UNIVERSITY OF ARKANSAS | College of Education & Health Professions
Education Reform

WORKING PAPER SERIES

When Students Don't Care: Reexamining International Differences in Achievement and Non-cognitive Skills

Gema Zamarro, Collin Hitt, and Ildefonso Mendez

June, 2017

EDRE Working Paper 2016-18

char·asse·in
THE CHARACTER ASSESSMENT INITIATIVE

When Students Don't Care: Reexamining International Differences in Achievement and Non-cognitive Skills

Gema Zamarro*

University of Arkansas

Collin Hitt

University of Arkansas & Southern Illinois University

Ildefonso Mendez

University of Murcia

First Version: November, 2015

This Version: June, 2017

*Corresponding author: Gema Zamarro, University of Arkansas, email: gzamarro@uark.edu. We thank conference participants at APPAM 37th Annual Fall Research Conference and at AEFP 41st Annual Conference for very useful feedback on earlier versions of this paper. We also would like to thank Katy Mazz and Lindsay Weixler for relevant discussions of our paper during these conferences. Finally, we also would like to thank Elise Swanson for her help producing the maps presented in this paper..

Abstract

Policy debates in education are often framed by using international test scores, such as the Programme for International Student Assessment (PISA). The obvious presumption is that observed differences in test scores within and across countries reflect differences in cognitive skills and general content knowledge, the things which achievement tests are designed to measure. We challenge this presumption, by studying how much of the within-country and between-country variation in PISA test scores is associated with student effort, rather than true academic content knowledge. Drawing heavily on recent literature, we posit that our measures of student effort are actually proxy measures of relevant non-cognitive skills related to conscientiousness. Completing surveys and tests takes effort and students may actually reveal something about their conscientiousness and diligence by the amount of effort they show during these tasks. Our previous work, and that of others validates this claim (e.g. Boe, May and Boruch, 2002; Borghans and Schils, 2012; Hitt, Trivitt and Cheng, 2016; Hitt, 2016; Zamarro et al., 2016). Using parametrizations of measures of survey and test effort we find that these measures help explain between 32 and 38 percent of the observed variation in test scores across countries, while explaining only a minor share of the observed variation within countries.

Keywords: Non-Cognitive Skills, PISA study, Survey Effort, Test Effort

JEL codes: I20, C80, C83

"U.S. 15-year-olds made no progress on recent international achievement exams and fell further in the rankings, reviving a debate about America's ability to compete in a global economy."

- ***The Wall Street Journal,*** December 3, 2012

"Finland's schools owe their newfound fame primarily to one study: the PISA survey, conducted every three years by the Organization for Economic Co-operation and Development (OECD)."

- ***The Atlantic Monthly,*** December 29, 2011

1. Introduction

Since their introduction, large scale international assessments have been used to make sweeping statements about the quality of countries' schools and the cognitive skills of their students. The Programme for International Student Assessment (PISA), the Third International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRSL), have become important sources, and for some countries the only sources, of information on student performance in key subjects such as math, reading and science.

The tests ostensibly measure student content knowledge or, more generally, cognitive ability. However, in reality, student performance is driven by more than just cognitive ability and content knowledge. Some students put forward less effort than others during exams. This is a commonsense observation. Test scores cannot tell us about the math, reading or science ability of students who don't pay attention or put forward effort while taking low-stakes tests. So, observed differences in test scores within and across countries might reflect in part differences in student effort on the tests, rather than just true academic content knowledge.

In this paper, we contend that it is possible to measure student effort on tests and similar tasks. Doing so allows us to adjust observed test scores for estimated student effort: students who

show low effort on tests probably possess greater math and reading ability than their test scores indicate. We examine student effort and performance on the 2009 wave of PISA.

Survey and test data available from PISA allow us to build a number of possible measures of student effort. For example, the random ordering of questions in different test booklets and the random assignment of booklets to students in PISA is a key feature of this data that we exploit in order to estimate measures of student effort during the performance of the test. We also build measures of student effort based on parametrizations of how much effort they put forward in answering the PISA student survey questionnaire. We argue that tests and questionnaires can be seen as a task and that by studying the effort that students put forward on them we can recover relevant character skills. In particular, we believe they could be used as proxies for character skills related to conscientiousness and diligence.

Thus, our motivation for conducting this study goes beyond the simple question of whether students try hard on tests. We are interested in within-country and cross-country differences in student's non-cognitive skills related to conscientiousness and diligence. Previous research (see e.g. Borghans and Schils, 2012; Hitt, Trivitt and Cheng, 2016; Hitt, 2016) has shown that measures of student effort, based on students' response patterns to surveys and tests, are predictive of important life outcomes, independent of measured cognitive ability. Drawing upon this literature, we argue that our measures of student effort derived from PISA can be understood as meaningful measures of student's non-cognitive skills related to conscientiousness and conscientiousness related skills. Students plausibly tell us something about their character - their conscientiousness, self-control, or persistence - through their effort on tests and surveys. Perhaps they tell us about how they approach the routines of schooling. Tests and surveys are tasks that resemble everyday schoolwork. By measuring effort on those tasks, we not only

correct PISA test scores for student effort, we believe we also identify potential indicators of international differences in relevant non-cognitive skills.

In particular, this paper aims to respond to the following research questions.

1. Does effort on tests and surveys vary across countries?
2. Does varying student effort impact our understanding of cross-country differences in test scores?

The rest of the paper goes as follows: Section 2 presents a simple conceptual framework for understanding the role that effort plays on the assessment of content knowledge and cognitive skills. Section 3 reviews previous relevant literature. Section 4 describes the PISA study and the data used in this paper. Section 5 further explains our proposed measures of student's effort in tests and surveys and describes our approach for obtaining corrected measures of PISA test performance. Section 6 describes our results. Section 7 concludes by highlighting implications of our findings.

2. Theoretical Framework

In order to understand the role that effort plays in the testing and survey process, it is useful to think about the assessment process. Below we briefly outline some of the very basic elements of standardized tests and student surveys.

What is a Standardized Test?

A standardized test is an instrument designed to measure student content knowledge with respect to specific subjects. The PISA 2009, for example, contained questions on reading, math and science. The test is largely multiple choice. Items vary in difficulty. The tests are often long:

the average PISA 2009 booklet contained approximately 60 questions, and was expected to take two hours. And importantly, the tests are low-stakes: student scores on PISA are anonymous and have no effect on the students themselves.

What is a Survey?

A survey is an instrument designed to gather information and opinions from students. It should be easily readable: surveys are typically constructed to be readable even to students whose reading ability is 3 to 5 years below the grade level. Surveys are often long. The PISA 2009 Student Survey contains approximately 170 items, almost all of which are multiple choice. The surveys are confidential and low stakes: the answers that students give have no effect on the students themselves.

How does effort relate to Test Scores?

Figure 1 presents a simple theoretical framework. A well-designed test should measure student cognitive ability or content knowledge. Realistically, we cannot observe actual cognitive ability or true content knowledge, especially not for students participating in PISA. What we observe is how well students perform on a test. Therefore, in order to conclude that test scores accurately measure true cognitive ability or content knowledge, one must assume that nothing moderates (or interferes with) the relationship between true ability and the performance on the test. We know this assumption is untrue.

<<Figure 1 Here>>

Student effort is a moderator between cognitive ability and test scores. Some students, simply put, don't try very hard on tests. This leads to an underestimate of those students' cognitive abilities. Previous literature has modeled student effort as a product of incentives (e.g.

Kautz et al., 2014). And indeed incentives have been shown to alter student performance on tests. But on PISA, as on most standardized tests, the explicit incentives are the same for all students; the 2009 PISA is a low stakes test for students. Therefore, if student effort differs across students taking PISA tests, it does so for a reason other than individual incentives.

This is not to say that students in all parts of the world view PISA tests identically. Some national or regional educational authorities attach great importance to their students' performance on PISA. In Spain, for instance, PISA tests are the sole measure of educational achievement in many Spanish regions. PISA tests serve a similar role to the National Assessment of Educational Progress in the United States. One might expect that, in Spain, the competition between Spanish regions may lead regional authorities to prepare specifically for PISA tests. This instance is likely an exception to the rule and as it will be seen in the results section bellow it appears that is not supported by the PISA data.

We argue that effort on PISA, and on standardized tests more generally, is driven by student's non-cognitive skills. Skills such as conscientiousness, persistence and self-control are, practically by definition, needed to complete long, mundane, low-stakes tasks. In Figure 1, we show a simple conceptual model where student effort is driven by such non-cognitive skills. Effort is a mediator of the relationship between non-cognitive skills and cognitive test scores. Put another way, non-cognitive skills impact test scores because students who possess high levels of self-control or conscientiousness try harder on low-stakes tests.

In making the case that non-cognitive skills drive student effort on tests, we rely largely on recent literature, which we describe in the next section. We cannot actually test this case, however, relying solely on data available from PISA. Just as with cognitive ability, we cannot actually observe the true non-cognitive skills of students participating in PISA. But we *can*

observe test-taking and survey-taking behaviors of students in PISA datasets, and we can point to recent literature that shows that test-taking and survey-taking effort are linked to later life outcomes, independent of test scores. This literature suggests strongly that student effort during assessments is not some behavior that is idiosyncratic to tests and surveys. Instead it suggests that student effort is indicative of relevant non-cognitive traits that have a broader and long-lasting impact.

