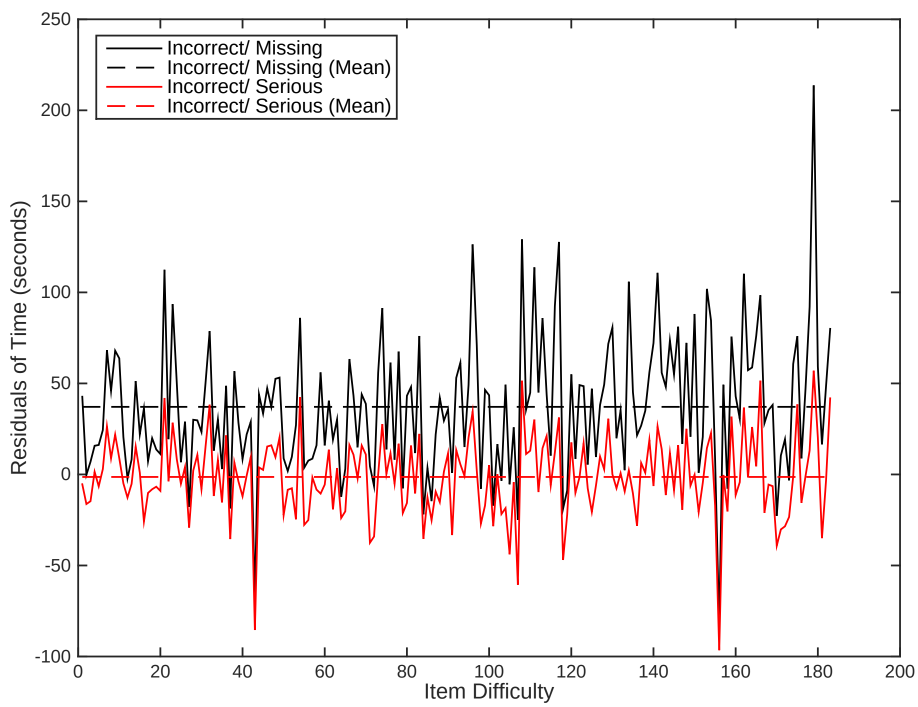
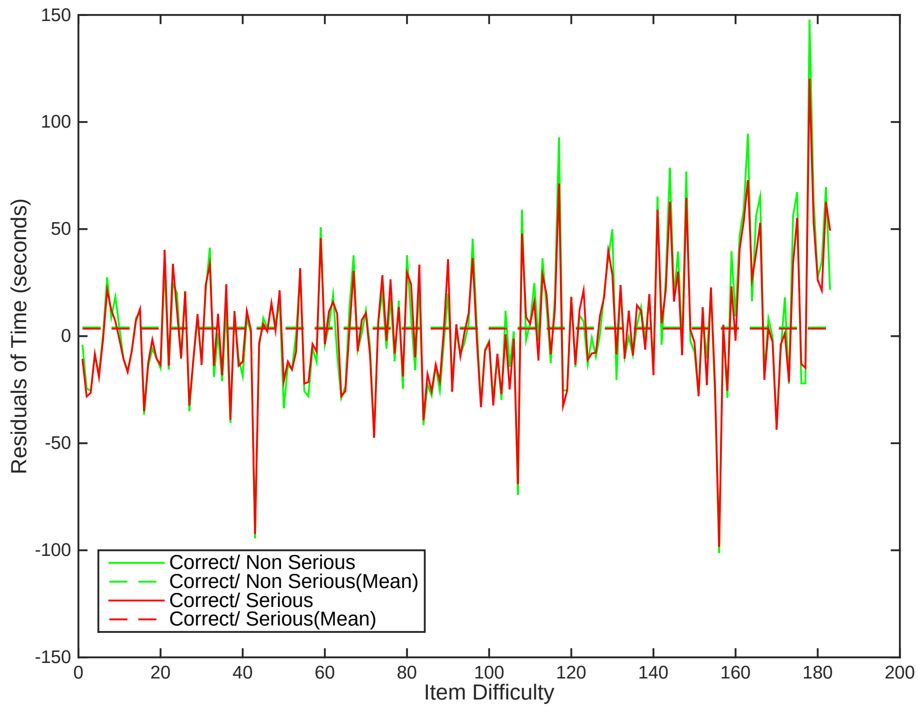


Figure 5: Time for Correct and Incorrect Answers for Serious and Non-serious Students After Removing Missing-item Students

