Trends in International Assessments and Outcomes in Adulthood*

Samuel Stemper†

May 26, 2024

Abstract

International assessments such as PISA and TIMSS are widely used to compare the academic proficiency of adolescents across countries and over time. Do scores on these assessments predict outcomes in adulthood? Combining data from PISA, TIMSS, PIAAC, and 18 representative global surveys, I study the relative predictive power of PISA and TIMSS scores among cohorts that took both tests during adolescence. Results suggest that cohorts with higher test scores perform better on assessments of adulthood skills, obtain higher levels of education, and have higher incomes as adults. I find suggestive evidence that PISA scores exhibit a relatively stronger relationship with education and income in adulthood compared to TIMSS scores.

*First version: November 29, 2023. This version: May 26, 2024. I am grateful to Thomas Kane and Joshua Goodman for helpful comments and suggestions. All remaining errors are mine.

†University of Auckland. Email: sam.stemper@auckland.ac.nz

Math skills play an important role in the academic and economic trajectory of individuals throughout their lives. Training to develop these skills is concentrated during childhood and adolescence. Policymakers often use international assessments such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) to monitor the academic proficiency of their students. Trends in these assessments are often used to evaluate the trajectory of education systems within a country; differences in test scores across nations are often used to identify promising strategies for education policy.

Do these assessments measure the skills that ultimately equip students for success in adulthood? This paper aims to address this question by investigating whether cohorts that score higher on international assessments during adolescence ultimately have more favorable educational and economic outcomes in adulthood.

This question is of particular significance for two reasons. First, PISA and TIMSS exams test distinct skills, even within the same subject. TIMSS emphasizes curriculum-based knowledge, focusing on material that students (ought to) learn in school. Alternatively, PISA measures students' ability to apply their knowledge in "real-world" scenarios, with less emphasis on curricular material.[1] Second, in many countries, PISA and TIMSS scores have moved in opposite directions since 2000. While TIMSS math scores have increased in most participating countries, PISA scores have stagnated or declined. This phenomenon is not explained by changes in the composition of participating countries or within-country time trends.

In this paper, I use variation in PISA and TIMSS test scores across cohorts and across countries to test the degree to which country-by-cohort average test scores predict outcomes in adulthood. I use data from the Programme for the International Assessment of Adult Competencies (PIAAC) to measure skills in adulthood and use harmonized international survey data to measure education and income. I focus on cohorts that took both assessments and evaluate the relative strength of these cohort-level measures in predicting the skills and outcomes of surveyed members of these cohorts later in life.

Across measures of skills, education, and income, I find that both PISA and TIMSS scores are positively associated with outcomes in adulthood. A 1 standard deviation cohort-level increase in PISA scores is associated with a 0.2 standard deviation increase in adulthood numeracy test

[1] This is one of many differences between these assessments. I discuss these differences more broadly in Section 1.

scores, a 1-year increase in years of education, and a 5 to 11 percentage-point increase in household income percentile. For TIMSS scores, effect sizes with respect to adulthood numeracy test scores are similar, but generally weaker with respect to education attainment and income. Among these measures of educational attainment and income, I find suggestive evidence for differences in these magnitudes, both in separate regressions as well as "horserace" regressions that include both scores.

This work relates to a broad set of literature on the relationship between measures of human capital and education, income, and growth. Of particular relevance to my study are Doty et al. (2022) and Égert et al. (2024), who use cohort-level variation (across US states and countries, respectively) in test scores to estimate the relationship between test scores and outcomes in adulthood. Methodologically, I employ a comparable approach to Doty et al. (2022). However, I differ from both papers in my use of multiple measures of skills to evaluate the relationship between skills and outcomes across different testing regimes.

More broadly, many studies examine the role of skills, as measured by test scores, in driving individual outcomes within countries or differences in economic growth across countries. Regarding the prior, Goldhaber and Özek (2019) and Hanushek (2012) summarize this literature, with the latter concluding that, in developed countries "[t]here is now considerable evidence that cognitive skills measured by test scores are directly related to individual earnings, productivity, and economic growth."[2] Regarding the latter, many studies study the role of country-level differences in test scores and education levels in driving differences in economic growth rates. This work typically uses variation in test scores across countries, rather than over time (Barro, 1991; Hanushek, 2013; Mankiw et al., 1992).[3]

I contribute to this literature by evaluating the relative predictive value of different measures of cohort skills according to two prominent testing regimes: PISA and TIMSS. As noted by Hanushek et al. (2015) and Hanushek et al. (2017), the effects of skills differ substantially across countries. Ideally, one would compare the predictive values of different skill measures within the same set of individuals. In the absence of *individuals* who took both exams, I focus on cohort-level

[2]My work relates specifically to math skills, which have been studied in more depth in relation to high school curriculum (Goodman, 2019; Joensen and Nielsen, 2009) and college majors (Kirkeboen et al., 2016). Altonji et al. (2012) summarize this literature in their review.

[3]Notable exceptions include Coulombe et al. (2004) and Coulombe and Tremblay (2006), who estimate growth regressions using variation across cohorts and across countries.

variation among cohorts that took both tests, following the approach in Doty et al. (2022).

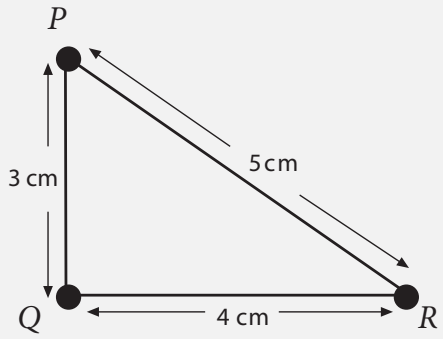
1 Background on PISA and TIMSS

PISA and TIMSS are both international assessments designed to evaluate and compare the educational performance of adolescent students across countries. The Organisation for Economic Co-operation and Development (OECD) conducts the PISA assessment every three years and tests 15-year-old students in mathematics, reading, and science. TIMSS, organized by the International Association for the Evaluation of Educational Achievement (IEA), assesses mathematics and science proficiency among 4th and 8th graders every four years. For both surveys, countries choose whether to participate in each round of assessment, and sampling and test administration are typically coordinated separately within each country.

1.1 Assessment Methodologies

PISA and TIMSS differ substantially in the material used to assess student skills. TIMSS is curriculum-based and tests students' knowledge of the material taught in school. Alternatively, PISA assesses students' ability to apply knowledge to "real-world" problems (Loveless, 2013). Sample questions from the 2011 Grade 8 TIMSS exam and the 2012 PISA exam illustrate this difference. As an example, in 2011, the TIMSS Grade 8 exam included the question below.

Example TIMSS Question



Which of these is the reason that triangle PQR is a right angle triangle?

A. $3^2 + 4^2 = 5^2$

B. $5 < 3 + 4$

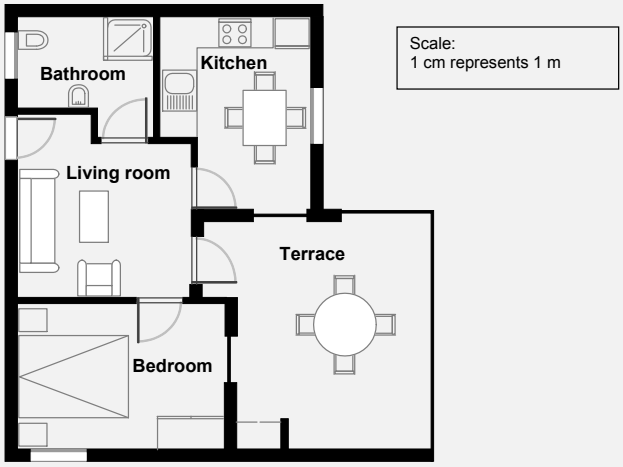
C. $3 + 4 = 12 - 5$

D. $3 > 5 - 4$

In contrast, the PISA 2012 exam included the question below.