Again, strictly speaking, our analysis of PISA data is limited to the relationship between our measures of student effort and test performance. So in this narrow sense, we hope to eventually correct PISA scores for student effort, allowing for more valid comparisons of student content knowledge and cognitive ability. But more generally, we believe that our research also produces information that will allow researchers to compare students' non-cognitive skills across and within countries, using effort on PISA as a proxy measure.

3. Survey and Test Effort as a Proxy Measure for Relevant Character Skills

A recent literature has argued that survey and tests can be seen as performance tasks and that parameterizations of survey and test effort can reveal meaningful measures of character skills related to conscientiousness.

Survey effort can be measured by analyzing response patterns within surveys. Hitt, Trivitt & Cheng (2016) studied the validity of item nonresponse rates as a proxy to measure character skills related to conscientiousness. They studied six nationally-representative, longitudinal datasets of American secondary school students and found that non-response patterns of students while in adolescence were predictive of their educational attainment and labor market outcomes later on as adults. This result hold even after controlling for cognitive ability measures available in these datasets. Similarly, in an unpublished working paper, Boe, May and Boruch (2002)

examined the role of student effort on cross-country comparisons using the Third International Mathematics and Science Study (TIMSS). They used item response rates on a student survey given as part of the TIMSS as their measure of student effort or as the authors call it, "student task performance." The authors found that more than 50 percent of the cross-country variation in test scores was explained by survey item nonresponse.

When asked to complete surveys, however, some individuals begin the survey and do not skip items but still exert low effort by hastily providing thoughtless and random answers. This behavior results in careless answering patterns, which is a behavior that can be detected and parameterized in alternative ways. These include the introduction of specific trick questions to measure careless answering or the analysis of response patterns on already collected data (see, Hitt, 2016; Meade & Bartholomew, 2012). In this respect, Hitt (2016) proposes and validates a measure of careless answering based on the study of response patterns to validated scales within the survey. The author used a nationally sample of American adolescent school-age children and found that careless answering patterns were predictive of education and labor outcomes of these students in their adulthood, after controlling for cognitive ability measures. Zamarro et al. (2016) studied the validity of careless answering measures as proxy of relevant character skills in a nationally-representative sample of American adults. The authors in this case found that careless answering patters were related to educational attainment, employment income, a greater likelihood of being employed in a high-skilled job, and self-reported measures of conscientiousness and neuroticism, even after controlling for cognitive ability. Other work in the area of survey research has also found that non response, inconsistent response patterns, and the amount of time respondents spend answering individual items are an indicator of respondent's

conscientiousness and conscientiousness related skills (See, Jensen & Soland, 2016; Roβmann & Gummer, 2016; Segal, 2012).

Concerning effort in tests, Borghans and Schils (2012) used the fact that students participating in PISA and in other studies were randomly given different test booklets, which differ in the order the questions were presented, to quantify the rate of decline in performance in the test as the test progresses, as a measure of student effort. The authors were then able to show that the rate of decline in performance over the course of the test's administration was related to non-cognitive factors such as motivation and ambition and turned to be a good predictor of final levels of educational attainment. On their analysis using PISA 2006, the authors also found that cross-country differences in motivation in the test explained 19 percent of the variance in PISA scores between countries. Similarly, Hernandez & Hershaff (2015) study the incidence of non-response on a statewide standardized test. They find that, conditional on test scores, skipping questions in the test among middle school students is related with important educational outcomes later on in high school and college, such as grade repetition, drop-out behavior, on-time graduation and attending a 4-year college.

Our work builds on this literature. Both measures of effort on the survey and the test might identify different levels of student effort. Those students who are disengaged and exert minimal effort might not show decline in the effort during the PISA test if they start and end the test with low levels of effort. However, these type of disengaged students might be better captured with measures of effort in the survey as they presumably would, most probable, exert low levels of effort in this task either by skipping questions or providing careless answers. Thus, we build measures of student's non-cognitive skills based on the pattern of response to both the tests and surveys that were part of the PISA study in 2009. We expand on the work of Borghans

& Schils (2012) and build measures of character skills related to conscientiousness based on the rate of decline in performance in the test as the test progresses. Following, Hitt, Trivitt, & Cheng (2016), Boe, May and Boruch (2002), and Hitt (2015) we also build proxies of conscientiousness and related skills by measuring the amount of effort students put forward on the survey that accompanies the PISA test. We contribute to the work of Borghans & Schils (2012) and Boe, May and Boruch (2002) by studying a different wave of data in PISA and by bringing together multiple measures of student effort as proxy for students' levels of conscientiousness. We do so with the aim to have a better understanding of what percentage of the observed differences across countries could be attributed to differences in student's effort. The next sections better describe the data used and our approach to construct these measures of student effort.

4. Data

Our data are from publicly available data sets published online by the Organization for Economic Co-operation and Development (OECD), the sponsoring agency of PISA. In particular, we focus on data from 2009. As we have mentioned, our study builds partly on research by Borghans and Schils (2012) that examines earlier waves of PISA.

In 2009, seventy-four countries and regional "economies" participated in PISA. In total, tests were administered to 515,958 students, comprising representative samples within their home countries. The 2009 PISA test was a standardized test of math, reading and science ability.[1]

[1] Scores were calculated separately for each content area, using an IRT framework that produces five "plausible values" for the level of ability of the student in each of the tested areas. The idea behind this approach is that each plausible value takes into account that a student's test score misestimates the student's true ability. Students of differing ability can receive the same raw

Each student took a test of approximately 60 items. Within each country, each participating student was randomly assigned one of several test booklets. Each booklet was comprised of test items drawn from an item bank of several hundred entries. Through 2006, all countries participating in PISA were issued the same set of thirteen booklets. That changed in 2009, according to the PISA 2009 Technical Manual:

> "*In PISA 2009 some countries were offered the option of administering an easier set of booklets. The offer was made to countries that had achieved a mean scale score in reading of 450 or less in PISA 2006, and to new countries that were expected – judging by their results on the PISA 2009 field trial conducted in 2008 – to gain a mean result at a similar level. The purpose of this strategy was to obtain better descriptive information about what students at the lower end of the ability spectrum know, understand and can do as readers. A further reason for including easier items was to make the experience of the test more satisfying for individual students with very low levels of reading proficiency***.**"

Our current analysis is limited to the 44 countries who took the standard, harder set of 13 booklets. Within those countries, we also exclude a relatively small group of students who received a booklet specially developed for schools that primarily serve students with disabilities. Our total sample is 311,484 students.

The PISA testing session lasted two hours.[2] After the survey, students were given an accompanying survey about learning environment, home factors, and student attitudes. The

score. Or, put differently, a student of a given ability could randomly receive any number of scores within a given range, due to error. Therefore, plausible values are, "random numbers drawn from the distribution of scores that could be reasonably assigned to each individual – that is, the marginal posterior distribution," according to the PISA 2009 Technical Manual. For simplicity of computation, however, we only use one of the assigned values for the analysis as collected in the variables PV1MATH, PV1READ and PV1SCIE.

[2] A one-hour test version was developed for schools serving special needs students. Students taking the one-hour test are excluded from our analysis.

surveys were administered immediately after the completion of the test, and were expected to take one hour to complete.

PISA provides each student's full test and survey record. This includes details of each student's response to each question. We use patterns of student item-level responses to build a number of plausible measures of student effort, which we discuss in the following section.

5. Methods

5.1 Measuring Student Effort

We explore three potential measures of student effort in PISA. The first of the measures is derived from student answer patterns on the test form, the other two are derived from answer patterns on the subsequent survey form. These two measures we argue allow us to better capture students who exhibit lower levels of engagement and effort. On the test form, we explore the rate of performance decline over the course of the test. On the survey form, which as mentioned above was administered immediately after the test, we explore: item nonresponse rates and a measure of careless answering patterns.

We will now describe each measure of effort in greater detail. Descriptive statistics for each measure can be found in Table 1.

<<Table 1 Here>>

5.1.1 PISA Test: Declining Effort

Across PISA tests, performance has been found to decline on average as students move from the beginning to end of the test (e.g. Borghans and Schils, 2012). Figure 2 presents the average performance on each question of the test as the test progresses, for a selected group of countries,

using data from PISA 2009. For all countries, performance declines as the test goes on. As can also be seen in this figure, the rate of performance-decline over the course of the test varies across countries. Even in countries with relatively high PISA test scores, such as Finland or the city of Hong Kong, a decline is observed. For some countries the decline in performance can be dramatic. That is the case of Greece as it is observed in Figure 2. Remember that, as question order is randomized across students as part of the PISA test, this observed decline cannot be attributed to the content of the final items on the test. Therefore, this observed decline in performance we believe is related to students' motivation and effort in the test, rather than a difference in question difficulty at the beginning versus the end of the test.

<<Figure 2 Here>>

In 2009, the order and assortment of test questions was randomized across PISA test booklets. Test booklets are then randomly assigned to students. So, across students, a given item varies in its position on the test. Some students begin with difficult questions, some with easier questions. The independence of question difficulty and question ordering allows us to calculate the effect that "order" has on the probability that a student answers a question correctly. Students who show no decline in motivation should have an equal probability of answering a given question correctly regardless of whether it appears at the beginning or the end of the test.