(a) Time for Correct Answers



(b) Time for Incorrect Answers

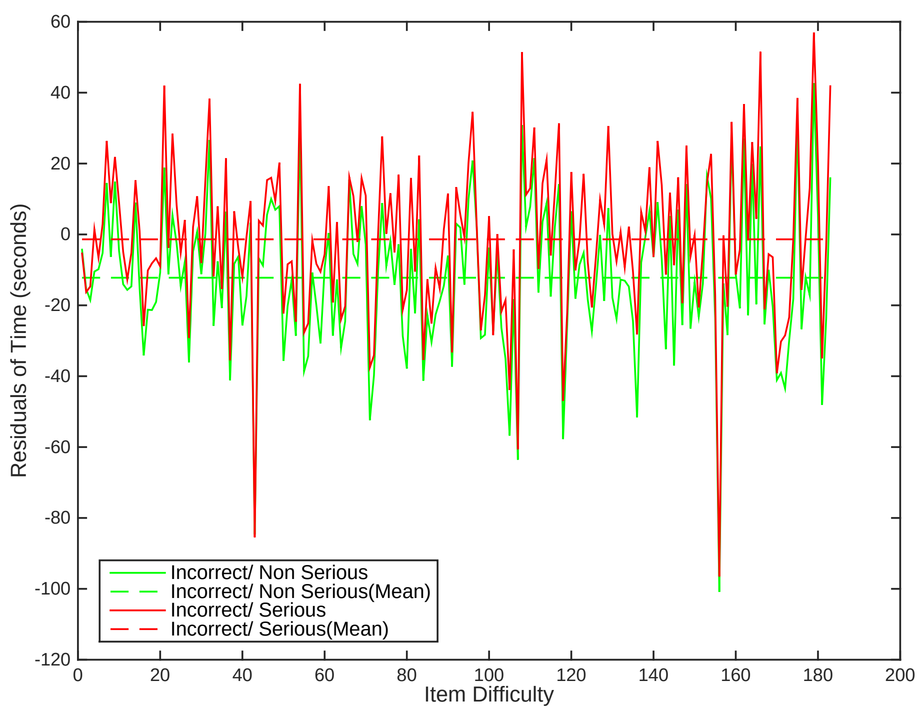


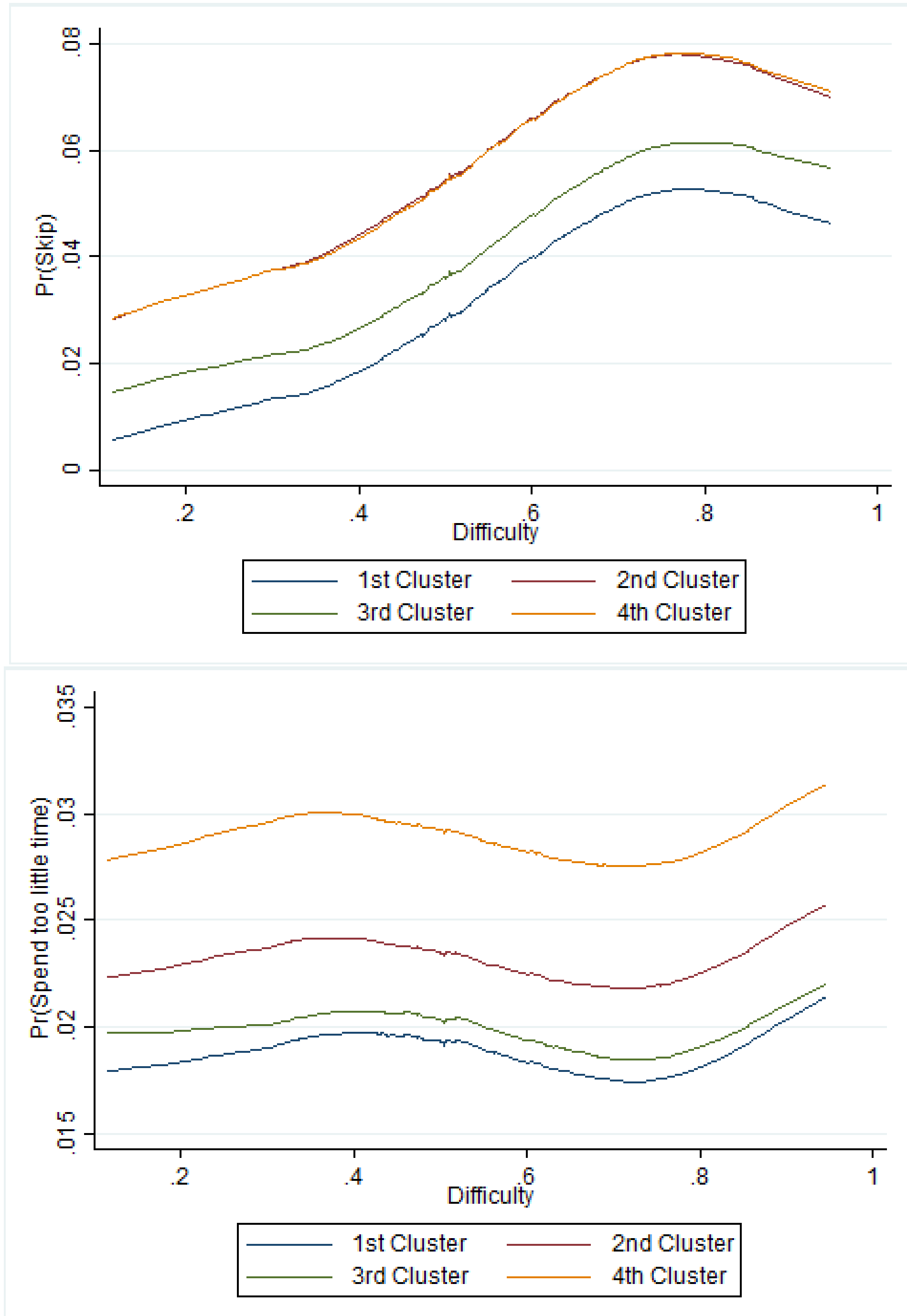
Table 2: Factors Related to Being Non-Serious

	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.

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

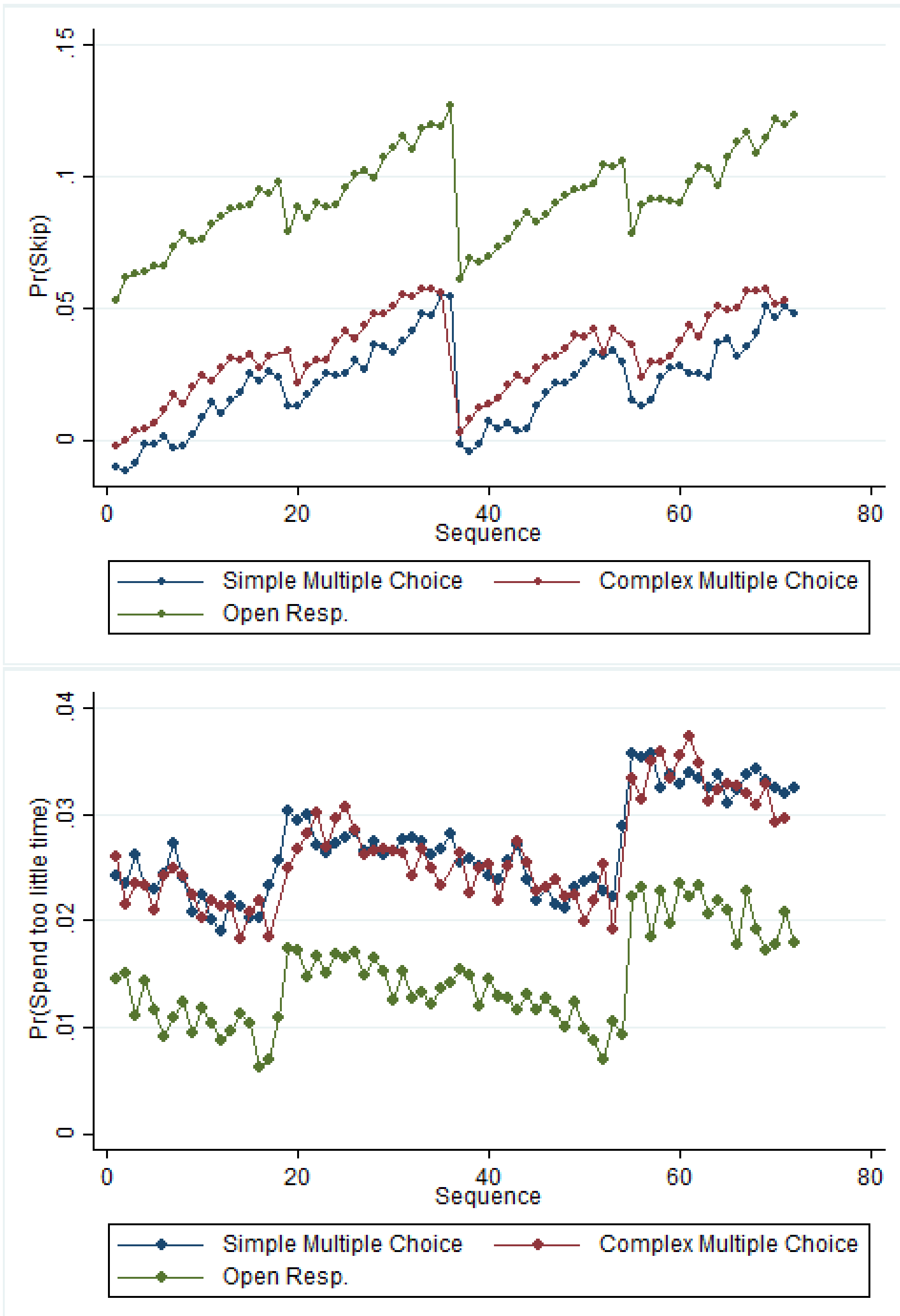


Table 3: Factors Affecting Pr(Skip) and Pr(Spend too little time) (Logit results)