Example PISA Question

This is the plan of the apartment that George's parents want to purchase from a real estate agency.



To estimate the total floor area of the apartment (including the terrace and the walls), you can measure the size of each room, calculate the area of each one and add all the areas together. However, there is a more efficient method to estimate the total floor area where you only need to measure 4 lengths. Mark on the plan above the four lengths that are needed to estimate the total floor area of the apartment.

Both questions above involve reasoning with geometric shapes. While the TIMSS question is a standalone mathematical problem involving knowledge of the Pythagorean theorem, the PISA question requires students to apply geometric reasoning to a real-world scenario. Appendix A provides links to publicly available questionnaires from PISA and TIMSS exams.

1.2 Differences in Tested Population

Beyond the content of the questions used in assessments, PISA and TIMSS also differ in the test-taking population. Most importantly, PISA has an age-based sample, testing 15-year-olds (regardless of grade), whereas TIMSS has a grade-based sample, testing students in grades 4 and 8 (Loveless, 2013). As a result, students tested during PISA assessments are typically one year older

than students tested during TIMSS Grade 8 assessments. More substantively, TIMSS data contains students who are either ahead of or behind the typical grade progression in their country. In Section 2, I describe my process to make these test-taking populations more comparable by restricting my focus to students who follow their country's typical grade progression.

1.3 Trends in PISA and TIMSS

PISA and TIMSS differ not only in the content and student population used in their assessments but also in their conclusions about the trajectory of global math skills. While TIMSS math scores have steadily increased over the past 25 years, PISA math scores have been flat or decreasing.

Figure 1 illustrates this phenomenon among the most well-represented countries in PISA and TIMSS assessments: the six countries that have participated in PISA and TIMSS Grade 8 assessments every year since their inception. These countries are Australia, Hong Kong, Hungary, Japan, Korea, and the United States. For each country, Figure 1 displays the country's average scores in the 7 PISA math assessments and the 7 TIMSS Grade 8 math assessments to date.[4] Lines show linear time trends, separately for PISA and TIMSS; the 18-year change associated with this time trend is shown in the corner of each panel.

Figure 1 demonstrates that, across many countries, PISA scores have fallen relative to TIMSS scores. These are large discrepancies: PISA math scores in Australia fell by over 0.4 standard deviations between 2000 and 2018. Over the same period, TIMSS Grade 8 scores were flat. For reference, Bloom et al. (2008) and Evans and Yuan (2019) find that a year of schooling typically increases test score performance by 0.3 and 0.2 standard deviations, respectively.[5] All 6 of the countries in Figure 1 exhibit relatively better trends on TIMSS tests than PISA tests; 4 of these differences are statistically significant.

In Appendix Table B.1, I formally test for differences in the trajectories of PISA and TIMSS assessments using student-level data from all PISA and TIMSS participating countries. On average, TIMSS Grade 8 math scores grew 0.012 standard deviations faster per year than PISA math scores. These patterns are not driven by changes in the composition of test-taking countries; regressions that include country fixed effects and country-specific time trends yield nearly identical results.

[4] These scores incorporate the sample restrictions and rescaling procedure described in Section 2 below.

[5] Bloom et al. (2008) use nationally-representative data from the U.S. The 0.3 estimate refers to the effect of a year of schooling for 8th to 9th graders, who are typically 13 to 15 years old. Evans and Yuan (2019) use a sample of test scores from low- and middle-income countries.

Over an 18-year period, this translates to a divergence of over 0.2 standard deviations between TIMSS Grade 8 math and PISA math scores.

This phenomenon, while significant and widespread, has not unfolded uniformly across all nations. Figure 2 illustrates country-level variation in long-term test score trends. Specifically, the horizontal axis displays estimated time trends in PISA math scores and the vertical axis displays the equivalent estimates for TIMSS Grade 8 math scores; these time trends are estimated identically to those in Figure 1. In this analysis, I limit the sample to countries with at least one PISA test and one TIMSS Grade 8 test in 2006 or earlier and at least one PISA test and one TIMSS Grade 8 test in 2012 or later. Consistent with the evidence presented above, most countries exhibit larger growth in TIMSS scores relative to PISA scores. This can be seen visually in Figure 2; most points fall above the 45-degree line. However, some countries deviate from this pattern, with greater growth in PISA scores relative to TIMSS scores; Italy is a notable example. Furthermore, there are countries that exemplify an extreme manifestation of this trend; for instance, in Korea, time trends in PISA scores suggest an 18-year decrease in PISA scores of more than 0.15 standard deviations, while TIMSS scores have increased by nearly 0.25 standard deviations.

The data and methods described in the section that follows use this variation in test scores—over time, across countries, and between assessments—to test whether these assessments are predictive of outcomes in adulthood.

2 Data and Methods

2.1 PISA and TIMSS Test Scores

I prepare country-by-cohort test score averages for PISA math and Grade 8 TIMSS math using publicly available student-level test score data.

As discussed above, PISA has an age-based sample whereas TIMSS has a grade-based sample. To make these test-taking populations more comparable, I restrict my PISA sample to students who are in the modal grade among test-takers in their country. To generate the equivalent sample in TIMSS data, I restrict my focus to students who are within the 1 year range centered at the modal age in TIMSS data. Altogether, these restrictions mean that both samples include students who are in the typical grade for their age.

Both PISA and TIMSS scores were scaled at their introduction to have mean 500 and standard deviation of 100. To place PISA and TIMSS scores on the same scale, I apply the procedure used in Gust et al. (2024). To do so, I first restrict my focus to countries that participated in both the TIMSS 2019 and PISA 2018 tests.[6] Among these countries, I then standardize the TIMSS Grade 8 math scores to have a mean of zero and a standard deviation of one and estimate conversion parameters to rescale these scores to match the mean and standard deviation of the PISA math scale. With these conversion parameters, I apply this rescaling to all data for Grade 8 TIMSS scores. I refer to these rescaled TIMSS scores as "PISA Scale" scores throughout this paper.

After this transformation, I compute the average birth year and average test score in math for each country and testing round (e.g. 2000 PISA, 2015 Grade 8 TIMSS, etc.). In some years, student-level data does not report students' year of birth; in these cases, I compute birth years based on each student's age and the test date.[7]

The set of countries identified by PISA and TIMSS administrators are generally comparable, with a few exceptions such as Belgium, which is administered separately in Flemish and French regions, and the UK, which is administered separately in England, Northern Ireland, and Scotland. Where possible, I separately identify these regions in PISA and other data.

Tests are typically administered every three years for PISA and every four years for TIMSS. Due to this periodicity, there is rarely direct overlap in individual birth years across the two tests. To generate larger overlapping test scores for both tests, I impute observations by taking the closest non-missing test score that is no more than two birth years away. In the case of ties, I take the later test score.[8]

For example, the United States participated in the Grade 8 TIMSS math exam in 1995 and 1999 (among other years). Students taking these exams were, on average, born in 1981 and 1985, respectively. Following the procedure described above, I assign the 1981 birth cohort's TIMSS Grade 8 score to students born between 1979 and 1982 and assign the 1985 birth cohort's score to

[6] These dates are the same as those used in Gust et al. (2024).

[7] I assume PISA tests are administered in May and TIMSS are administered in July.