Our measure of test effort expands on the work by Borghans and Schils (2012), who examined data from the 2006 wave of PISA. Within each country, using a linear regression model, they examined the relationship between question position and the probability that it is answered correctly. Their approach generated country-level estimates of the decline in performance over the course of the test. The effects of order vary by country, suggesting motivation varies by country.

In this paper, we expand the methods used by Borghans and Schils (2012). In particular, we also first estimate country-level estimates but using a linear random coefficient model, which we argue would fit better the data than the simple regression model that Borghans and Schils (2012) employed. Our country-level estimates, then are obtained from the following model:

$$y_{ij} = \alpha_0 + \alpha_0^i + \beta_1 O_{ij} + \beta_1^i O_{ij} + \gamma_j + \varepsilon_{ij} \quad (1)$$

Where the dependent variable y_{ij} takes value 1 if the answer of student *i* to question *j* was correct, value 0.5 if they got half credit for that question, and 0 if the answer was wrong. The independent variable of interest O_{ij} is the sequence order of the test question, rescaled such as the first question is numbered as 0 and the last question as 1. The constant α_0 then represents the average performance of students in the very first question on the test. By introducing a random intercept in the model (α_0^i) we allow for different students to deviate from the average performance in the first question. The intercept coefficient β_1 presents the average decline in test performance from the first to the last question of the test. By introducing a random slope in the model (β_1^i) we allow for different students to deviate from the average decline in performance. The introduction of this random intercept (α_0^i) and random slope component (β_1^i) has the advantage of better taking into account the structure of the data and allow for estimations of how individual students differ from the average observed pattern. This is the main difference between our model specification and that of Borghans and Schils (2012) who excluded this components and estimated just an average constant and slope. γ_j are question fixed effects to control for the difficulty level or nature of each question (e.g. Multiple choice or open question).

The model presented in (1) is then estimated for each country separately using Maximum Likelihood methods assuming a normal distribution for the random coefficients and allowing for

the random constant (α_0^i) and random intercept (β_1^i) components to be correlated. This process provides us with estimates of the country average performance in the first question (α_0), country average decline in test performance (β_1), estimated question dummies effects, the standard deviations of individual random effects (α_0^i and β_1^i), and the estimated correlations among them.

Secondly, we obtain student-level estimates of the rate of decline in the course of the test by comparing each student's performance at the end of the test with his/her performance at the beginning. Table 1 shows the average number of items correct on the first ten and last ten items of the test. Across the sample, average performance declines from 5.85 items correct on the first ten items to 4.46 items correct on the final ten items. That said, for some students the decline in performance may reflect the fact that their booklets randomly contained relatively difficult items toward the end of the exam. Indeed, ANOVA estimates find that 16.3 percent of the variation in decline from the first ten to last ten items is explained by booklet number. In order to account for the effects of booklet on decline in student performance, we simply estimate decline in performance, regression-adjusted for booklet number. Per Table 1, the adjusted rate of decline in performance from the first ten to the last ten items is 1.37 points.

Our approach for computing student-level rates of decline in the course of the test might seem simplistic. By averaging performance over the first ten and last ten items, and then comparing those averages to one another, we are assuming that the rate of decline is fairly steady across the course of the test. If, for instance, student performance actually drops within the first ten items and remains steady thereafter, this measure will fail to fully capture decline in performance. The plots shown in Figure 2 do not point to such a pattern and the rate of decline in performance on average appears to take place steadily over the course of the test. Moreover, the country-level average decline of performance using this simple method lead to similar values

than the country-level estimates obtained from the random coefficient model described in (1). The mean value of the estimated country-level decline parameter (β_1) estimated from equation (1) is -0.122 (which can be found in Table 4.B, discussed in greater detail later in the text). That is, at the country level, the average estimated effect of moving an item from being the very first question on the test to being the very last item on test would be a 12.2 percentage point decline in the probability of answering the item correctly. This estimate is similar to the average value we obtain based on our student-level estimates based on our simple method described above. In this case we obtain that sudents on average present a 13.7 percentage point decrease in the average probability of answering correctly from the first ten to the last ten items on the test.

5.1.2 PISA Survey: Survey Effort Measures

Item Nonresponse: The item nonresponse rate on a survey is often defined as the rate at which students skip questions, or answer "I don't know." For decades survey methods researchers have seen survey item nonresponse as a measure of disengagement in the survey process, presuming of course the survey is well designed.

In PISA surveys, "I don't know" is virtually never offered as an answer choice. Survey item nonresponse rates are then measured as the rate at which students skip questions. Per Table 1, the average survey item nonresponse rate is 3 percent. The standard deviation of survey item nonresponse is 5 percent within country and 1 percent between countries.

Careless Answering Patterns: Commonsense intuition tells us that some students might not skip questions at all, but instead just fill in the "bubbles". We attempt to identify this type of behavior. We term "careless" answers as a series of answers on the student survey that appear inconsistent with one another.

We use a novel method developed by Hitt (2016), in order to distinguish between legitimate answers and answers that appear to have been entered carelessly. We exploit the fact that a large number of items on the PISA Student Survey are part of larger multi-item scales that use a Likert-type response format. For example, as part of a scale to assess "attitude toward school" students are asked the extent to which they agree with a number of statements. The first item is, "School has done little to prepare me for adult life when I leave school." A subsequent item is, "School has taught me things which could be useful in a job." A priori, one would think, students who agree with the first statement should be unlikely to agree with the later statement.

When inquiring about the "attitude toward school" or some other concept, survey administrators ask multiple, similar questions for a simple reason. Asking multiple, simple questions about a related concept yields more reliable information than asking only a single question. In a well-constructed scale, answers to each of the questions should be reasonably well correlated with one another. If they weren't, one could hardly argue that the questions were actually measuring the same concept. Standard psychometric tests such as Cronbach's alpha and item-rest correlations are often used to report whether items within a scale are in fact correlated.

In a scale deemed consistent and reliable, in psychometric terms, item-answers within a given scale are correlated with one another. That is to say, answers to any given item ought to be predicted reasonably well by answers to the other items on the scale. We examine the frequency with which students give answers that appear inconsistent, or more specifically, unpredictable, given their answers on the other related questions that are part of the scale.

Following Hitt (2016), we conduct a separate bivariate regression for every Likert-type item on the PISA Student Survey. In total, we examine 84 items across 12 scales. Every item is regressed on the average of answers given to the remaining items on the same scale. For

example, student responses to the first item of the "attitude towards school" scale are regressed on average score of the remaining items of that same scale. That is, we follow this type of "item-rest" bivariate regression equation, adapted from Hitt (2016).

$$Y_{ijs} = \beta_0 + \beta_1 \overline{Y}_{is,-j} + \eta_{ijs} \quad (2)$$

Where Y_{ijs} is the answer to item *j* within scale *s* provided by student *i*. The coefficient of interest is β_1 which is the coefficient of the variable $\overline{Y}_{is,-j}$, the average of the rest of the items in scale *s* (all items not *j),* by student *i.* β_0 is a constant and η_{jst} is an error term. These bivariate regressions are mathematically equivalent to the item-rest correlations used in psychometric evaluations of scales (Hitt 2016).[3]

We store the estimated student-level residuals η_{jst} from each regression. Each residual literally measures the extent to which a given student gave an unpredictable answer, as judged by the regression model (which is based on the answer patterns of all students) and that student's answers to other items on the scale.

We then standardize the absolute value of each residual, with a mean of zero and a standard deviation of one. The average of these standardized scores, obtained from responses to each item of the 12 identified scales, is then combined into a composite "careless answering" score. Displayed in Table 1, the unit of change for the careless answering score does not have a conversational interpretation. A lower score signifies that on average a student's individual answers were well predicted by his/her other answers. A higher score signifies that the student

[3] When the regression is standardized, the constant term β_0 drops out and the slope coefficient β_1 becomes mathematically identical to the coefficient of a Pearson product-moment correlation.

consistently gave answers that did not appear consistent. The mean careless answering score is zero, with a standard deviation of 0.24 within country and 0.07 between countries.

5.2 Studying the Role of Effort in International Comparisons of Student's Performance

To study the degree the international variation in student effort on PISA helps explain international differences in PISA test scores we estimate a random-effect multilevel analysis following this model:

$$y_{it} = \gamma_1 E_i + \alpha_c + \varepsilon_{it} \quad (3)$$

Where y_{it} is the PISA score for student *i* on test *t*, E_i is the array of measures of student effort described in previous section, α_c is a country level random effect and ε_{it} is the error term assumed to be normally distributed. Estimates of the model described in (3) allow us to study the relationship between effort and test scores: across the overall sample, within country, and across countries.

Using the estimates from the model described in (3) above, we then compute PISA scores adjusted for student effort adding the estimated residuals from these regressions and the country specific effects.