	Skip	Spend too little time
Sequence	0.0675***	-0.0150***
	(0.0004)	(0.0005)
Difficulty	0.6159***	0.6643***
	(0.0126)	(0.0150)
Cluster 2	0.5932***	0.0265***
	(0.0053)	(0.0068)
Cluster 3	0.2714***	-0.0197***
	(0.0056)	(0.0069)
Cluster 4	0.6757***	0.2926***
	(0.0052)	(0.0064)
Complex Multiple Choice	0.3543***	-0.1220***
	(0.0066)	(0.0056)
Open Response	1.6828***	-0.7622***
	(0.0062)	(0.0072)
Log(math score)	-3.1288***	-1.7452***
	(0.0121)	(0.0167)
Log per capita GDP	6.6012***	8.0347***
	(0.1043)	(0.1477)
(Log per capita GDP)^2	-0.3084***	-0.3668***
	(0.0050)	(0.0070)
ESCS	0.0122***	0.0109***
	(0.0023)	(0.0028)
ESCS^2	-0.0325***	-0.0026
	(0.0012)	(0.0016)
Grade	-0.0401***	-0.0538***
	(0.0024)	(0.0033)
Female	-0.0090**	-0.2169***
	(0.0036)	(0.0048)
Anxiety	0.0015	-0.0236***
	(0.0028)	(0.0036)
Ambition	-0.1155***	0.0106***
	(0.0030)	(0.0041)
Skipping class/Arriving late	0.0505***	0.0342***
	(0.0010)	(0.0014)
Additional learning time (hrs/week)	-0.0026***	0.0040***
	(0.0001)	(0.0002)
Time on classes	0.0019***	0.0014***
	(0.0001)	(0.0002)
Standardized test frequency	0.0259***	0.0209***
	(0.0024)	(0.0032)
Teacher-developed tests frequency	-0.0112***	-0.0154***
	(0.0018)	(0.0024)
Stake of Standardized test	-0.0060***	0.0000
	(0.0003)	(0.0004)
Stake of Teacher-developed tests	-0.0099***	0.0030***
	(0.0004)	(0.0005)
School average science score	-1.6290***	0.3706***
	(0.0181)	(0.0253)
Observations	9,058,210	9,058,210

Table 4: Country Scores After Different Imputations

Country	OS	SENA	FIS	Imputed SENA	t-statistics		
					Difference	Difference	Difference
	(1)	(2)	(3)	(4)	(1)- (3)	(2)- (3)	(3)- (4)
Singapore	564.9	567.3	570.9	570.8	4.31	2.14	0.02
	(1.06)	(1.02)	(1.32)	(1.4)			
Chinese Taip	547.8	549	553.3	553	2.79	1.26	0.1
	(2.42)	(2.4)	(2.52)	(2.63)			
Estonia	546.8	550.8	555.1	555.1	3.87	1.42	0.02
	(1.96)	(2.01)	(2.29)	(2.31)			
Japan	546.5	554.4	557	560.3	5.4	0.62	-0.76
	(2.75)	(2.83)	(3.04)	(3.17)			
Finland	544.2	549.3	552.4	554	2.69	0.99	-0.51
	(2.04)	(2.07)	(2.24)	(2.33)			
Hong Kong	543.6	546.5	551.2	551.1	2.79	1.25	0.04
	(2.57)	(2.55)	(2.71)	(2.79)			
USA (Massachusrtts)	540.2	544.8	548.5	548.5	3.91	0.44	0
	(6.34)	(6.08)	(5.93)	(5.9)			
Canada	538.8	542.7	546.6	546.9	1.58	1.41	-0.08
	(1.88)	(1.85)	(2.1)	(2.11)			
Macao	535	541.3	544	544.4	3.37	1.43	-0.2
	(1)	(0.98)	(1.61)	(1.5)			
Slovenia	529.9	532.5	536.9	537.2	3.88	2.05	-0.11
	(1.31)	(1.31)	(1.73)	(1.93)			
B-S-J-G (China)	529.4	532.8	537.5	537.7	2.95	0.77	-0.03
	(4.27)	(4.23)	(4.38)	(4.43)			
Netherlands	526.4	527.2	530.9	530.7	3.61	1.13	0.04
	(2.25)	(2.23)	(2.31)	(2.35)			
Korea	526.1	530.3	532.1	535.1	2.97	0.42	-0.68
	(3.02)	(3.02)	(3.06)	(3.18)			
United Kingd0m	524.1	527.2	532.7	532.7	4.62	1.64	0.01
	(2.25)	(2.24)	(2.48)	(2.6)			
Germany	523	529.1	535	535.2	2.78	1.53	-0.05
	(2.5)	(2.51)	(2.89)	(2.94)			
Australia	518.2	524.8	528.4	529.3	2.71	1.52	-0.32
	(1.4)	(1.4)	(1.93)	(1.83)			
New Zealand	517.5	527.6	532.7	533.8	4.34	1.36	-0.27
	(2.47)	(2.35)	(2.97)	(2.72)			
Ireland	517	520.3	524.7	524.7	3.13	1.36	-0.01
	(2.15)	(2.13)	(2.39)	(2.49)			
Poland	515.9	520.2	525.8	526.1	2.01	1.51	-0.06
	(2.48)	(2.47)	(2.82)	(3)			
Denmark	515.5	521.2	525.4	526.1	2.02	1.23	-0.2
	(2.16)	(2.24)	(2.51)	(2.55)			
Switzerland	514.9	523.3	526.7	529	3.18	0.81	-0.54
	(2.64)	(2.72)	(3.09)	(3.03)			
USA (North Carolina)	513.9	518.5	521.7	521.8	4.36	0.46	-0.02
	(5.1)	(4.95)	(4.97)	(4.92)			
Belgium	512	520.8	525.2	526.2	2.37	1.27	-0.27
	(2.06)	(2.16)	(2.7)	(2.6)			
Austria	511.8	515.4	521.7	521.8	3.23	1.79	-0.03

(continued on next page)