[8] The choice of how many years to impute reflects a tradeoff between sample size (greater imputation allows for larger overlapping samples) and cohort-level precision (cohorts further away from the tested cohort are more dissimilar). In robustness, I show that I obtain similar estimates using differing levels of cohort-level imputation, suggesting that my results are driven by long-term test score changes over time rather than round-to-round fluctuations in test score performance.

students born between 1983 and 1986.[9]

Finally, I keep only country-by-cohort observations with both PISA and TIMSS scores. This restriction limits the set of cohorts that appear in my data but ensures that statistical tests using PISA scores have the same sample as tests using TIMSS scores, and vice versa.

I link this country-by-cohort data with individual-level outcomes—namely, adult PIAAC test scores in numeracy and education and income data from harmonized international surveys. Because this individual data is not available for all birth years and all countries, this generates two slightly different sets of birth cohort-country combinations. I refer to these different samples as the "PIAAC Sample" and the "SDR Sample." Appendix Figure B.1 shows the number of observations in each country-cohort combination across both samples.

2.2 PIAAC Scores

To assess the numeracy skills of adults, I use individual-level test results from PIAAC. Similar to PISA and TIMSS, PIAAC is an international survey of skills. Unlike PISA and TIMSS, PIAAC assesses the skills and competencies of adults aged 16 to 65. PIAAC rounds took place in 2012, 2014, and 2017. Each country only participated in one round of PIAAC testing, with one exception: the United States administered PIAAC tests in all three years.[10]

I focus on PIAAC numeracy scores. The OECD defines numeracy as "the ability to access, use, interpret, and communicate mathematical information and ideas, in order to engage in and manage the mathematical demands of a range of situations in adult life" (OECD et al., 2009). Appendix A includes links to publicly available documents that provide examples of questions included in the PIAAC numeracy assessment. The OECD reports PIAAC scores on a scale ranging from 0 to 500. This scale is based on "information-processing tasks of increasing complexity," and is calculated such that "[a]t each point on the scale, an individual with a proficiency score of that particular value has a 67% chance of successfully completing test items located at that point" (OECD, 2013). To simplify interpretation, I scale individual numeracy test scores such that they have mean 0 and standard deviation 1.

[9]Students born in 1987 are assigned the score of the next cohort: students who took the Grade 8 TIMSS math exam in 2003.

[10]The United States' 2014 PIAAC round is known as the National Supplement to the Main Study. This round was not administered to a nationally representative sample and was instead meant to "enhance and expand" the Round 1 data. See NCES, PIAAC Participating Countries.

I link PIAAC participants to their cohort's PISA and TIMSS scores based on country and birth year, which I calculate based on the year of the test and the respondent's age. For a small number of countries, respondent ages are reported in 5-year ranges (e.g. 20-24); in these cases, I take the midpoint and round up to the nearest integer.

2.3 Harmonized International Survey Data

For education and income, I use harmonized international survey data from the Survey Data Recycling ("SDR") project.[11] The SDR project collects publicly available data from numerous large-scale international survey projects, such as the International Social Survey Programme and the European Social Survey, and harmonizes responses to common demographic questions, among others.[12] The Survey Data Recycling project selects surveys based on the following criteria: surveys must be "designed as cross-national [...]; samples intended to be representative of the adult population of given country or territory; projects contain questions about political attitudes and behaviors; projects are freely available in the public domain; and their documentation [...] is provided in English."[13] SDR raw data contains 4.4 million responses from 23 survey projects. I link these responses to their cohort's PISA and TIMSS test scores; the resulting data contains over 115,000 responses from 18 survey regimes.[14] These surveys took place between 1998 and 2017.

I consider three main educational outcomes: years of education as well as whether the respondent completed secondary or tertiary school. Many surveys do not ask for years of education directly but ask for the age a respondent finished education, the year in which a respondent finished education, or the highest level of education they received. The SDR project harmonizes across all of these response types to generate a value that is comparable across survey rounds.[15]

[11] Specifically, I use the Survey Data Recycling (SDR) v.2.0 database. This data and supporting documentation is available here: SDR2 Database.

[12] Slomczynski and Tomescu-Dubrow (2018) contains more information on this harmonization process.

[13] Individually, these surveys have been used extensively in economics and political science research. For example, Aghion et al. (2010) and Mayda (2006) use data from the International Social Survey Programme, Hainmueller and Hiscox (2007) and Roth and Wohlfart (2018) use data from the European Social Survey, and Besley and Persson (2019) and Guriev and Zhuravskaya (2009) use data from the World Values Survey.

[14] The reduction in sample size is accounted for mostly by birth year: 3.7 million respondents in SDR raw data were born prior to 1982, the first birth cohort for which PISA and TIMSS data are available in any country.

[15] SDR documentation describes their process as follows: "To construct the target variable T_EDU_YEARS, we use source items about respondents': (a) exact number of education years completed; (b) age when finished full-time education; (c) year when finished school; (d) years of education derived by survey providers from various source variables. We rely on the English language and Spanish language questionnaires and codebooks describing the source survey data.

Typical questions on respondent's years of education are: "How many years of formal education have you received?" (ABS), "How many years of schooling have you completed?"(AMB/2010-2016), "About how many years of education

Additionally, I consider household income percentiles. In reporting their household income, many surveys ask respondents to select from a set of pre-coded income brackets. To account for different reporting schemes across surveys and countries, SDR data estimates each respondent's position in the national income distribution by converting these brackets into percentiles within each national survey. More technically, income brackets are sorted in ascending order (from the lowest to the highest household income) and are assigned values of the mid-point from the cumulative distribution.

One concern with using household income, rather than personal income or wages, is that household income may be affected by household composition. While I do not have data on particular aspects of household composition (e.g. whether respondents are living with their parents and/or their spouse), I later show that my main results are robust to flexible controls for household size.

Finally, I restrict the sample to respondents who have non-missing PISA and TIMSS cohort scores and are age 16 or older at the time of the interview.[16]

2.4 Data Description

Table 1 summarizes the main variables across my two samples. Across both samples, slightly more than half of respondents are women, and the average age is roughly 23 years.[17] In my SDR data, the average respondent had 13 years of education. At the time of the interview, 83 percent of respondents aged 19 or older had completed secondary school and 37 percent aged 24 and older had completed a bachelor's degree.

As noted above, these samples contain slightly different sets of country-by-cohort pairs. Appendix Figure B.1 summarizes the set of countries and birth cohorts in my samples. There are 28 unique countries in the PIAAC sample, 15 containing at least 3 unique PISA scores and 3 unique Grade 8 TIMSS scores. There are 51 unique countries in the SDR sample, 27 containing at least

have you completed, whether full-time or part-time?" (ESS). Typical questions on the age of completing education are: "How old were you when you finished your full-time education?" (CDCEE), "At what age did you finish your education (full-time education)?" (LB). In two cases, respondents were asked to indicate the year of their education completion: "When did you obtain this degree?" (LITS/1), "When did you obtain this qualification?" (LITS/2)."

[16] A small number of respondents report ages that don't align with their reported year of birth and survey year. I drop all observations for which the difference between (a) a respondent's predicted age (survey year minus birth year) and (b) a respondent's report age is larger than 2.

[17] I additionally include age statistics for individuals aged 19 and above and 24 and above. These samples are used in analyses related to secondary school completion, bachelor's completion, and household income percentiles.

3 unique PISA scores and 3 unique Grade 8 TIMSS scores. Table 1 additionally lists the set of 18 survey sources included in my sample.