As an alternative approach, one could follow Borghans and Schils (2012) and simply use the estimated country average performance in the very first question in the PISA test estimated from our random coefficient model specification in (2) as a measure of performance purged of decline in test performance effects. However, although one could argue that this measure is not affected by fatigue in the test, it can be affected by different rates of nonresponse or other measures of test effort. Some students show low effort throughout the test, from the very onset. We have argued that our survey-based measures

help identify such students. Therefore we prefer the approach outlined in (3) above, which takes into account all of the information we've collected on student effort.

6. Results

This section presents the results obtained following the methods described above. We start with a descriptive analysis of our measures of student effort in PISA before answering our two research questions: 1) Does effort on tests and surveys vary across countries? And 2) Does varying student effort impact our understanding of cross-country differences in test scores?

6.1 Descriptive Analysis of the Proposed Measures of Student Effort

In the previous section we laid out a number of plausible measures of student effort on tests and surveys. We now examine the extent to which they are related to student test scores, and the extent to which these variables are related to one another.

Table 2A displays pairwise student-level correlations between PISA 2009 math, reading and science scores and our measures of student effort. All correlation coefficients shown are statistically significant ($p<0.01$). Each of our effort variables are constructed such that a higher value signifies lower effort (i.e. higher detrimental behavior). All correlations are negative, as expected. Of the survey-based effort measures, item nonresponse and careless answering patterns are all negatively related to test scores. On the PISA math score, the correlations are -0.27 and -0.08, respectively. The magnitudes on reading and science tests are similar, an interesting fact we will discuss momentarily. The rate of performance decline is also negatively related to total score, an unsurprising fact. The correlation coefficient is -0.09.

<<Tables 2A and 2B Here>>

As stated above, the pattern of results is noticeably similar across test subjects. One critique of our measures of student effort is that cognitive ability could be the real driver of student engagement on tests. If this was true, one would expect that reading ability above all else would be a driver of nonresponse and careless answers. Students who cannot read at all cannot read surveys and tests. And yet the correlations of our effort variables are hardly higher with reading than with other topics. This finding indicates that student effort impacts each test score similarly - something that would be true if our measures captured student effort, and likely would not be true if our measures were driven by reading limitations.

The correlations between our effort measures are also interesting. Neither survey item nonresponse nor careless-answering is strongly correlated with decline-in-performance on the test. Again, the variable "decline" is based on comparisons of performance at the beginning of the test versus the end of the test. The survey is administered immediately after the test. One might argue that cognitive fatigue causes students to decline in performance after the test. If this was the case, then one would expect fatigue to impact student effort on the subsequent hour-long survey - and therefore one would expect "decline" to be correlated with survey item nonresponse and careless answering. Yet the results tell a different story. Decline in performance on the test is very weakly related to survey item nonresponse and careless answering. This is consistent with the notion that survey effort in fact signals a lack of effort throughout the entire assessment process. "Decline" captures diminished effort over the course of the test. Some students, however, never display much effort in the first place - their performance starts off low, stays low, and they show little effort on the survey. This type of students would not be well capture by measures of decline in test performance as they would show little decline. However, a measure

that identifies such students would not be correlated with decline, but would be correlated with overall test score. That seems to be the case for our survey-based measures of effort.

Careless answering patterns are not strongly correlated with survey item nonresponse. This again is unsurprising. Within a given question, giving a careless answer and not responding at all are mutually exclusive options. Over the course of the assessment, it's possible that different students take different approaches. Some just skip questions frequently, while others complete every question but do so with little care – few seem to switch back and forth between skipping items and answering carelessly.

While the student-level correlations between effort measures are weak, the correlations at the country level are much stronger. These country level correlations are presented in Table 2B. While students who decline in performance are not the same students who skip items or who give careless answers, such students are concentrated together within countries. We delve further into the country-level concentrations of student effort in the following section.

<<Tables 3A, 3B and 3C Here>>

Tables 3A, 3B and 3C are regression estimates, where student-level PISA test scores are regressed on each of our effort measures. All results are standardized, and significant, at $p<0.01$. The first three columns are standardized bivariate regressions, with a single regressor. The coefficients across the first three columns are identical to the corresponding correlation coefficients in Table 2A. Of primary interest in Tables 3A, 3B and 3C, is the estimated R-squared. No individual measure of effort explains more than a minor share of the individual variation in PISA test scores. Nevertheless, when used in combination, our measures of effort explain a substantially greater share of the overall variance than any standalone variable. The

fourth column of Tables 3A, 3B and 3C contains all measures of student effort in a single regression. For math, reading and science scores, respectively, 9.6, 10.9 and 10.3 percent of the respective student-level variation in PISA scores is explained by our measures of effort.

6.2 International Comparisons of Student Effort Measures

We now turn to our first research question: Does effort on tests and surveys vary across countries? As shown in the descriptive statistics of the previous section, the variance between countries is smaller than the variance within countries. Nonetheless, there is a measurable between-country difference in each effort measure. To test the significance of the between-country variance, we conducted a one-way ANOVA of each effort measure, with country as the independent grouping variable. The model F-statistic is statistically significant in every case, with between 1 and 7 percent of the overall variation explained by country dummy variables.

<<Tables 4A and 4B Here>>

<<Figure 3 Here>>

We'll now focus at some length on decline in performance over the course of the test. Tables 4A and 4B present our estimates of country-level decline in performance during the course of the test obtained from the random coefficient model described in (1). Figure 3 displays the estimates of these regressions for a selected group of countries.

As can be seen in this table and figure, and as it was anticipated in the descriptive averages presented in Figure 2, we observe a considerable amount of heterogeneity across countries not only in initial performance in the test but also on our country average estimate of the rate of decline in performance as the test progresses. Some high performing countries like Finland start at a high performance level and remain at a higher level as the test progresses.

Other high performing countries like South Korea do not start at especially high levels in the response of the very first question of the test but present relatively low rates of decline as the test progresses, which makes them end up at a very good final position in performance by the end.

Interestingly, there are countries like Spain that have a performance on the first question of the test that is above average and at the level of the high performing country of South Korea. However, Spain's higher rate of decline in performance as the test progresses quickly drags down their cumulative scores. Decline in performance in the test is especially dramatic for the case of Greece, the country in our sample that presented the highest estimated rate of decline. It is important to stress here that this was also the country that presented the highest rate of decline in performance in PISA 2006 according to the estimates presented in Borghans and Schils (2012). This is reassuring as it suggests that our estimates of the country-level average rates of decline on test performance are capturing permanent country-specific noncognitive skills and are not the result of just one specific year of the PISA study. Finally, we also observe countries like the U.S that start at relatively lower levels of performance but that, given their relatively higher effort in the course of the test, finish in a middle position.

It should also be stressed that decline in performance in the test is not that highly correlated with initial performance in the test. Across countries, the correlation coefficient between country-level initial performance and rate of decline is about 0.2. Figure 4 plots the rate of decline as a function of initial performance. As it can be seen in this figure, there are countries like Hungary that start the test at relatively lower levels of performance but present higher levels of effort during the test and thus decline less during the course of the test. On the other side of the spectrum, there are high performing countries like Russia or Japan that start at a very high level of performance but decline relatively more than other high performing countries.

<<Figure 4 Here>>

In addition, the last column of Table 4A presents the estimated correlation between the individual specific random intercept and random slope components of the model. It is interesting to observe that, although overall the correlation seems to be small if we obtain the average for all countries together, these estimated correlations vary substantially across countries. This indicates that in countries with lower estimated correlations (e.g. Korea, Japan, The Netherlands or New Zealand) both high performing and low performing students present rates of decline in test performance that are similar. In other countries we observe bigger negative correlations (see e.g. Russia, Greece, Indonesia, Thailand) indicating that higher performing students present much higher rates of decline in test performance than lower performing students. This result goes in line with the notion that lower performing students stay disengaged from the beginning of the test to the end.

Finally, we also observe cross-country variation on the levels of survey effort. Figure 5 presents in a map country-average levels of item non-response and careless answering. As we can see in this figure countries differ in their degree of each of these behaviors. Some countries like the U.S. present relatively higher levels of item non-response but relatively lower levels of careless answering while others like Spain present relatively higher levels of both item non-response and careless answering. In contrast, the city of Hong Kong in China, presents relatively low levels of both behaviors.

<<Figure 5 Here>>

6.3 International Comparisons of Student's Performance Accounting for Effort

Tables 5A, 5B and 5C display the estimates of the within-country, between-country and overall variance in PISA scores explained by our measures of effort, following the random effect multi-level analysis model described in (3).

<<Tables 5A, 5B and 5C Here>>

Within the top three rows are estimates, by column, of the proportion of the variance of PISA scores in a given subject area explained by each measure of effort. The overall R-squared in each model corresponds with the R-squared numbers in Tables 3A, 3B and 3C. As discussed, for every variable, the overall R-squared is modest. However, the between-country estimates of our multi-level model tell a very different story.

Altogether, our measures of effort explain a substantial portion of the variation in between-country test scores. The first column displays results for decline in performance: whereas only 0.5 percent of the variation in student-level test scores is explained by decline in performance, 24.8 percent of the between country variation in math test scores is explained by decline-in-performance. Decline-in-performance explains 18.7 percent of the international variation in reading scores, and 27.5 percent of international variation in science scores. These estimates are in line with those presented by Borghans and Schils (2012) in their analysis of PISA 2006.