Country	OS	SENA	FIS	Imputed SENA	t-statistics Difference	Difference	Difference
	(1)	(2)	(3)	(4)	(1)-(3)	(2)-(3)	(3)-(4)
	(2.33)	(2.31)	(2.66)	(2.84)			
Norway	510.4	517.2	523.2	524	3.25	1.8	-0.19
	(2.1)	(2.07)	(2.6)	(2.62)			
Czech Republic	507.7	512.5	518.8	519.1	2.56	2.02	-0.06
	(1.89)	(1.94)	(2.45)	(2.61)			
United States	506.8	512.6	515.6	516.1	1.38	0.7	-0.1
	(3.01)	(3.01)	(3.21)	(3.08)			
Spain (Regions)	506.6	513.6	518.5	518.8	2.9	1.99	-0.11
	(1.25)	(1.28)	(2.09)	(1.98)			
France	504.9	512.9	519	520.2	2.11	1.88	-0.31
	(1.92)	(1.92)	(2.63)	(2.66)			
Spain	504.6	512	517.1	517.4	5.91	1.62	-0.06
	(1.85)	(1.94)	(2.52)	(2.47)			
Portugal	502.4	516.5	520.1	521.4	4.76	1.02	-0.33
	(1.93)	(2.07)	(2.92)	(2.51)			
Latvia	501.6	506	508.9	509.9	3.9	1.16	-0.36
	(1.58)	(1.55)	(1.96)	(1.98)			
Sweden	499.2	510.5	517.3	518.3	5.72	1.32	-0.19
	(3.19)	(3.3)	(3.9)	(3.77)			
Italy	493.9	500.1	506.3	506.9	1.38	1.61	-0.14
	(2.41)	(2.45)	(2.98)	(3.05)			
Lithuania	491.9	495	499.6	500	3.94	1.26	-0.1
	(2.46)	(2.45)	(2.71)	(2.82)			
Luxembourg	491.7	499.4	505.3	505.9	3.84	2.61	-0.2
	(0.97)	(0.94)	(2.09)	(2.05)			
Hungary	491.4	499	503	505.3	4.22	1.08	-0.56
	(2.3)	(2.33)	(2.85)	(2.9)			
Croatia	489.7	493.9	500.5	500.6	2.65	1.79	-0.03
	(2.38)	(2.39)	(2.81)	(3.1)			
Russian Federation	485.9	500.3	504.6	506	5.06	0.92	-0.27
	(3.09)	(3.02)	(3.61)	(3.34)			
Iceland	484	492.1	496.3	498.1	4.94	1.47	-0.56
	(1.57)	(1.67)	(2.34)	(2.29)			
Slovak Republic	476.9	484.6	488.7	490.1	3.94	1.17	-0.36
	(2.25)	(2.26)	(2.71)	(2.81)			
Israel	476.6	487.2	492.4	493.8	3.53	1.05	-0.28
	(3.07)	(3.14)	(3.8)	(3.52)			
Greece	468.2	472.7	478.2	478.6	3.36	1.09	-0.07
	(3.3)	(3.42)	(3.7)	(3.81)			
Bulgaria	459.4	468.7	475.1	476.4	3.22	1.08	-0.21
	(3.83)	(3.95)	(4.39)	(4.38)			
Chile	457.2	467	471.8	472.9	4.01	1.26	-0.27
	(2.12)	(2.16)	(3.07)	(2.83)			
United Arab	456.3	459.1	462.5	463	3.58	1.06	-0.13
	(2.17)	(2.2)	(2.39)	(2.5)			
Turkey	446.9	448.5	453.4	453.5	2.88	0.9	-0.02
	(3.72)	(3.73)	(3.92)	(4.06)			
Uruguay	443	456.6	461.7	464.2	2.33	1.32	-0.56
	(1.98)	(2.03)	(3.29)	(3.03)			
Qatar	437.7	442.4	447.8	448.5	1.92	2.67	-0.26

(continued on next page)

Country	OS	SENA	FIS	Imputed SENA	t-statistics Difference	Difference	Difference
	(1)	(2)	(3)	(4)	(1)-(3)	(2)-(3)	(3)-(4)
	(0.73)	(0.72)	(1.91)	(1.98)			
Thailand	433.5	441.3	442.5	444.1	4.55	0.31	-0.38
	(2.5)	(2.7)	(2.97)	(2.84)			
Costa Rica	429	440.2	442.6	444.3	1.21	0.67	-0.45
	(1.98)	(1.95)	(2.9)	(2.35)			
Colombia	427.6	437.4	438.9	441	2.56	0.45	-0.58
	(1.94)	(1.99)	(2.75)	(2.36)			
Montenegro	424.4	436	442.5	446.9	2.02	2.03	-1.04
	(0.96)	(1.02)	(3.02)	(3.07)			
Mexico	422.3	435.4	436.7	438.5	4.87	0.37	-0.46
	(2.02)	(1.91)	(3.12)	(2.22)			
Peru	404.6	420.7	422.7	424.8	1.33	0.47	-0.46
	(2.03)	(2.04)	(3.78)	(2.61)			
Brazil	400	428.6	429.1	434.8	4.89	0.09	-0.96
	(1.76)	(2.04)	(5.09)	(2.87)			
Tunisia	395.6	410.4	413.6	417	0.96	0.83	-0.75
	(1.96)	(1.79)	(3.45)	(2.99)			
Dominican Republic	365.4	378.8	385.2	386.1	1.1	1.49	-0.17
	(1.69)	(1.75)	(3.93)	(3.01)			

Note: Standard errors are in parentheses.
OS: Original score calculated by assigning zero to all skipped items.
SENA: Original score calculated by treating skipped items at the end are not administered.
FIS: Fully imputed score.
Imputed SENA: Imputed score when skipped items at the end are ignored.

Table 5: Country Ranks After Different Imputations

Country	Original Rank (1)	All Countries Serious (2)	Only One Country Serious (3)	All Other Countries Serious (4)
Singapore	1	1	1	1
	(1,1)	(1,1)	(1,1)	(1,1)
Chinese Taipei	2	4	2	7
	(2,6)	(2,7)	(2,2)	(4,9)
Estonia	3	3	2	7
	(2,6)	(2,6)	(2,2)	(5,9)
Japan	4	2	2	8
	(2,6)	(2,5)	(2,2)	(4,9)
Finland	5	5	2	8
	(2,7)	(3,7)	(2,2)	(6,9)
Hong Kong	6	6	2	9
	(2,8)	(3,8)	(2,5)	(6,9)
USA (Massachusetts)	7	7	2	9
	(2,11)	(2,12)	(2,8)	(5,17)
Canada	8	8	4	9
	(7,8)	(7,9)	(2,7)	(9,12)
Macao	9	9	6	11
	(9,9)	(9,9)	(3,7)	(9,15)
Slovenia	10	11	9	16
	(10,11)	(10,12)	(7,10)	(12,19)
B-S-J-G (China)	11	10	9	16
	(9,15)	(9,16)	(5,11)	(10,25)
Netherlands	12	16	10	18
	(10,15)	(12,18)	(9,12)	(16,25)
Korea	13	15	10	18
	(10,15)	(10,18)	(9,13)	(13,26)
United Kingdom	14	13	10	22
	(12,15)	(11,17)	(9,12)	(16,28)
Germany	15	12	10	23
	(12,16)	(10,16)	(7,12)	(16,31)
Australia	16	17	12	29
	(16,20)	(16,19)	(10,14)	(23,32)
New Zealand	17	14	10	29
	(16,22)	(11,17)	(9,12)	(22,32)
Ireland	18	22	14	31
	(16,22)	(17,25)	(11,16)	(23,32)
Poland	19	19	14	31
	(16,24)	(17,25)	(10,16)	(23,32)
Denmark	20	20	14	32
	(16,24)	(17,25)	(10,16)	(25,32)
Switzerland	21	18	12	32
	(16,25)	(16,25)	(10,16)	(25,33)
USA (North Carolina)	22	25	16	32
	(15,3)	(16,32)	(10,23)	(23,36)
Belgium	23	21	14	32
	(19,25)	(17,25)	(10,26)	(29,33)
Austria	24	24	16	32

(continued on next page)