2.5 Methodology

My method estimates the linear relationship between a cohort's test scores in adolescence and their outcomes in adulthood. My baseline regression takes the form below.

$$y_{icbt} = \beta_0 + \beta_1 TestScore_{cb} + \gamma_c + \delta_{age} + \zeta_t + \theta_{gender} + \varepsilon_{icbt} \tag{1}$$

y_{icbt} denotes an outcome (e.g. years of education, household income percentile) for individual i in country c born in year b and surveyed in year t. For all outcomes, I estimate separate regressions using either PISA or TIMSS scores individually, as well as a "horserace" regression that includes both scores. Additionally, when possible, I test whether my results are sensitive to including region-by-age fixed effects—which allow age fixed effects to vary across regions of the world, as specified by the World Bank—and country-by-survey year fixed effects—which control for time-specific country factors.

My coefficient of interest, β_1, measures the association between cohort-level average test scores and individual outcomes in PIAAC or SDR data. Importantly, these estimates may not necessarily reflect the *causal* effect of skills on these outcomes. There are many reasons for variation in national test scores: namely, changes in the composition of students, changes in the quality of the education they received, and changes in their environment, among others. These factors may have independent effects on outcomes in adulthood that are entirely unrelated to test scores. As such, the estimates in this paper should be viewed as evaluating the predictive validity of test scores on later-life outcomes, rather than the causal effect of skills.

For regressions using PIAAC data, I weight each observation by $w_{ict} / \sum_{i \in ct} w_{ict}$, where w_{ict} is individual i in country c in year t's final sampling weight, and $\sum_{i \in ct} w_{ict}$ denotes the sum of sampling weights in country c in year t. In addition, in these regressions, I account for imputation error by using Rubin's rule and the PIAAC plausible values (Rubin, 1987). For regressions using SDR data, I weight estimates using the SDR-provided weights, which rescale sampling weights from each national survey such that the sum within each survey equals the number of respon-

dents. Following Doty et al. (2022), I cluster standard errors by country and birth year.

3 Results

Table 2 shows the results of regressions using PIAAC data, which assess the relationship between adolescent test scores and numeracy skills in adulthood. Regression results shown in Columns 1 to 3 include age and age-squared terms and fixed effects for country, gender, and test year.[18] Column 1 displays the relationship between a cohort's PISA scores and that cohort's PIAAC scores in adulthood, which indicates that a 1 standard deviation increase in a cohort's average PISA score is associated with a 0.21 standard deviation increase in that cohort's PIAAC scores in adulthood ($p < 0.1$). In Column 2, I repeat this specification, replacing PISA scores with TIMSS scores. Here, I find that a 1 standard deviation increase in a cohort's TIMSS average math score is associated with a 0.26 increase in that cohort's PIAAC scores ($p < 0.05$). Below the estimate in Column 2, I display the results of a test of equality of coefficients for the estimates in Columns 1 and 2, which indicates that the difference between the PISA estimate and TIMSS estimate is not significant ($p = 0.42$). In Column 3, I estimate the effects of PISA and TIMSS scores in the same regression and find that estimates are reasonably similar to those estimated independently: associations between TIMSS and PIAAC are larger than associations between TIMSS and PIAAC, though differences between these effect sizes are statistically insignificant.

Columns 4 to 6 of Table 2 employ an additional specification by incorporating region-by-age interaction terms. The inclusion of these controls does not substantially alter the estimates; coefficients on TIMSS scores are slightly larger than corresponding coefficients on PISA scores, but these differences are not statistically significant ($p = 0.63$, $p = 0.67$). These estimates demonstrate that cohort-level variation in adolescent math skills indeed translates into later-life differences in numeracy skills.[19] To assess the persistence of these patterns with respect to education and income in adulthood, I next turn to harmonized international survey data.

Table 3 shows the results of these estimates. Regression results shown in Columns 1 to 3

[18] I use continuous terms for age, rather than fixed effects, due to the coarse rounding of ages for some countries in PIAAC data, described above.

[19] In similar specifications using TIMSS data and combined student test score data from PISA and the World Bank, Égert et al. (2024) find similar-sized estimates. Specifically, Égert et al. (2024) estimate log-log regressions of PIAAC scores on PISA scores extended backward with two vintages of World Bank data. Their estimated coefficients are between 0.2 and 0.7.

include fixed effects for country, age, gender, survey year, and survey wave, and results shown in Columns 4 to 6 add region-by-age fixed effects. The presence of multiple surveys within the same country over time allows for the inclusion of country-by-survey year fixed effects, which I add in Columns 7 to 9.

Panels A, B, and C of Table 3 display effects on measures of educational attainment. In Panel A, I estimate associations between adolescent test scores and years of education. Across numerous specifications, these results suggest that a 1 standard deviation increase in cohort-level PISA scores is associated with a 0.6 to 1.1-year increase in years of education. TIMSS scores do not exhibit any systematic relationship with years of education; these estimates are consistently small and negative. I test for differences between these two estimates—PISA and TIMSS effects—and display corresponding p-values below coefficient estimates in Panel A; these estimates suggest that PISA scores exhibit a statistically larger effect on years of education than TIMSS scores.

Panels B and C of Table 3 show effects on discrete levels of education: completion of secondary school and tertiary school, respectively. Panel B shows that both PISA and TIMSS test scores are associated with higher rates of secondary school completion. Effect sizes suggest that a 1 standard deviation increase in PISA scores increases secondary school completion by roughly 10 to 13 percentage points. Increases in TIMSS scores are not systematically associated with an effect on secondary school completion. The results in Panel C suggest a positive relationship between test scores and tertiary school completion, but these effects are statistically imprecise for both PISA and TIMSS scores.

Finally, in Panel D of Table 3 I report the effects on household income percentiles. Among respondents aged 24 or older, a 1 standard deviation increase in cohort PISA scores is associated with a 5 to 11 percentile increase in household income. TIMSS scores exhibit slightly smaller but highly significant effects, suggesting that a 1 standard deviation increase in TIMSS scores is associated with a 3 to 6 percentile increase in household income. The difference between the estimated effects of PISA and TIMSS scores varies in statistical significance across specifications. In my most stringent specification (in Columns 7 to 9), the effect of PISA scores is statistically significantly different from the effect of TIMSS scores.

I subject these results to a number of robustness tests, which appear in Appendix B.

As noted above, periodicity in the timing of PISA and TIMSS tests means that there is rarely

direct overlap in individual birth years across the two tests. In my main estimates, I impute observations by taking the closest non-missing test score that is no more than two birth years away. This choice balances a tradeoff between sample size (greater imputation allows for larger overlapping samples) and cohort-level precision (cohorts further away from the tested cohort are more dissimilar).

In Figures B.2 and B.3, I show how my estimates change as I vary the number of adjacent cohorts imputed in my sample. Figure B.3 shows that the standard errors of my coefficients tend to decrease as I increase the number of adjacent cohorts in my sample, and this decrease is most substantial when I move from 1 year of adjacency to 2 years. Figure B.2 indicates that my coefficient estimates are quite stable across estimates using various degrees of imputation.

Figures B.4 and B.5 repeat this analysis with respect to education and income, with similar conclusions: coefficient standard errors drop substantially when moving from 1 to 2 years of imputation, and coefficients are generally stable across estimates.

Altogether, these results suggest that it is likely the long-run changes in test scores (e.g. decade-to-decade variation) that drive my results, rather than year-to-year variation in skill measures.

Next, I show that my results concerning income are qualitatively similar when I control for fixed effects for household size. These results are shown in Appendix Table B.2.

Finally, I confirm that my estimates with respect to education and income are not driven by any one country or any one survey. To do so, I reproduce my estimates 10 times, once after excluding each of the largest 10 countries individually. The results are shown in Appendix Figure B.6, which shows that my results are stable, regardless of which country is excluded. I repeat this process, instead dropping individual surveys; these results are shown in Appendix Figure B.7.