Survey item nonresponse is an even stronger predictor of international variation in test scores. In standalone models, survey item nonresponse explains 41.3, 33.0 and 37.8 percent of the international variation in PISA math, reading and science scores, respectively. These estimates are largely consistent with the findings of Boe, May and Baruch (2002), who examined TIMSS scores and found that 53 percent of the international variation in math scores was

attributable to item response rates on a corresponding survey. This results provides some evidence to the idea that there are students who are disengaged in the whole testing and survey process and that different shares of such students help explain a significant part of the cross-country variation in PISA test scores.

Careless answers on the survey are by far the weakest predictor of test scores, within and across countries. In all subjects, careless answering explains only about 2 percent of the international variation in test scores, among the countries in our analytical sample.

The final column in tables 5A, 5B and 5C displays estimates when all measures of student effort are included in the random effects model. Of the between country variation in PISA test scores, our combined measures of student effort on the test explain 38.5 percent of the variation in math, 32.4 percent of the variation in reading, 37.3 percent of the variation in science.

Given the popular use of PISA scores to make international comparisons, it is useful to examine how the international distribution of test scores changes after adjusting for student effort. Table 6A displays the summary statistics of the raw and adjusted scores at the student level while Table 6B shows the results aggregated at the country level, both obtained using our simple approach for adjustments based on the estimates presented in Tables 5A, 5B, and 5C and described in section 5.2 above. Both at the student and country level, the overall distribution of test scores tightens. At the student level, the standard deviation in math scores, for example, shrinks from 97 in the raw scores to 92.2 in the adjusted scores. As shown in these simple descriptive statistics, the gap between the highest and lowest performing countries in our sample is driven partly by student effort. Looking at the country-level statistics we observe that the range

of PISA scores shrinks, after adjustments for student effort, from 227.3 to 203 in Math, from 153.6 to 128.7 in Reading and from 192.1 to 167.4 in Science.

<<Tables 6A and 6B Here>>

7. Conclusion and Discussion

We have examined measures of student effort on the PISA study. We have shown that these measures differ by country, and have shown that the distribution of international test scores can change substantially once adjusting the effort that students put forward. However, the information contained in PISA datasets does not allow us to directly test one final question: does effort on tests and surveys provide a proxy measure of student noncognitive skills?

Using only data available from PISA, we can only posit that these effort-based measures of effort are proxies for noncognitive skills, such as conscientiousness and persistence. However, previous research provides compelling evidence that our measures of effort actually capture student noncognitive skills.

Beyond their analysis of PISA scores, Borghans and Schils (2012) also examined student motivation on tests that were administered as part of a longitudinal study of British youth. At the baseline year, when respondents were 16 years old, a math test was given that had similar psychometric properties to PISA. Borghans and Schils (2012) found that the estimated decline in performance on this test was predictive of later labor market outcomes, including employment and wages, independently of final scores on the test.

The fact that decline in performance contains independent information that is predictive of objective measures of well-being shows that student motivation on tests is not idiosyncratic to the testing session. It suggests strongly that decline in performance captures noncognitive skills.

Similarly, recent research examines whether item nonresponse is a proxy measure for noncognitive skills such as conscientiousness (e.g. Hedengren and Strattman, 2012). The most robust examination of survey item nonresponse as an indicator of noncognitive skills can be found in Hitt, Trivitt and Cheng (2016). Within six longitudinal surveys of adolescents from the United States, the frequency with which students skip questions or answer "don't know" is found to be predictive of later educational attainment or labor market outcomes, independent of controls for cognitive ability. The fact that, after adjusting for cognitive ability, survey item nonresponse rates are still associated with later outcomes suggests strongly that item nonresponse is tied to relevant noncognitive abilities. Careless answering, similar to item nonresponse, has been validated as a proxy measure of noncognitive skills in the literature (see e.g. Hitt, 2016; Zamarro et al., 2016).

Overall, this research combined with our findings suggests strongly that international differences in test scores are driven partly by international differences in noncognitive skills. Our analysis produces country-level estimates of student effort and persistence, separate from adjusted PISA scores. We are hopeful that the effort-based measures we develop can be used in future research of noncognitive skills. There is growing interest in international comparisons of noncognitive skills. Our results can be used to inform this research. Noncognitive skills research relies heavily on student self-reported scales. These scale scores provide valuable but imperfect information about student noncognitive skills; self-reported scales are prone to a number of biases, and of course are affected by differences in student effort on surveys.

Our results suggest that standardized test scores reflect more than student learning, they reflect the character traits of students taking the tests. As designed, test scores provide valuable but imperfect information on student cognitive abilities. But testing data can also contain

information about the effort that each student put forward on the test. As researchers seek to examine international differences in noncognitive skills, they may be able to exploit the measures of effort we have laid out here.

In summary, we calculated international differences in test-effort and survey-effort, using innovative measures, which we argue proxy as measures of noncognitive skills related to conscientiousness. Adding survey effort measures to test decline measures allow us to better capture student effort as those students completely disengaged from the whole testing process would show little or no decline in the test if they do not try at all but can be captured by lower levels of effort in the survey. We then decompose international differences in test scores based upon our novel measures of noncognitive skills, finding that between 32 and 38 percent of the between country variation in PISA scores is driven by our measures of effort.

The policy implications of international and regional gaps in test scores are based in large part on what test scores are seen to represent. Our work examines the extent to which these differences in test scores are really driven by differences in math, science and literacy skills, rather than by differences of another sort – differences in how students approach the routine tasks of school and work. Our analysis expands on methods from previous research and applies them to a new sample of students, those participating in the 2009 wave of PISA. Our findings are remarkably consistent with those by Borghans and Schils (2012) and Boe, May and Baruch (2002), who used a different wave of PISA or a different dataset. A substantial portion of the international variation in test scores is driven by student effort on the test itself.

References

Boe, E. E., May H., and Boruch R.F. (2002). Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Acheievement Differences at the National, Classroom, and Student Levels. *Center for* Research and Evaluation in Social Policy, CRESP-RR-2002-TIMSS1. http://eric.ed.gov/?id=ED478493.

Borghans, L., and Schils T. (2012). The Leaning Tower of Pisa. Working Paper. Accessed February 24, 2016. http://www.sole-jole.org/13260.pdf.

Hedengren, D., and Stratmann T. (2012). The Dog That Didn't Bark: What Item Nonresponse Shows about Cognitive and Non-Cognitive Ability. Available at SSRN 2194373. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2194373.

Hernandez, M., and Hershaff J. (2015). Skipping Questions in School Exams: The Role of Non-Cognitive Skills on Educational Outcomes. Mimeo. Accessed October 14, 2016. http://www.edpolicy.umich.edu/files/wp-hernandez-hershaff-skipping-questions-dec-2015.pdf.

Hitt, C. E. (2016). Just Filling in the Bubbles: Using Careless Answers Patterns on Surveys as a Proxy Measure of Noncognitive Skills, EDRE Working Paper 2015-6. Fayetteville, AR: Department of Education Reform, University of Arkansas

Hitt, C.E., Trivitt, J.R., Cheng, A. (2016). When You Say Nothing at All: The Predictive Power of Student Effort on Surveys. *Economics of Education Review*, 52, 105-119.

Jensen, N., & Soland, J. (2016). Understanding the Impact of Student Test Effort on Teacher-Value Added Estimates. Paper presented at the Association for Education Finance and Policy 41st Annual Conference, March 17-19. Denver, CO.

Kautz, T., Heckman J.J., Diris R., Weel B.T., and Borghans L. (2014). Fostering and Measuring Skills: Improving Cognitive and Non-Cognitive Skills to Promote Lifetime Success. National Bureau of Economic Research Working Paper 20749.

Roβmann, J. and Gummer T. (2016). Using Paradata to Predict and Correct for Panel Attrition. *Social Science Computer Review*, 34(3), 312-332.

Segal, C. (2012). Working When No One is Watching: Motivation, Test Scores, and Economic Success. *Management Science*, 58(8), 1438-1457.

Zamarro, G., Cheng, A., Shakeel, M., & Hitt, C. (2016). Comparing and Validating Measures of Character Skills: Findings From a Nationally Representative Sample. EDRE Working Paper 2016-08. Fayetteville, AR: Department of Education Reform, University of Arkansas.