Country	Original Rank (1)	All Countries Serious (2)	Only One Country Serious (3)	All Other Countries Serious (4)
	(19,26)	(19,30)	(12,19)	(29,34)
Norway	25	23	15	32
	(22,28)	(18,26)	(12,17)	(31,35)
Czech Republic	26	27	16	33
	(25,30)	(24,32)	(15,22)	(32,37)
United States	27	32	20	33
	(23,32)	(26,32)	(16,26)	(32,38)
Spain (Regions)	28	29	16	33
	(26,30)	(26,31)	(16,22)	(32,37)
France	29	28	16	36
	(26,32)	(24,32)	(15,23)	(32,39)
Spain	30	31	18	36
	(26,32)	(26,32)	(16,23)	(32,29)
Portugal	31	26	16	37
	(29,33)	(23,31)	(14,22)	(33,39)
Latvia	32	33	26	37
	(30,33)	(33,34)	(23,29)	(34,39)
Sweden	33	30	18	39
	(29,34)	(23,32)	(14,26)	(33,41)
Italy	34	34	29	40
	(34,38)	(33,37)	(23,33)	(38,42)
Lithuania	35	39	33	41
	(34,38)	(37,40)	(29,34)	(39,42)
Luxembourg	36	35	29	41
	(35,37)	(34,36)	(26,33)	(40,42)
Hungary	37	37	31	41
	(34,38)	(34,39)	(26,34)	(39,42)
Croatia	38	38	33	41
	(34,39)	(35,40)	(29,34)	(40,42)
Russian Federation	39	36	29	42
	(35,40)	(33,39)	(25,34)	(40,42)
Iceland	40	40	34	42
	(39,50)	(39,40)	(33,36)	(41,42)
Slovak Republic	41	42	39	43
	(41,42)	(41,42)	(34,41)	(42,45)
Israel	42	41	35	43
	(41,42)	(40,42)	(33,40)	(42,45)
Greece	43	43	41	45
	(43,43)	(43,45)	(40,43)	(43,47)
Bulgaria	44	44	43	47
	(44,46)	(43,45)	(41,44)	(45,48)
Chile	45	45	43	47
	(44,46)	(44,45)	(41,44)	(45,48)
United Arab Emirates	46	46	44	47
	(44,46)	(46,47)	(44,45)	(46,48)
Turkey	47	48	47	49
	(47,48)	(48,49)	(44,47)	(48,53)
Uruguay	48	47	44	49
	(47,48)	(46,47)	(44,47)	(48,53)
Qatar	49	49	47	53
	(49,49)	(49,49)	(47,48)	(52,54)

(continued on next page)

Country	Original Rank (1)	All Countries Serious (2)	Only One Country Serious (3)	All Other Countries Serious (4)
Thailand	50	51	49	54
	(49,51)	(49,53)	(47,50)	(52,55)
Costa Rica	51	50	49	55
	(51,52)	(50,53)	(47,50)	(54,55)
Colombia	52	53	49	55
	(51,53)	(50,54)	(48,50)	(54,56)
Montenegro	53	52	49	55
	(53,53)	(50,52)	(47,50)	(54,56)
Mexico	54	54	50	56
	(53,54)	(53,54)	(49,51)	(55,56)
Peru	55	56	54	57
	(55,55)	(56,56)	(51,55)	(57,57)
Brazil	56	55	51	57
	(56,56)	(55,55)	(49,55)	(57,57)
Tunisia	57	57	55	57
	(57,57)	(57,57)	(55,55)	(57,57)
Dominican Republic	58	58	58	58
	(58,58)	(58,58)	(58,58)	(58,58)

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	0.06	0.04
		(0.05)	(0.08)
	De-meaned E	0.26	0.32
		(0.02)	(0.06)
	De-meaned P	0.68	0.64
		(0.04)	(0.04)