4 Conclusion

In this paper, I estimate the degree to which outcomes in adulthood are explained by cohort-level variation in test scores. Comparing results from two major international testing regimes—PISA and TIMSS—I find that math scores from both tests are strong predictors of adult skills, education levels, and incomes. Additionally, the evidence suggests that PISA scores have a stronger connection with education and income levels in adulthood than TIMSS scores. These results have several implications for policymakers and researchers alike.

Most substantially, these results highlight a concern for numerous countries that have witnessed stagnant or falling PISA scores alongside comparatively stronger growth in TIMSS. To the degree that PISA scores are more predictive of future educational and economic success, this highlights a concern for young cohorts who have generally performed worse than cohorts before them.

Concerning research, this work stands in contrast to efforts to harmonize results from numerous international assessments into a standard measure of education quality (e.g. Angrist et al. (2021)). My results here suggest that future research should be cautious about a single definition of human capital which is measured uniformly across diverse testing regimes. Instead, researchers should embrace a richer model of skills to study both the potential drivers of skills as well as the impact of skills on economic and non-economic outcomes. Recent work by Hermo et al. (2022), which distinguishes between "reasoning" skill and "knowledge" skill, is one such example.

References

Aghion, Philippe, Yann Algan, Pierre Cahuc, and Andrei Shleifer, "Regulation and distrust," *The Quarterly journal of economics*, 2010, *125* (3), 1015–1049.

Altonji, Joseph G, Erica Blom, and Costas Meghir, "Heterogeneity in human capital investments: High school curriculum, college major, and careers," *Annu. Rev. Econ.*, 2012, *4* (1), 185–223.

Angrist, Noam, Simeon Djankov, Pinelopi K Goldberg, and Harry A Patrinos, "Measuring human capital using global learning data," *Nature*, 2021, *592* (7854), 403–408.

Barro, Robert J, "Economic growth in a cross section of countries," *The quarterly journal of economics*, 1991, *106* (2), 407–443.

Besley, Timothy and Torsten Persson, "Democratic values and institutions," *American Economic Review: Insights*, 2019, *1* (1), 59–76.

Bloom, Howard S, Carolyn J Hill, Alison Rebeck Black, and Mark W Lipsey, "Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions," *Journal of Research on Educational Effectiveness*, 2008, *1* (4), 289–328.

Coulombe, Serge and Jean-François Tremblay, "Literacy and Growth," *B.E. Journals in Macroeconomics*, 8 2006, *6* (2), 4.

_ , Jean-Francois Tremblay, and Sylvie Marchand, "Literacy Scores, Human Capital and Growth Across Fourteen OECD Countries," *Statistics Canada Catalogue*, 2004, (89-552-mpe, no. 11).

Doty, Elena, Thomas J Kane, Tyler Patterson, and Douglas O Staiger, "What Do Changes in State Test Scores Imply for Later Life Outcomes?," Technical Report, National Bureau of Economic Research 2022.

Égert, Balázs, Christine De la Maisonneuve, and David Turner, "A new macroeconomic measure of human capital exploiting PISA and PIAAC: Linking education policies to productivity," *Education Economics*, 2024, pp. 1–17.

Evans, David and Fei Yuan, "Equivalent years of schooling: A metric to communicate learning gains in concrete terms," *World Bank Policy Research Working Paper*, 2019, (8752).

Goldhaber, Dan and Umut Özek, "How much should we rely on student test achievement as a measure of success?," *Educational Researcher*, 2019, *48* (7), 479–483.

Goodman, Joshua, "The labor of division: Returns to compulsory high school math coursework," *Journal of Labor Economics*, 2019, *37* (4), 1141–1182.

Guriev, Sergei and Ekaterina Zhuravskaya, "(Un) happiness in transition," *Journal of economic perspectives*, 2009, *23* (2), 143–168.

Gust, Sarah, Eric A Hanushek, and Ludger Woessmann, "Global universal basic skills: Current deficits and implications for world development," *Journal of Development Economics*, 2024, *166*, 103205.

Hainmueller, Jens and Michael J Hiscox, "Educated preferences: Explaining attitudes toward immigration in Europe," *International organization*, 2007, *61* (2), 399–442.

Hanushek, Eric A, "The economic value of education and cognitive skills," *Handbook of education policy research*, 2012, pp. 39–56.

_ , "Economic growth in developing countries: The role of human capital," *Economics of education review*, 2013, *37*, 204–212.

_ **, Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann**, "Returns to skills around the world: Evidence from PIAAC," *European Economic Review*, 2015, *73*, 103–130.

_ **,** _ **,** _ **, and** _ , "Coping with change: International differences in the returns to skills," *Economics Letters*, 2017, *153*, 15–19.

Hermo, Santiago, Miika Päällysaho, David Seim, and Jesse M Shapiro, "Labor market returns and the evolution of cognitive skills: Theory and evidence," *The Quarterly Journal of Economics*, 2022, *137* (4), 2309–2361.

Joensen, Juanna Schrøter and Helena Skyt Nielsen, "Is there a causal effect of high school math on labor market outcomes?," *Journal of Human Resources*, 2009, *44* (1), 171–198.

Kirkeboen, Lars J, Edwin Leuven, and Magne Mogstad, "Field of study, earnings, and self-selection," *The Quarterly Journal of Economics*, 2016, *131* (3), 1057–1111.

Loveless, Tom, "International tests are not all the same," *Brookings: Research*, 2013.

Mankiw, N Gregory, David Romer, and David N Weil, "A contribution to the empirics of economic growth," *The quarterly journal of economics*, 1992, *107* (2), 407–437.

Mayda, Anna Maria, "Who is against immigration? A cross-country investigation of individual attitudes toward immigrants," *The review of Economics and Statistics*, 2006, *88* (3), 510–530.

OECD, *Reporting the results* 2013.

_ **, PIAAC Numeracy Expert Group et al.**, "PIAAC numeracy: a conceptual framework," Technical Report, OECD education working paper, Bd. 35 2009.

Roth, Christopher and Johannes Wohlfart, "Experienced inequality and preferences for redistribution," *Journal of Public Economics*, 2018, *167*, 251–262.

Rubin, Donald B., *Multiple imputation for nonresponse in surveys*, John Wiley Sons., 1987.

Slomczynski, Kazimierz M and Irina Tomescu-Dubrow, "Basic principles of survey data recycling," *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*, 2018, pp. 937–962.