Figure 1: Theoretical Framework

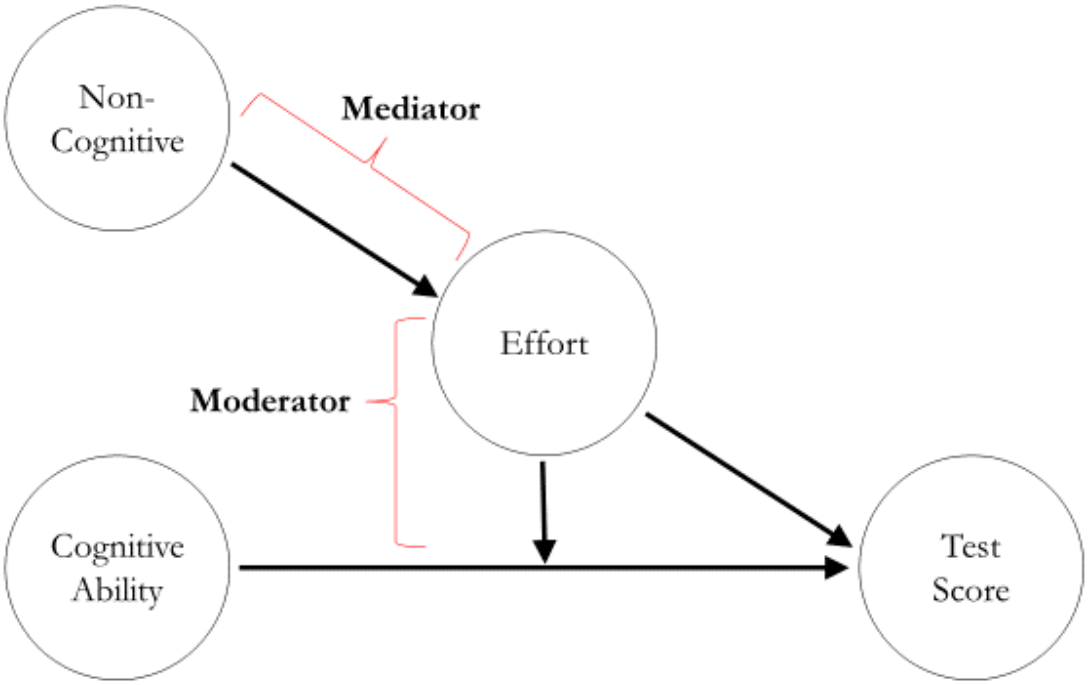


Figure 2: Average Performance by Question Position in Selected Countries

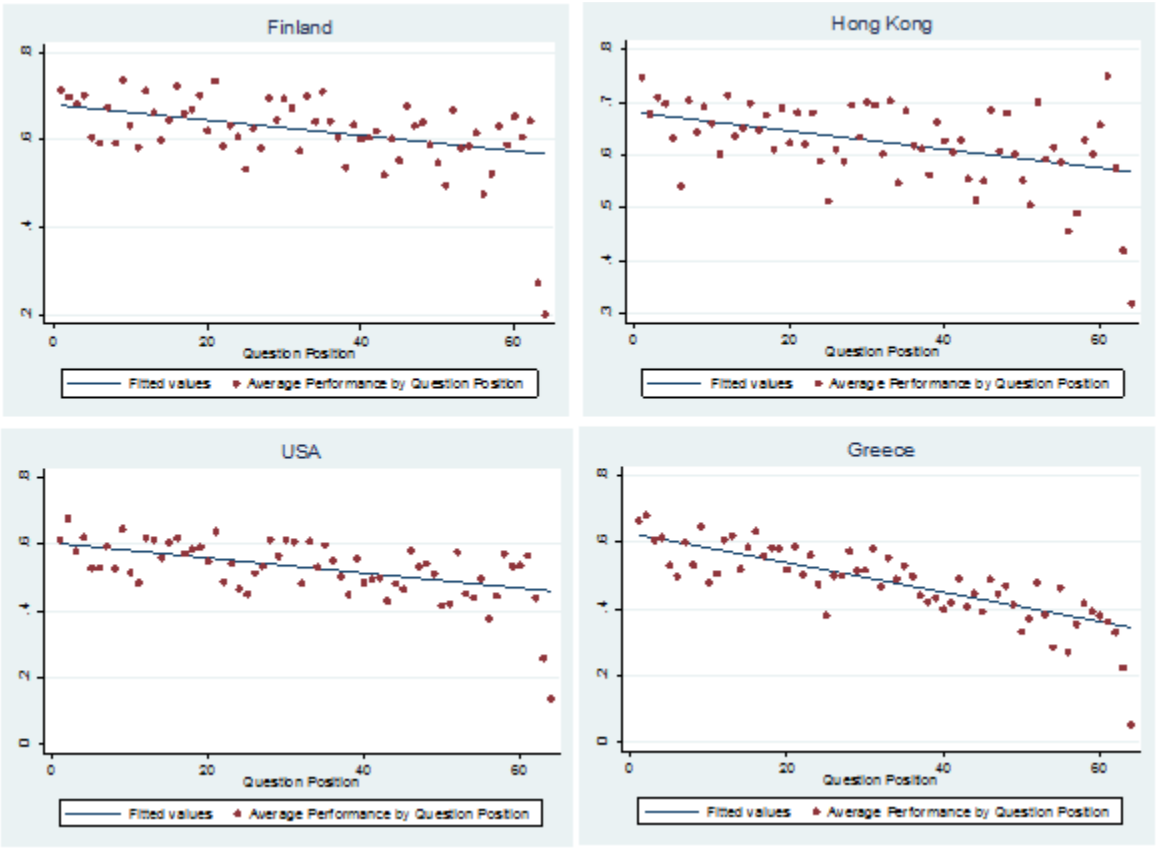
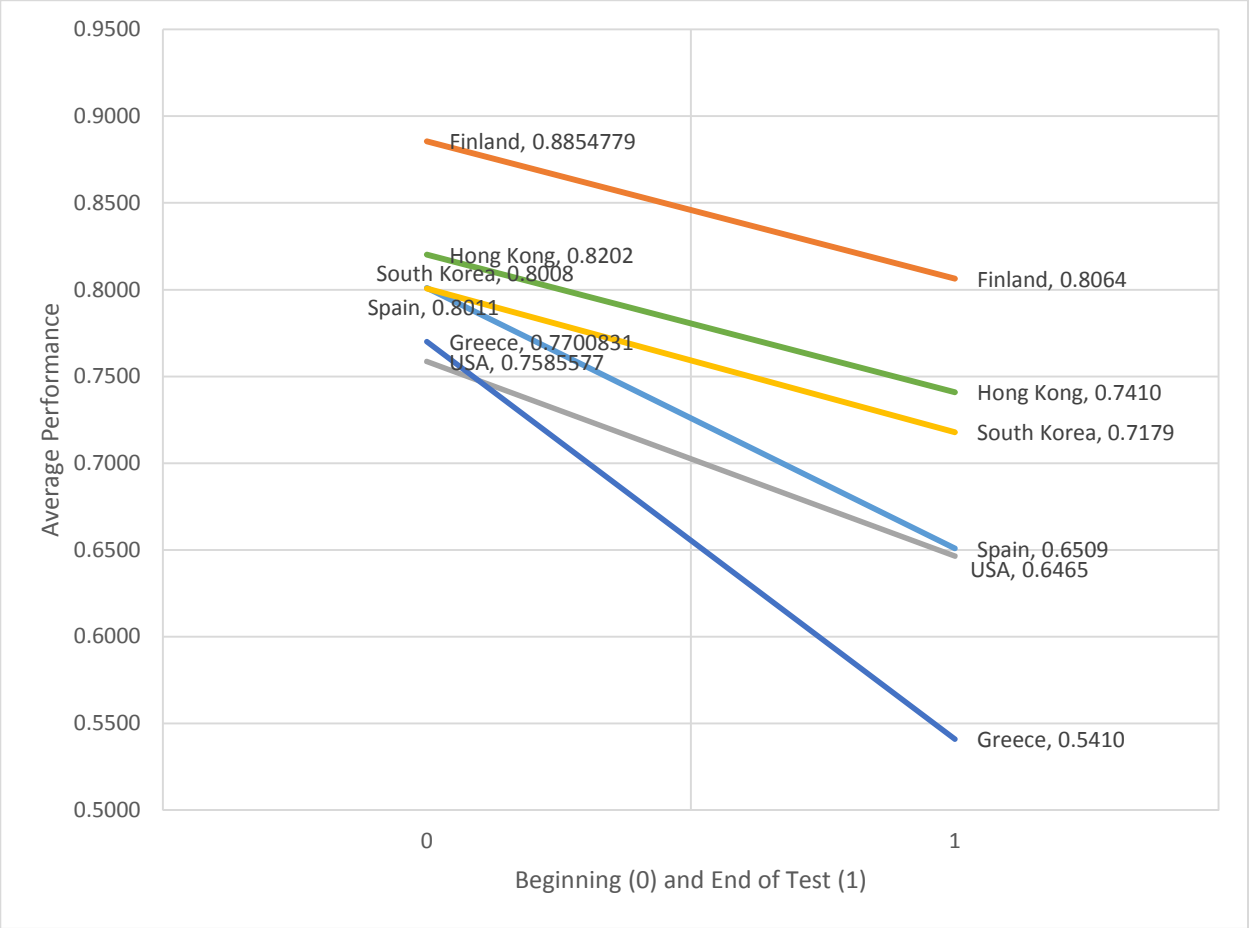
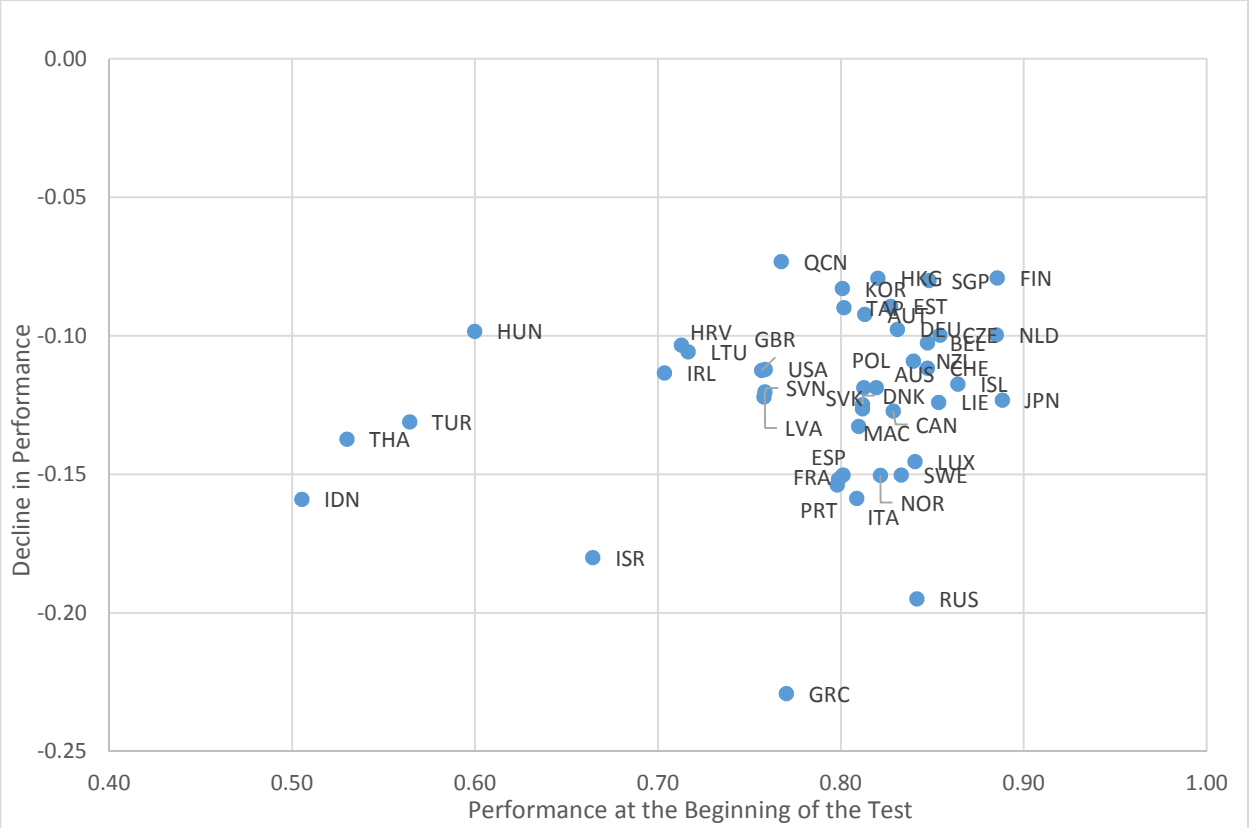