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

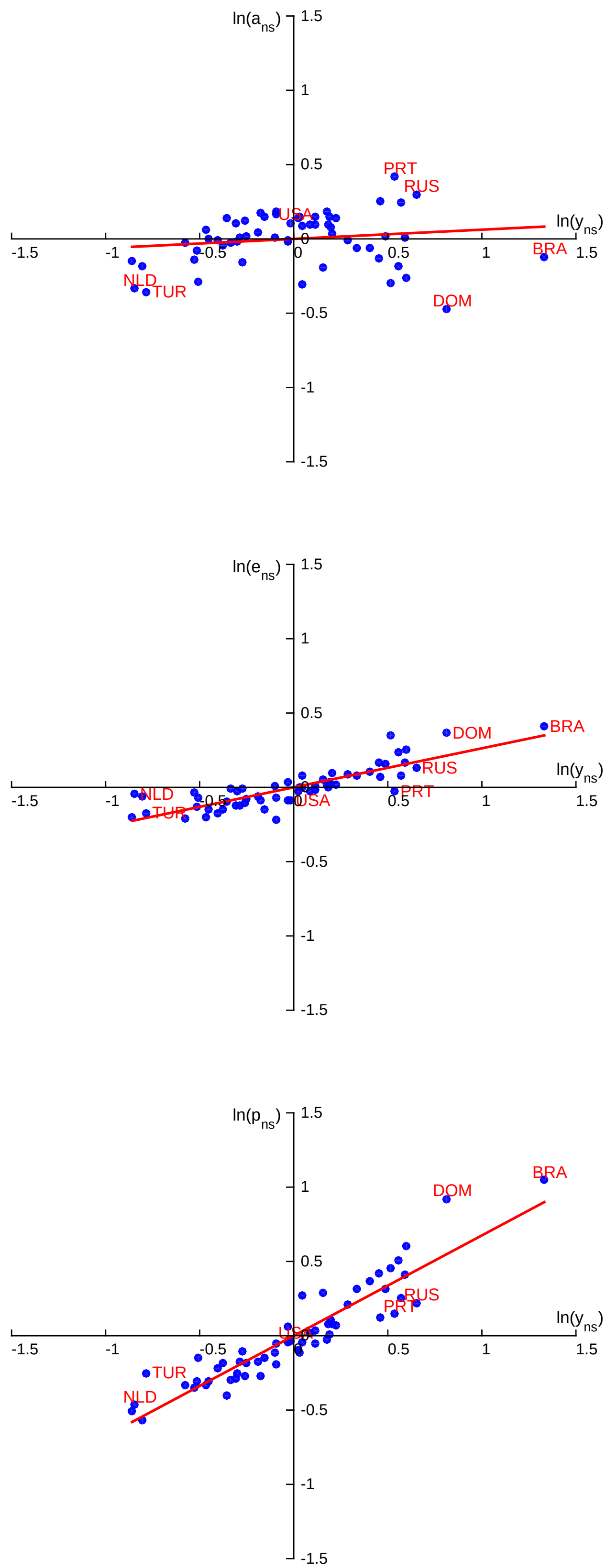


Table 7: Decomposed Factors for Non-Serious Students

Country	$IFC(\%)$	$Y_{ns}(\%)$	A_{ns}	E_{ns}	P_{ns}
Brazil	6.72%	6.25%	0.25	0.37	0.67
Dominican Republic	4.32%	3.73%	0.18	0.35	0.59
Russian Federation	4.09%	3.16%	0.38	0.28	0.29
Uruguay	4.07%	2.98%	0.29	0.29	0.36
Montenegro	4.01%	2.89%	0.24	0.31	0.39
Sweden	3.98%	2.93%	0.36	0.27	0.30
Tunisia	3.98%	2.77%	0.21	0.35	0.37
Peru	3.93%	3.00%	0.22	0.32	0.43
Portugal	3.78%	2.82%	0.43	0.24	0.27
Israel	3.40%	2.69%	0.29	0.29	0.32
Bulgaria	3.38%	2.20%	0.28	0.27	0.29
New Zealand	3.34%	2.61%	0.37	0.27	0.27
France	3.08%	2.02%	0.29	0.27	0.25
Mexico	3.05%	2.59%	0.25	0.29	0.36
Costa Rica	2.98%	2.48%	0.27	0.27	0.34
Luxembourg	2.98%	2.02%	0.31	0.25	0.26
Chile	2.97%	2.30%	0.27	0.27	0.32
Norway	2.82%	1.97%	0.34	0.25	0.23
Belgium	2.79%	2.07%	0.33	0.25	0.25
Spain	2.71%	1.98%	0.31	0.25	0.25
Iceland	2.71%	2.00%	0.33	0.26	0.24
Switzerland	2.63%	1.85%	0.31	0.24	0.24
Germany	2.58%	1.69%	0.33	0.24	0.21
Slovak Repubic	2.56%	1.73%	0.31	0.25	0.22
Italy	2.54%	1.60%	0.28	0.25	0.22
Spain (Region)	2.52%	1.80%	0.31	0.24	0.24
Australia	2.46%	1.85%	0.33	0.25	0.22
Hungary	2.44%	1.50%	0.29	0.25	0.21
Denmark	2.40%	1.71%	0.33	0.25	0.21
Colombia	2.38%	1.92%	0.24	0.26	0.31
Czech Republic	2.38%	1.37%	0.30	0.23	0.20
Croatia	2.36%	1.29%	0.29	0.23	0.20
Japan	2.32%	1.38%	0.34	0.23	0.18
Qatar	2.20%	1.73%	0.21	0.27	0.31
Poland	2.17%	1.24%	0.29	0.22	0.20
Austria	2.16%	1.22%	0.28	0.24	0.18
Greece	2.12%	1.25%	0.24	0.24	0.21
Macao	1.98%	1.51%	0.34	0.20	0.22
United States	1.98%	1.62%	0.32	0.23	0.23
United Kingdom	1.93%	1.18%	0.28	0.24	0.17
Thailand	1.92%	1.60%	0.28	0.23	0.25
USA (Massachusetts)	1.86%	1.50%	0.34	0.23	0.19
Estonia	1.85%	1.22%	0.32	0.22	0.18
Lithuania	1.82%	0.97%	0.25	0.24	0.17
Canada	1.79%	1.28%	0.32	0.22	0.18
Finland	1.79%	1.16%	0.33	0.22	0.16
USA (North Carolina)	1.70%	1.41%	0.33	0.21	0.20
Ireland	1.67%	1.10%	0.28	0.21	0.19
Slovenia	1.66%	0.99%	0.26	0.22	0.17
B-S-J-G (Chi	1.64%	1.13%	0.27	0.21	0.20
Hong Kong	1.63%	1.03%	0.30	0.20	0.17
Latvia	1.61%	1.05%	0.28	0.21	0.17
Turkey	1.44%	0.75%	0.20	0.21	0.18
United Arab	1.38%	0.99%	0.21	0.23	0.20
Singapore	1.32%	0.93%	0.28	0.20	0.17
Korea	1.31%	0.74%	0.24	0.23	0.13
Chinese Taipei	1.22%	0.70%	0.24	0.20	0.14
Netherlands	0.96%	0.71%	0.20	0.24	0.15

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 Fraction of Non-serious items Across Subjects

Table A.1 shows the fraction of no-response items and the fraction of non-reached items for the subject of science, reading and math. The fraction of no response items for the reading and math tests are a bit higher on average than science. Moreover, the fraction of no-response and non-reached items are highly correlated across subjects. For example, the correlation between the fraction of no-response items for science and for reading is 0.98, showing that non seriousness is common across subjects of the test as might be expected.

A.2 Time Spent, Accuracy and Position

Table A.2 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.

The upper part of Table A.3 shows proportion correct for all items (not just answered ones) across positions. Serious students have higher proportion correct than each category of

Table A.1: Fraction of Non-serious items Across Subjects

Country	Fraction of Non-reached items (%)			Fraction of No-response items (%)		
	science	reading	math	science	reading	math
Singapore	0.62	0.46	0.58	1.30	2.22	2.26
Chinese Taipei	0.58	0.37	0.58	1.98	3.83	3.25
Estonia	0.92	0.39	0.83	1.83	3.30	4.34
Japan	0.97	0.76	1.19	2.78	6.44	5.94
Finland	0.75	0.52	1.25	2.13	3.73	5.84
Hong Kong	0.65	0.39	0.56	1.60	3.22	2.81
USA (Massachusetts)	0.45	0.18	0.42	1.18	1.84	2.49
Canada	1.02	0.68	1.18	2.09	3.47	4.06
Macao	0.31	0.13	0.27	0.98	2.02	1.85
Slovenia	1.11	0.47	1.38	3.27	5.78	6.36
B-S-J-G (China)	0.87	0.60	0.61	2.02	4.25	2.54
Netherlands	0.71	0.29	0.92	1.61	2.08	3.31
Korea	1.06	0.53	1.16	2.51	4.86	4.53
United Kingdom	1.39	0.84	1.39	3.31	5.89	6.24
Germany	1.38	0.84	1.34	3.43	5.51	7.18
Australia	1.37	0.85	1.49	3.20	5.28	5.97
New Zealand	1.46	0.98	1.34	3.38	5.70	6.38
Ireland	1.05	0.42	0.89	2.10	3.13	4.31
Poland	1.14	0.40	1.15	3.02	5.51	5.29
Denmark	1.57	1.11	1.55	3.30	5.24	5.39
Switzerland	1.50	1.07	1.60	3.47	6.34	5.94
USA (North Carolina)	0.43	0.21	0.53	1.22	2.19	1.64
Belgium	1.35	0.60	1.11	3.06	5.11	5.88
Austria	1.34	0.54	1.35	4.00	6.74	7.24
Norway	1.75	1.28	2.09	3.59	5.34	6.92
Czech Republic	1.25	0.44	1.31	3.84	6.46	6.85
United States	0.61	0.43	0.69	1.44	2.52	2.28
Spain (Regions)	1.21	0.65	1.43	2.88	4.48	6.85
France	2.19	1.68	2.30	4.75	7.69	8.08
Spain	1.21	0.60	1.55	2.91	4.62	6.80
Portugal	1.37	0.38	1.43	3.40	6.46	6.95
Latvia	0.82	0.26	0.88	2.25	3.78	4.78
Sweden	2.06	1.54	2.58	4.76	6.35	8.31
Italy	1.70	0.77	1.63	4.08	5.78	7.42
Lithuania	1.41	0.65	1.24	3.77	6.48	6.73
Luxembourg	1.57	0.76	1.44	4.27	7.36	6.79
Hungary	1.18	0.37	1.20	3.89	7.76	6.82
Croatia	1.28	0.39	1.64	4.35	6.92	8.87
Russian Federation	1.37	0.68	1.24	3.47	5.05	5.56
Iceland	1.67	0.91	1.67	3.75	6.12	5.91
Slovak Republic	1.31	0.56	1.15	4.20	7.47	5.76
Israel	1.96	1.06	2.07	4.37	6.90	7.91
Greece	1.73	0.98	1.59	3.95	6.76	7.28
Bulgaria	2.15	1.15	1.07	6.14	10.35	8.68
Chile	2.26	1.10	1.46	4.05	7.04	9.30
United Arab Emirates	1.68	1.22	1.23	3.11	5.33	4.36
Turkey	1.28	0.56	1.24	4.26	8.63	6.58
Uruguay	2.87	2.39	2.02	6.44	10.00	11.79
Qatar	3.73	3.57	3.02	4.95	8.74	7.20
Thailand	0.35	0.17	0.42	1.89	3.57	2.73
Costa Rica	1.27	0.81	0.94	3.22	6.45	7.18
Colombia	2.32	1.40	1.45	2.78	5.06	5.30
Montenegro	2.94	1.99	2.96	9.54	15.87	14.52
Mexico	1.09	0.52	0.72	1.98	3.59	4.50
Peru	1.07	0.58	0.71	3.46	6.33	8.30
Brazil	1.91	1.40	1.55	5.57	10.30	9.54
Tunisia	5.11	4.43	2.89	7.20	12.14	9.55
Dominican Republic	14.97	11.68	7.89	7.94	13.19	13.55