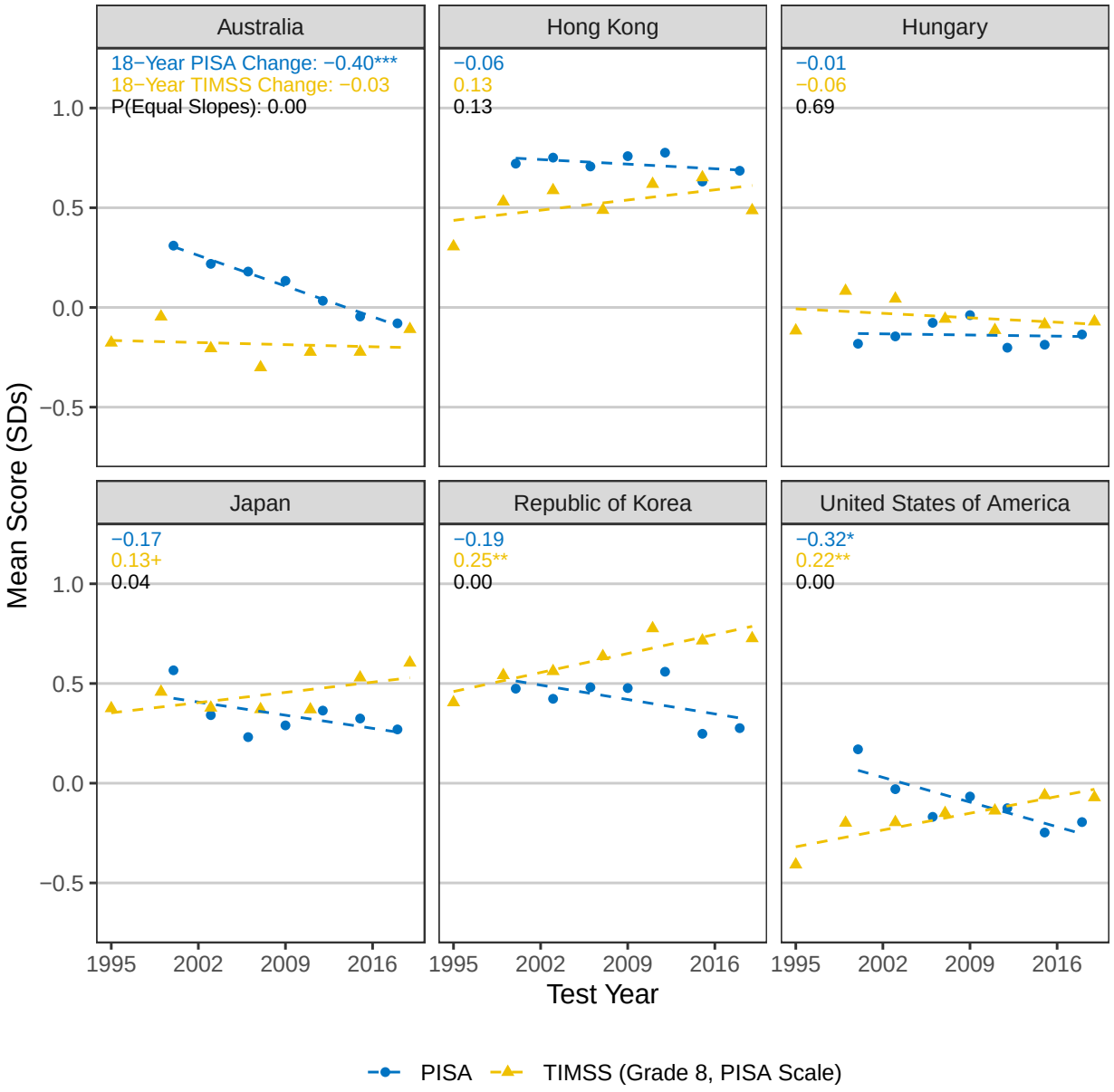


Figure 1: Trends in PISA and TIMSS Grade 8 Math Scores Across Continuously Participating Countries

Note: Figure displays average PISA and TIMSS Grade 8 math scores among the 6 countries that have participated in every round of PISA and TIMSS Grade 8 testing since their inception in 1995 and 2000, respectively. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Text at the top left of each panel summarizes the results of the following regression, run separately for each country and test (PISA or TIMSS Grade 8):

$$AvgScore_{ct} = \beta_0 + \beta_1 Year_t + \varepsilon_{ct}$$

The displayed number corresponds to $18 \times \beta_1$. + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001.

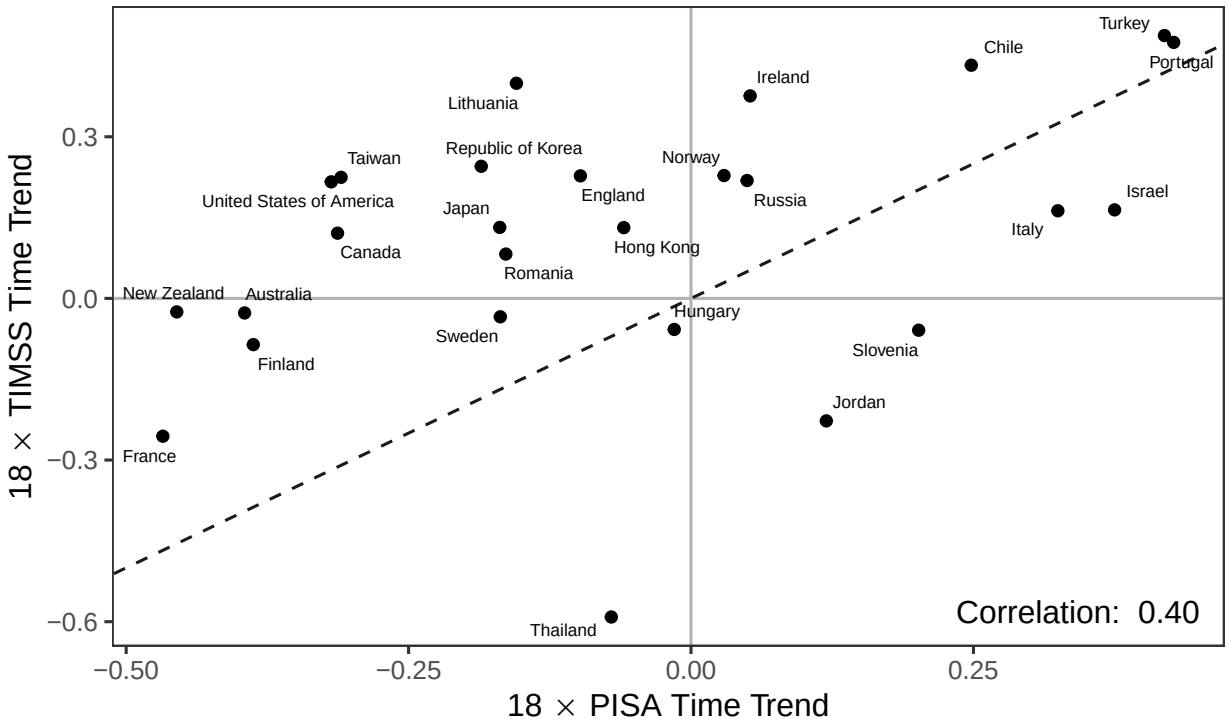


Figure 2: Growth in PISA and TIMSS Grade 8 Math Scores

Note: Figure displays estimated time trends for average PISA and TIMSS Grade 8 math scores across countries. Displayed countries are those with at least one PISA test and one TIMSS Grade 8 test in 2006 or earlier and at least one PISA test and one TIMSS Grade 8 test in 2012 or later. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Time trends are estimated separately for PISA and TIMSS Grade 8 math scores using the regression equation:

$$AvgScore_{ct} = \beta_0 + \beta_1 Year_t + \varepsilon_{ct}$$

The displayed number corresponds to $18 \times \beta_1$. Dashed line is the 45-degree line.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
Panel A: PIAAC Sample					
Female	31,433	0.525	0.499	0	1
Age	31,433	23.537	4.474	16	32
Survey Year	31,433	2,013.150	1.469	2,012	2,017
Year of Birth	31,433	1,989.613	4.741	1,982	2,001
PISA Math Score	31,433	−0.008	0.405	−0.922	1.029
TIMSS Math Score (Grade 8, PISA Scale)	31,433	−0.226	0.375	−1.260	0.779
PIAAC Numeracy Score	31,433	0.056	0.942	−4.298	3.477
Panel B: SDR Sample					
Female	104,836	0.520	0.500	0	1
Age	104,836	22.982	4.216	16	35
Age |19 +	89,441	23.928	3.827	19	35
Age |24 +	42,717	27.245	2.732	24	35
Survey Year	104,836	2,010.148	4.102	1,998	2,017
Year of Birth	104,836	1,987.145	4.157	1,982	2,001
PISA Math Score	104,836	−0.116	0.497	−1.810	1.029
TIMSS Grade 8 Math Score (PISA Scale)	104,836	−0.308	0.456	−2.019	0.779
Years of Education	90,852	13.122	3.036	0	22
Complete Secondary School |19 +	89,441	0.820	0.384	0	1
Complete Bachelors |24 +	42,717	0.337	0.473	0	1
Percentile of HH Income |24 +	31,403	52.888	27.970	0	100
Source: Afrobarometer	104,836	0.005	0.072	0	1
Source: Americas Barometer	104,836	0.008	0.089	0	1
Source: Arab Barometer	104,836	0.017	0.129	0	1
Source: Asia Europe Survey	104,836	0.003	0.051	0	1
Source: Asian Barometer	104,836	0.028	0.164	0	1
Source: Caucasus Barometer	104,836	0.006	0.079	0	1
Source: Comparative National Elections Project	104,836	0.009	0.096	0	1
Source: Consolidation of Democracy	104,836	0.001	0.025	0	1
Source: Eurobarometer	104,836	0.001	0.028	0	1
Source: European Quality of Life Survey	104,836	0.046	0.209	0	1
Source: European Social Survey	104,836	0.246	0.431	0	1
Source: European Values Study	104,836	0.028	0.166	0	1
Source: International Social Survey Programme	104,836	0.427	0.495	0	1
Source: Latinobarometro	104,836	0.037	0.189	0	1
Source: Life in Transition Survey	104,836	0.054	0.226	0	1
Source: New Baltic Barometer	104,836	0.001	0.037	0	1
Source: New Europe Barometer	104,836	0.006	0.074	0	1
Source: World Values Survey	104,836	0.078	0.268	0	1

Note: Table displays summary statistics for PIAAC and SDR data in Panels A and B, respectively. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2.

Table 2: Relationship Between Adulthood Numeracy and Adolescent Test Scores by Country Birth Cohort

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable: PIAAC Numeracy Test Score						
PISA	0.205+		0.158	0.229		0.185
	(0.109)		(0.093)	(0.133)		(0.117)
TIMSS (Grade 8, PISA Scale)		0.256*	0.204*		0.277*	0.235*
		(0.115)	(0.083)		(0.117)	(0.088)
Num.Obs.	31433	31433	31433	31433	31433	31433
R2	0.141	0.142	0.142	0.142	0.142	0.143
p-value: PISA = TIMSS	-	0.417	0.585	-	0.631	0.670
Country FEs	✓	✓	✓	✓	✓	✓
Age and Age^2	✓	✓	✓	✓	✓	✓
Gender FEs	✓	✓	✓	✓	✓	✓
Test Year FEs	✓	✓	✓	✓	✓	✓
Region-by-Age and Region-by-Age^2				✓	✓	✓

Note: Table displays regression results estimating the relationship between average PISA and TIMSS math scores and PIAAC numeracy scores. In the table, PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Robust standard errors clustered at the country and birth year level and adjusted for multiple imputation using Rubin's rule in parentheses. Observations are weighted by $w_{ict} / \sum_{i \in ct} w_{ict}$, where w_{ict} is individual i in country c in year t's final sampling weight, and $\sum_{i \in ct} w_{ict}$ denotes the sum of sampling weights in country c in year t. p-values shown in Columns 2 and 5 reflect the results of a test of equality of coefficients for the estimates in Columns 1 and 2, and Columns 4 and 5, respectively. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3: Relationship Between Adulthood Outcomes and Adolescent Test Scores by Country Birth Cohort

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Years of Education; Sample: 16+ Years Old									
PISA	0.882*		0.855*	0.641		0.636	1.024+		1.033+
	(0.358)		(0.346)	(0.443)		(0.427)	(0.555)		(0.538)
TIMSS (Grade 8, PISA Scale)		0.270	0.141		0.114	0.028		0.049	-0.060
		(0.180)	(0.126)		(0.202)	(0.157)		(0.244)	(0.186)
Num.Obs.	90852	90852	90852	90852	90852	90852	90852	90852	90852
R2	0.271	0.270	0.271	0.274	0.274	0.274	0.296	0.295	0.296
p-value: PISA = TIMSS	-	0.034	0.034	-	0.091	0.093	-	0.020	0.015
Panel B: Complete Secondary School; Sample: 19+ Years Old									
PISA	0.126*		0.119*	0.110+		0.105+	0.117*		0.115*
	(0.054)		(0.049)	(0.059)		(0.054)	(0.044)		(0.041)
TIMSS (Grade 8, PISA Scale)		0.064	0.052		0.050	0.039		0.028	0.017
		(0.049)	(0.037)		(0.045)	(0.034)		(0.033)	(0.026)
Num.Obs.	89441	89441	89441	89441	89441	89441	89441	89441	89441
R2	0.099	0.098	0.099	0.102	0.102	0.103	0.132	0.131	0.132
p-value: PISA = TIMSS	-	0.230	0.200	-	0.187	0.165	-	0.015	0.009
Panel C: Complete Bachelors; Sample: 24+ Years Old									
PISA	-0.026		-0.026	-0.013		-0.014	0.000		-0.002
	(0.069)		(0.069)	(0.067)		(0.066)	(0.059)		(0.059)
TIMSS (Grade 8, PISA Scale)		0.030	0.030		0.029	0.029		0.015	0.015
		(0.022)	(0.022)		(0.027)	(0.026)		(0.021)	(0.021)
Num.Obs.	42717	42717	42717	42717	42717	42717	42717	42717	42717
R2	0.079	0.079	0.079	0.081	0.081	0.081	0.108	0.108	0.108
p-value: PISA = TIMSS	-	0.477	0.460	-	0.563	0.542	-	0.802	0.783
Panel D: Percentile of Household Income; Sample: 24+ Years Old									
PISA	5.856+		6.007*	5.764*		5.668*	10.663**		10.300**
	(2.763)		(2.619)	(2.248)		(2.150)	(3.070)		(2.912)
TIMSS (Grade 8, PISA Scale)		5.016***	5.098**		4.771***	4.719**		3.683*	3.207+
		(1.126)	(1.299)		(1.067)	(1.253)		(1.638)	(1.757)
Num.Obs.	31403	31403	31403	31403	31403	31403	31403	31403	31403
R2	0.083	0.083	0.084	0.085	0.086	0.086	0.103	0.102	0.103
p-value: PISA = TIMSS	-	0.807	0.776	-	0.742	0.742	-	0.013	0.004
Country FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Age and Gender FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Survey Year FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Survey Wave FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Region-by-Age FEs				✓	✓	✓	✓	✓	✓
Country-by-Survey Year FEs							✓	✓	✓

Note: Table displays regression results estimating the relationship between average PISA and TIMSS math scores and adulthood outcomes in SDR data. In the table, PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Robust standard errors clustered at the country and birth year level in parentheses. Observations are weighted by SDR-provided weights. p-values shown in Columns 2, 5, and 8 reflect the results of a test of equality of coefficients for the estimates in Columns 1 and 2, Columns 4 and 5, and Columns 7 and 8, respectively. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

A Example PISA and TIMSS Test Questions

Publicly-available documents with sample PISA, TIMSS, and PIAAC assessment questions are available at the links below:

- OECD: Sample PIAAC questions and questionnaire
- NCES: TIMSS Released Assessment Questions
- OECD: PISA Test Questions