Figure 3: The Estimated Decline in Performance during the PISA test, by Country



Note: Estimates above were obtained using random coefficient regression estimates by country, including a random constant and slope. The graph shows estimated performance at the beginning and at the end of the test. Details of this model are explained in Section 5.

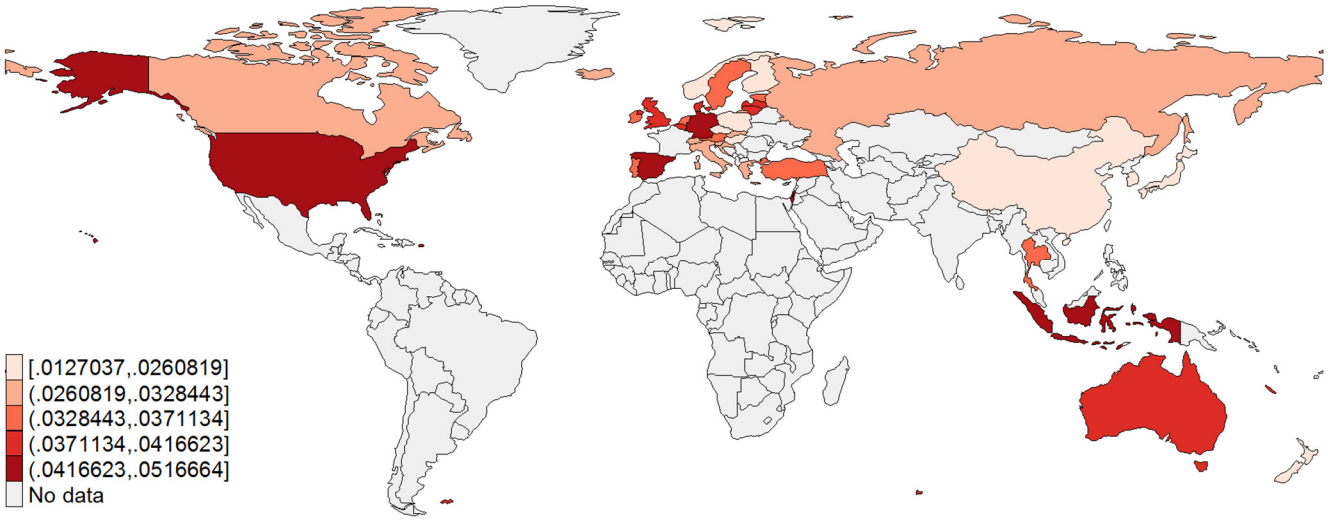
Figure 4: Rate of Decline in Test Performance as a Function of Initial Performance



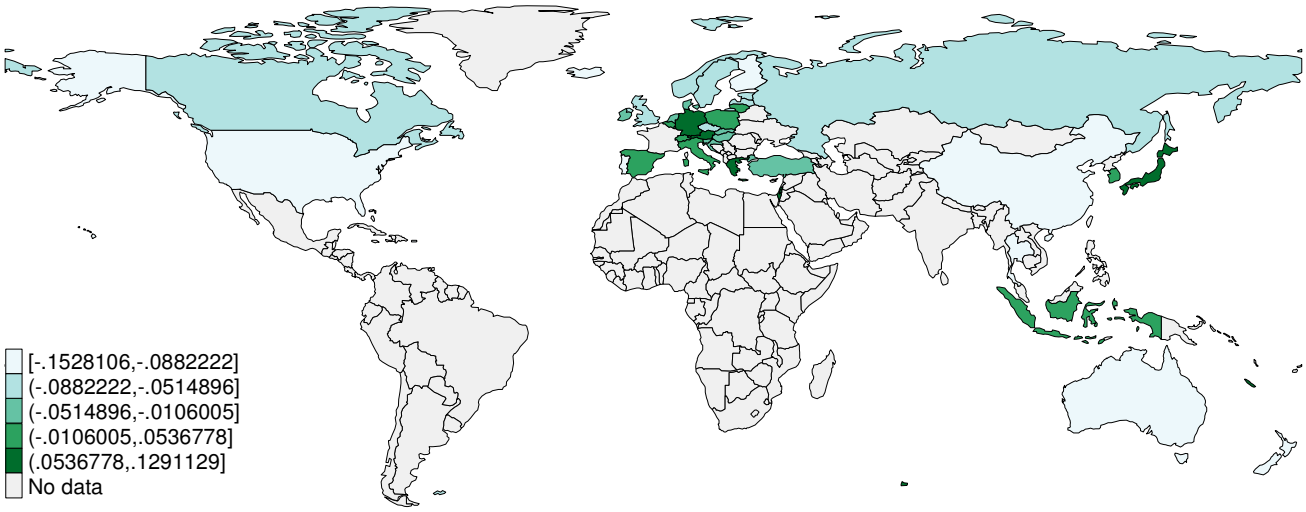
Note: Estimates above were obtained using random coefficient regression estimates by country, including a random constant and slope. Details of this model are explained in Section 5.

Figure 5: Survey Effort Measures in the PISA Survey, by Country

Item Non-Response



Careless Answering



Note: Item non-response rates and careless answering measures were obtained at the student level as described in Section 5.1.2 and then averaged at the country level. Data from Hong Kong was used to represent China in the maps.