Note: In this table non-reached items include non-reached open response items and no-response items including no-response open response items.

Table A.2: 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

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.3. 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.

A.3 Factors Related to Being Non-Serious

In Table A.4 and Table A.5, we present correlates of each non-seriousness criterion for cutoff level of 10% (as defined in Section 3.1) and 6%, respectively. The results are consistent across different cutoff levels.

A.4 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.6 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

Table A.3: 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

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.5 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.7. Binary variables are clearly identified. All others are continuous indices. Details of these are available in the PISA technical report, (OECD (2015b)), 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. We also include country fixed effect in the imputations.

A.6 Descriptive Statistics

Table A.8 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 on average in the 10th grade and half the students 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.[41] 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.

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

A.7 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 we will decompose Y_s into its component parts.

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

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.9.

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.4: Factors Related to Being Non-Serious for Each Criterion (For Cutoff level of 10%)

Variables	non reached	no response	missing	little time
Log(math score)	-0.050***	-0.215***	0.057***	-0.148***
	(0.006)	(0.008)	(0.009)	(0.010)
ESCS	0.003***	0.006***	-0.006***	0.001
	(0.001)	(0.001)	(0.002)	(0.002)
ESCS^2	-0.000	0.001**	0.004***	-0.001
	(0.000)	(0.001)	(0.001)	(0.001)
Grade	-0.001	-0.003***	0.010***	-0.006***
	(0.001)	(0.001)	(0.002)	(0.002)
Female	0.003**	-0.005***	0.021***	-0.018***
	(0.001)	(0.002)	(0.003)	(0.002)
Anxiety	0.002	-0.004***	0.011***	-0.004**
	(0.001)	(0.001)	(0.002)	(0.002)
Ambition	-0.005***	-0.005***	-0.009***	0.003
	(0.001)	(0.001)	(0.002)	(0.002)
Skipping Class/Arriving Late	0.001	0.002***	-0.000	0.002***
	(0.000)	(0.001)	(0.001)	(0.001)
Out-of-school learning (hrs/week)	-0.000**	0.000*	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Time on classes	-0.000	0.000**	-0.001***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Standardized test frequency	-0.001	0.000	0.002	0.002
	(0.001)	(0.001)	(0.002)	(0.002)
Teacher-developed tests frequency	0.003***	0.001	0.008***	-0.002**
	(0.001)	(0.001)	(0.001)	(0.001)
Stakes of Standardized tests	-0.000	0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Stakes of teacher-developed tests	-0.000	-0.000**	0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)
Log (school average science score)	-0.050***	-0.056***	-0.040***	0.068***
	(0.007)	(0.010)	(0.014)	(0.014)
Log per capita GDP	-0.032	0.690***	-4.583***	1.192***
	(0.054)	(0.071)	(0.105)	(0.093)
(Log per capita GDP)^2	0.001	-0.034***	0.217***	-0.057***
	(0.003)	(0.003)	(0.005)	(0.005)
Observations	283,674	283,674	283,674	283,674
R-squared	0.0105	0.0489	0.0836	0.0122

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

Table A.5: Factors Related to Being Non-Serious for Each Criterion (For Cutoff level of 6%)

Variables	non reached	no response	missing	little time
Log(math score)	-0.050***	-0.111***	0.030***	-0.114***
	(0.006)	(0.006)	(0.008)	(0.009)
ESCS	0.003***	0.002***	-0.004***	-0.000
	(0.001)	(0.001)	(0.001)	(0.001)
ESCS^2	-0.000	0.000	0.005***	-0.001
	(0.000)	(0.000)	(0.001)	(0.001)
Grade	-0.001	-0.003***	0.007***	-0.004***
	(0.001)	(0.001)	(0.002)	(0.001)
Female	0.003**	-0.004***	0.012***	-0.015***
	(0.001)	(0.001)	(0.002)	(0.002)
Anxiety	0.002	-0.003***	0.010***	-0.003*
	(0.001)	(0.001)	(0.002)	(0.002)
Ambition	-0.005***	-0.001	-0.009***	0.002
	(0.001)	(0.001)	(0.002)	(0.002)
Skipping Class/Arriving Late	0.001	0.001*	-0.000	0.001**
	(0.000)	(0.000)	(0.001)	(0.001)
Out-of-school learning (hrs/week)	-0.000**	0.000	-0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Time on classes	-0.000	0.000*	-0.001***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Standardized test frequency	-0.001	-0.000	0.003**	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Teacher-developed tests frequency	0.003***	0.000	0.005***	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Stakes of Standardized tests	-0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Stakes of teacher-developed tests	-0.000	-0.000	0.000	-0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
Log (school average science score)	-0.050***	-0.025***	-0.047***	0.042***
	(0.007)	(0.007)	(0.011)	(0.012)
Log per capita GDP	-0.032	0.336***	-3.365***	0.909***
	(0.054)	(0.048)	(0.088)	(0.074)
(Log per capita GDP)^2	0.001	-0.016***	0.159***	-0.043***
	(0.003)	(0.002)	(0.004)	(0.004)
Observations	283,674	283,674	283,674	283,674
R-squared	0.0105	0.0294	0.0720	0.0113

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

Table A.6: Factors affecting Pr(Skip) and Pr(Spend too little time) (Individual Characteristics)

	Skip	Spend too little time
Log (math score)	-0.0790***	-0.0411***
	(0.0221)	(0.0024)
Log per capita GDP	0.5659**	0.2814***
	(0.2389)	(0.0176)
(Log per capita GDP)^2	-0.0279**	-0.0133***
	(0.0118)	(0.0009)
ESCS	0.0002	0.0003
	(0.0040)	(0.0003)
ESCS^2	-0.0003	-0.0001
	(0.0019)	(0.0001)
Grade	-0.0049	-0.0012***
	(0.0040)	(0.0003)
Female	0.0077	-0.0043***
	(0.0061)	(0.0005)
Anxiety	0.0172***	-0.0010**
	(0.0048)	(0.0004)
Ambition	-0.0120**	0.0010**
	(0.0053)	(0.0004)
Skipping class/Arriving late	0.0019	0.0005***
	(0.0019)	(0.0002)
Additional learning time (hrs/week)	0.0000	0.0001***
	(0.0002)	(0.0000)
Time on classes	0.0003	0.0000
	(0.0002)	(0.0000)
Standardized test frequency	0.0033	0.0007**
	(0.0039)	(0.0003)
Teacher-developed tests frequency	0.0010	-0.0006***
	(0.0026)	(0.0002)
Stake of Standardized test	-0.0004	-0.0000
	(0.0005)	(0.0000)
Stake of Teacher-developed tests	-0.0006	-0.0001
	(0.0007)	(0.0001)
School average science score	-0.0911***	0.0114***
	(0.0330)	(0.0028)
Observations	290,271	290,234
R-Squared	0.00236	0.0207

Table A.7: 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
COUNTRY	Country fixed effects

Table A.8: 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

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

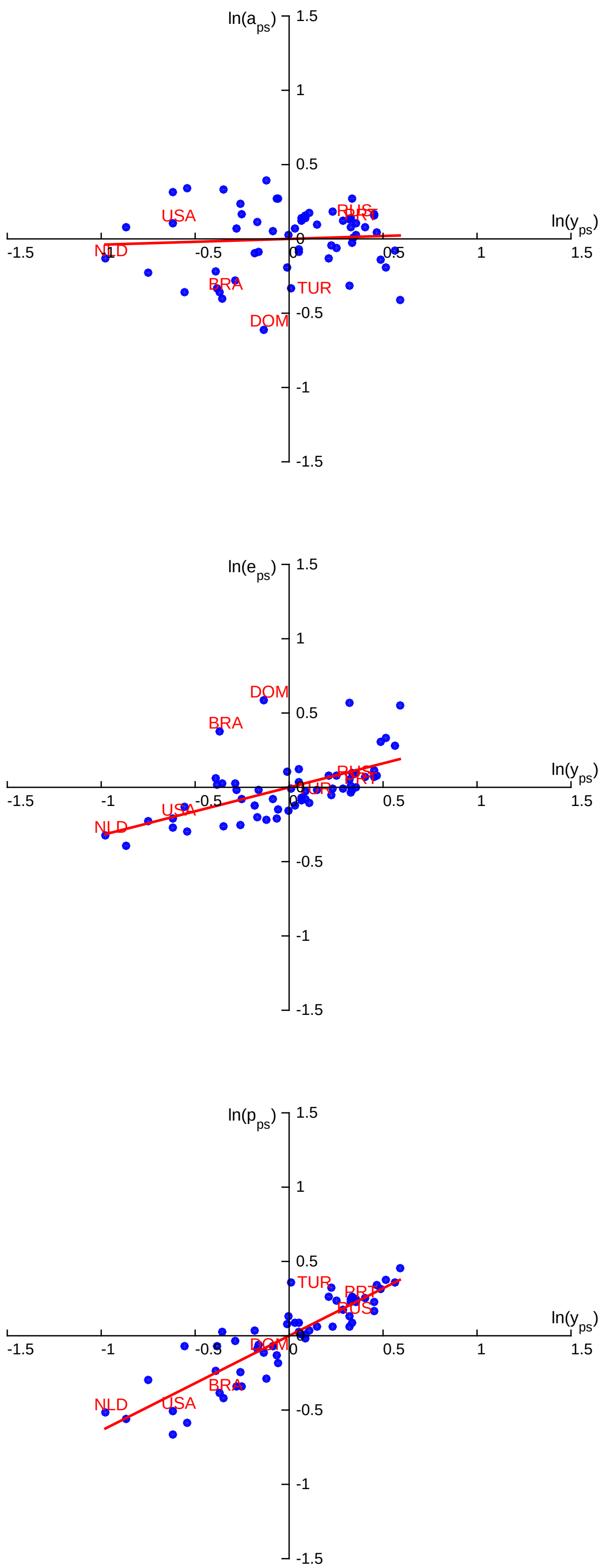


Table A.9: Decomposed Factors for Partially Serious Students

Country	$Yps(\%)$	Aps	Eps	Pps
Tunisia	1.21%	0.19	0.13	0.49
Bulgaria	1.18%	0.27	0.10	0.45
Montenegro	1.12%	0.24	0.10	0.45
Uruguay	1.09%	0.25	0.10	0.43
Croatia	1.07%	0.30	0.08	0.44
France	1.06%	0.34	0.08	0.39
Sweden	1.06%	0.34	0.08	0.37
Czech Republic	1.01%	0.31	0.08	0.40
Luxembourg	0.96%	0.30	0.08	0.39
Portugal	0.96%	0.32	0.08	0.40
Hungary	0.95%	0.29	0.08	0.40
Japan	0.94%	0.38	0.07	0.34
Italy	0.94%	0.28	0.08	0.41
Austria	0.94%	0.31	0.08	0.40
Poland	0.93%	0.33	0.07	0.39
Peru	0.93%	0.21	0.13	0.33
Russian Federation	0.93%	0.33	0.08	0.36
Germany	0.90%	0.32	0.07	0.37
Greece	0.87%	0.27	0.08	0.39
Norway	0.85%	0.34	0.07	0.33
Lithuania	0.84%	0.28	0.07	0.43
Slovak Repubic	0.83%	0.25	0.08	0.41
Switzerland	0.78%	0.32	0.07	0.33
United Kingdom	0.75%	0.34	0.07	0.32
New Zealand	0.73%	0.33	0.07	0.31
Spain	0.73%	0.34	0.07	0.32
Spain (Region)	0.72%	0.33	0.07	0.32
Belgium	0.72%	0.32	0.07	0.32
Israel	0.71%	0.26	0.08	0.32
Iceland	0.71%	0.27	0.08	0.34
Denmark	0.70%	0.31	0.07	0.34
Turkey	0.68%	0.21	0.07	0.45
Slovenia	0.67%	0.29	0.06	0.36
Chile	0.67%	0.24	0.08	0.34
Finland	0.63%	0.37	0.06	0.26
Estonia	0.63%	0.38	0.06	0.27
Australia	0.62%	0.30	0.07	0.29
Hong Kong	0.60%	0.42	0.06	0.23
Dominican Republic	0.59%	0.16	0.13	0.28
Korea	0.57%	0.26	0.07	0.29
Ireland	0.57%	0.32	0.06	0.29
Latvia	0.56%	0.26	0.07	0.32
Chinese Taipei	0.52%	0.34	0.07	0.22
Canada	0.52%	0.36	0.06	0.24
B-S-J-G (China)	0.51%	0.31	0.07	0.22
Costa Rica	0.50%	0.22	0.08	0.30
Macao	0.48%	0.40	0.06	0.21
Qatar	0.47%	0.19	0.08	0.32
Brazil	0.47%	0.20	0.11	0.21
Colombia	0.46%	0.21	0.08	0.29
Mexico	0.45%	0.23	0.08	0.25
Singapore	0.39%	0.40	0.06	0.17
United Arab	0.38%	0.20	0.07	0.29
United States	0.36%	0.32	0.06	0.19
USA (Massachusetts)	0.36%	0.39	0.06	0.16
Thailand	0.32%	0.23	0.06	0.23
USA (North Carolina)	0.28%	0.31	0.05	0.18
Netherlands	0.25%	0.25	0.05	0.19