Table 1: Measures of Student Motivation during PISA Assessment, Summary Statistics

Variable		Mean	SD	Min	Max	Observations
Test: First Ten Score	overall	5.94	2.47	0.00	10.00	N = 311,484
	between		0.62	3.56	7.21	n = 44
						T-bar =
	within		2.41	-1.26	12.39	7,079.18
Test: Last Ten Score	overall	4.48	2.73	0.00	10.00	N = 311,484
	between		0.75	2.08	6.27	n = 44
						T-bar =
	within		2.64	-1.78	11.51	7,079.18
Test: Adjusted "Decline"	overall	1.37	2.37	-11.04	10.30	N = 311,484
	between		0.27	-0.52	0.84	n = 44
						T-bar =
	within		2.36	-10.63	10.47	7,079.18
Survey: Item Nonresponse	overall	0.03	0.05	0.00	0.95	N = 311,484
	between		0.01	0.01	0.05	n = 44
						T-bar =
	within		0.05	-0.02	0.96	7,079.18
Survey: Careless Answers	overall	-0.03	0.25	-1.04	5.19	N = 309,425
	between		0.07	-0.15	0.13	n = 44
						T-bar =
	within		0.24	-1.04	5.16	7,032.39

Table 2A: Student-level Correlations Between Test Scores and Measures of Motivation

	1	**2**	**3**	**4**	**5**	**6**
1. Math Score	1.00					
2. Reading Score	0.82	1.00				
3. Science Score	0.88	0.88	1.00			
4. Test: Decline	-0.07	-0.10	-0.09	1.00		
5. Survey: Items Missing	-0.27	-0.29	-0.28	0.03	1.00	
6. Survey: Careless Answers	-0.08	-0.08	-0.08	-0.02	0.06	1.00

Note: All coefficients significant at $p<0.001$

Table 2B: Country-level Correlations Between Test Scores and Measures of Motivation

	1	**2**	**3**	**4**	**5**	**6**
1. Math Score	1.00					
2. Reading Score	0.90	1.00				
3. Science Score	0.95	0.93	1.00			
4. Test: Decline	-0.50	-0.43	-0.52	1.00		
5. Survey: Items Missing	-0.64	-0.57	-0.62	0.36	1.00	
6. Survey: Careless Answers	-0.14	-0.14	-0.14	0.24	0.25	1.00

Note: All coefficients significant at $p<0.01$, except for all coefficients in Row 6, which are not statistically significant at $p<0.10$

Table 3A: Regression of PISA Math Score on Effort

Test: Decline	-0.071			-0.063
Survey: Items Missing		-0.274		-0.290
Survey: Careless Answers			-0.085	-0.069
R-squared	0.005	0.075	0.007	0.096
N	311,484	311,484	309,425	309,425

Note: All coefficients significant at p<0.001

Table 3B: Regression of PISA Reading Score on Effort

Test: Decline	-0.099			-0.092
Survey: Items Missing		-0.291		-0.305
Survey: Careless Answers			-0.078	-0.062
R-squared	0.010	0.085	0.006	0.109
N	311,484	311,484	309,425	309,425

Note: All coefficients significant at p<0.001

Table 3C: Regression of PISA Science Score on Effort

Test: Decline	-0.089			-0.082
Survey: Items Missing		-0.283		-0.298
Survey: Careless Answers			-0.077	-0.061
R-squared	0.008	0.080	0.006	0.103
N	311,484	311,484	309,425	309,425

Note: All coefficients significant at p<0.001

Table 4A*:* Results, Random Coefficients Estimates of Item Order on Performance

	Country	α_0	β_1	$SD(\alpha_{0i})$	$SD(\beta_{1i})$	$Corr(\alpha_{0i}, \beta_{1i})$
1	JPN	0.888	-0.123	0.181	0.169	-0.073
2	FIN	0.885	-0.079	0.167	0.155	-0.174
3	NLD	0.885	-0.100	0.171	0.128	-0.040
4	ISL	0.864	-0.117	0.192	0.187	-0.238
5	CZE	0.854	-0.100	0.190	0.164	-0.161
6	LIE	0.853	-0.124	0.176	0.180	-0.218
7	SGP	0.848	-0.080	0.189	0.156	-0.055
8	BEL	0.847	-0.103	0.192	0.160	-0.144
9	CHE	0.847	-0.112	0.188	0.161	-0.198
10	RUS	0.841	-0.195	0.202	0.235	-0.459
11	LUX	0.840	-0.145	0.208	0.184	-0.243
12	NZL	0.840	-0.109	0.185	0.162	0.007
13	SWE	0.833	-0.150	0.195	0.196	-0.240
14	DEU	0.831	-0.098	0.191	0.163	-0.173
15	CAN	0.829	-0.127	0.182	0.180	-0.188
16	EST	0.827	-0.089	0.172	0.156	-0.272
17	NOR	0.821	-0.150	0.183	0.187	-0.242
18	HKG	0.820	-0.079	0.171	0.144	-0.086
19	AUS	0.819	-0.119	0.191	0.162	-0.056
20	AUT	0.813	-0.092	0.199	0.158	-0.215
21	POL	0.812	-0.119	0.182	0.177	-0.284
22	DNK	0.812	-0.125	0.182	0.161	-0.171
23	SVK	0.812	-0.126	0.195	0.166	-0.326
24	MAC	0.810	-0.133	0.176	0.202	-0.445
25	ITA	0.809	-0.159	0.194	0.204	-0.318
26	TAP	0.802	-0.090	0.185	0.151	-0.200
27	ESP	0.801	-0.150	0.192	0.196	-0.315
28	KOR	0.801	-0.083	0.145	0.124	0.022
29	FRA	0.799	-0.152	0.204	0.200	-0.204
30	PRT	0.798	-0.154	0.192	0.198	-0.399
31	GRC	0.770	-0.229	0.198	0.227	-0.448
32	QCN	0.767	-0.073	0.156	0.115	-0.085
33	USA	0.759	-0.112	0.188	0.149	-0.101
34	LVA	0.758	-0.120	0.174	0.173	-0.361
35	SVN	0.758	-0.122	0.187	0.147	-0.135
36	GBR	0.757	-0.112	0.186	0.145	-0.084
37	LTU	0.716	-0.106	0.187	0.153	-0.320
38	HRV	0.713	-0.103	0.186	0.149	-0.315
39	IRL	0.703	-0.113	0.186	0.165	-0.167

40	ISR	0.664	-0.180	0.221	0.222	-0.347
41	HUN	0.600	-0.098	0.174	0.142	-0.195
42	TUR	0.564	-0.131	0.181	0.167	-0.359
43	THA	0.530	-0.137	0.180	0.163	-0.426
44	IDN	0.505	-0.159	0.152	0.181	-0.546

Table 4B: Random Coefficient Models Estimates of Test Decline

Overall	$\boldsymbol{\alpha_0}$	$\boldsymbol{\beta_1}$	$\boldsymbol{SD(\alpha_0^i)}$	$\boldsymbol{SD(\beta_1^i)}$	$\boldsymbol{Corr(\alpha_0^i, \beta_1^i)}$
Mean	0.784	-0.122	0.185	0.169	-0.227
SD	0.090	0.032	0.014	0.026	0.133

Table 5A: Variation in PISA Math Scores Explained by Student Motivation, Multilevel Model.

	Test: Decline	Survey: Item Nonresponse	Survey: Careless Answers	Combined
within country	0.003	0.066	0.006	0.084
between countries	0.248	0.413	0.018	0.385
Overall	0.005	0.075	0.007	0.096
Country n	44	44	44	44
Student n	311,484	311,484	309,425	309,425

Table 5B: Variation in PISA Reading Scores Explained by Student Motivation, Multilevel Model.

	Test: Decline	Survey: Item Nonresponse	Survey: Careless Answers	Combined
within country	0.008	0.079	0.005	0.101
between countries	0.187	0.330	0.020	0.324
Overall	0.010	0.085	0.006	0.109
Country n	44	44	44	44
Student n	311,484	311,484	309,425	309,425

Table 5C: Variation in PISA Science Scores Explained by Student Motivation, Multilevel Model.

	Test: Decline	Survey: Item Nonresponse	Survey: Careless Answers	Combined
within country	0.006	0.073	0.005	0.092
between countries	0.275	0.378	0.018	0.373
Overall	0.008	0.080	0.006	0.103
Country n	44	44	44	44
Student n	311,484	311,484	309,425	309,425

Table 6A: Student-level PISA Scores, Raw and Adjusted for Student Motivation

Variable	Mean	Std. Dev.	Min	Max
Math	501.4	97.0	48.1	1022.2
Math, Adjusted	520.5	92.2	83.9	1112.06
Reading	495.2	93.9	6.7	871.1
Reading, Adjusted	514.7	88.6	69.4	1083.5
Science	504.5	96.1	0.8	883.8
Science, Adjusted	524.0	91.0	62.0	1117.5

Note: N = 309,425

Table 6B: Country-level PISA Scores, Raw and Adjusted for Student Motivation

Variable	Mean	Std. Dev.	Min	Max
Math	501.5	38.0	372.8	600.1
Math, Adjusted	520.6	34.0	404.1	607.1
Reading	494.4	27.1	402.4	556.0
Reading, Adjusted	514.0	23.8	434.2	562.9
Science	504.1	32.6	383.1	575.2
Science, Adjusted	523.7	28.9	414.7	582.1

Note: N = 44

Appendix

Table A1: Country names and abbreviations

Abbre.	*Country name*	*Abbre.*	*Country name*
USA	United States of America	ITA	Italy
AUS	Australia	JPN	Japan
AUT	Austria	KOR	South Korea
BEL	Belgium	LIE	Liechtenstein
CAN	Canada	LTU	Lithuania
CHE	Switzerland	LUX	Luxembourg
CZE	Czech Republic	LVA	Latvia
DEU	Germany	MAC	Macao-China
DNK	Denmark	NLD	Netherlands
ESP	Spain	NOR	Norway
EST	Estonia	NZL	New Zealand
FIN	Finland	POL	Poland
FRA	France	PRT	Portugal
GBR	United Kingdom	QCN	Shanghai-China
GRC	Greece	RUS	Russian Federation
HKG	Hong Kong-China	SGP	Singapore
HRV	Croatia	SVK	Slovak Republic
HUN	Hungary	SVN	Slovenia
IDN	Indonesia	SWE	Sweden
IRL	Ireland	TAP	Chinese Taipei
ISL	Iceland	THA	Thailand
ISR	Israel	TUR	Turkey