B Additional Tables and Figures

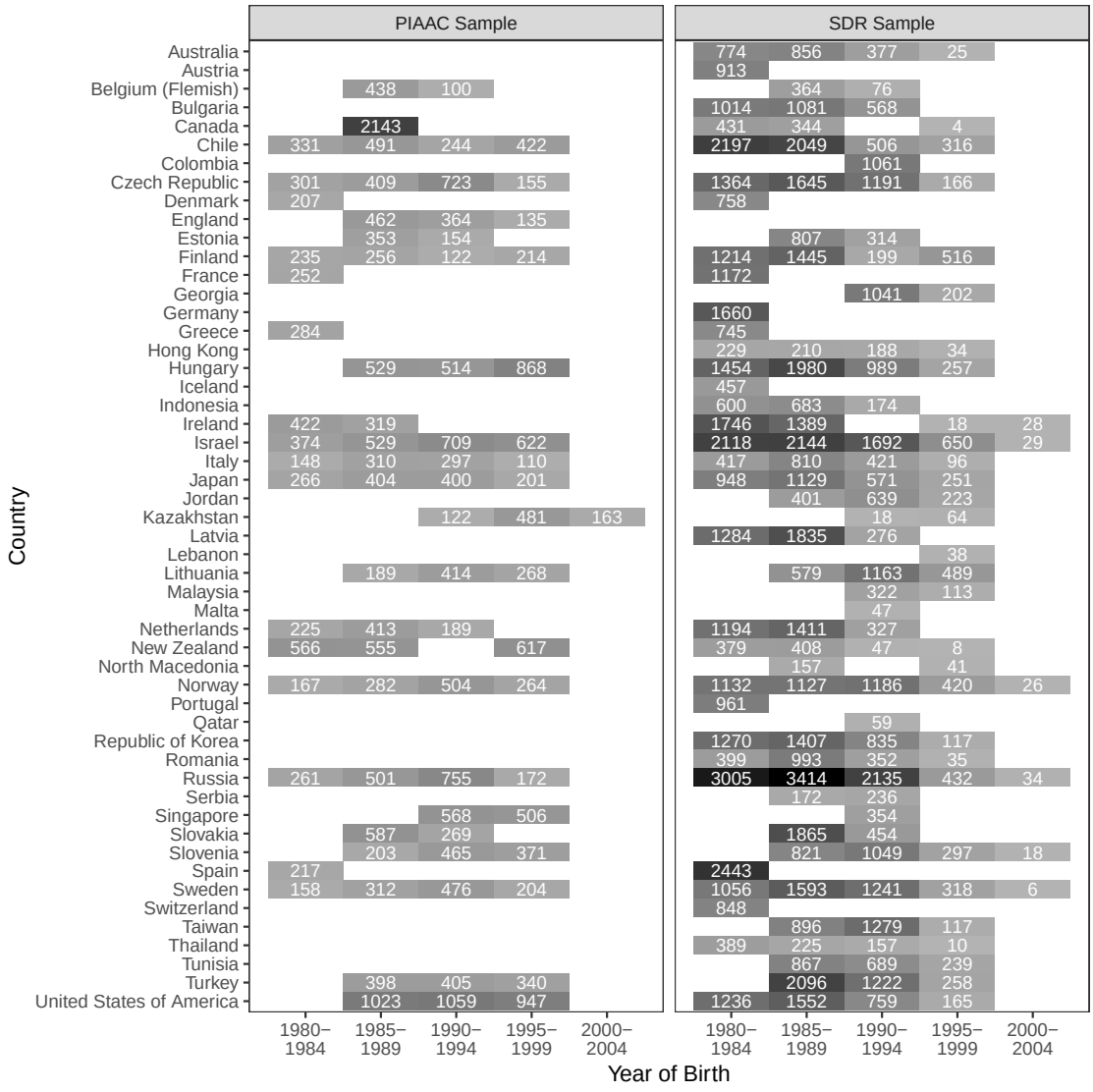


Figure B.1: Data Coverage by Country and Year of Birth

Note: Figure displays the number of observations in each country-by-year of birth cell, separately for the PIAAC and SDR samples.

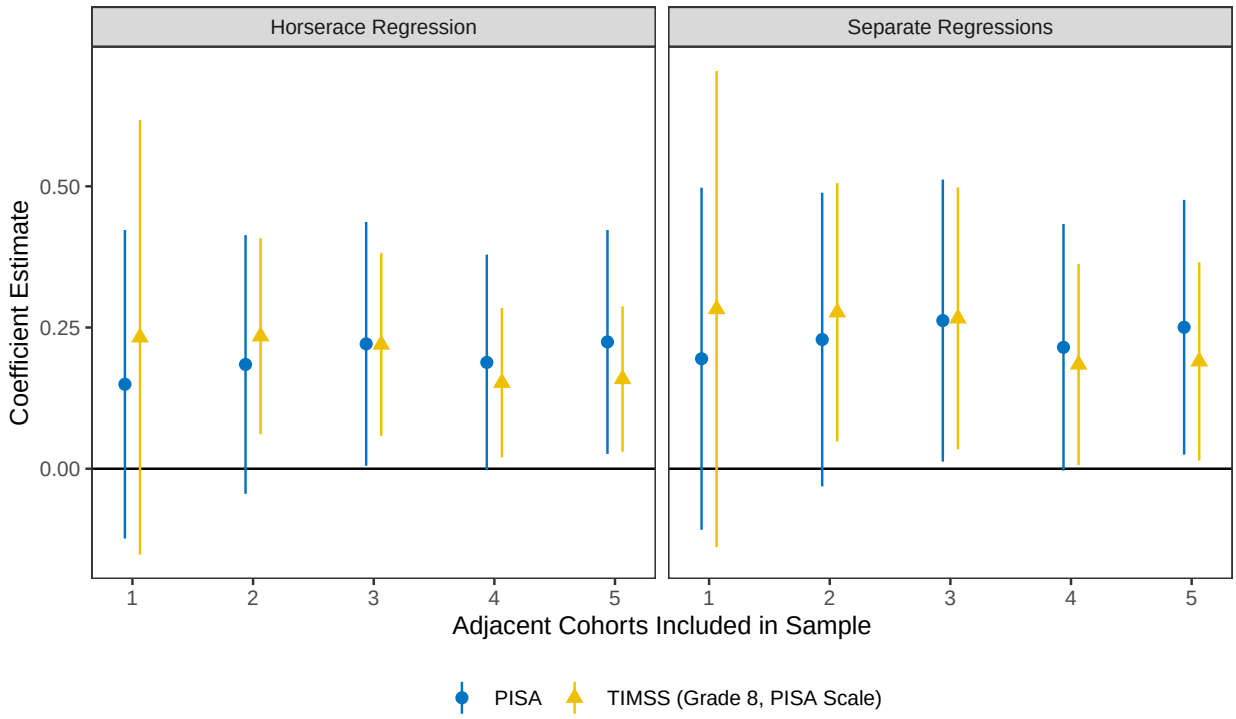


Figure B.2: Relationship Between Adulthood Numeracy and Adolescent Test Scores by Country Birth Cohort: Coefficient Estimates and Adjacent Cohorts Included in Sample

Note: Figure displays regression coefficients estimating the relationship between average PISA and TIMSS math scores and PIAAC numeracy scores. PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Plotted coefficients show the estimated effect of PISA and TIMSS scores in data that includes the specified number of adjacent cohorts. The "Horserace Regression" panel summarizes the results of 5 separate regressions; each regression includes PISA and TIMSS as independent variables. The "Separate Regressions" panel summarizes the results of 10 separate regressions; each regression includes either PISA or TIMSS as an independent variable. All regressions include controls for age and age-squared, fixed effects for gender, test year, and region-age and region-age-squared interactions. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Error bars show 95% confidence intervals. Observations are weighted by $w_{ict} / \sum_{i \in ct} w_{ict}$, where w_{ict} is individual i in country c in year t's final sampling weight, and $\sum_{i \in ct} w_{ict}$ denotes the sum of sampling weights in country c in year t.

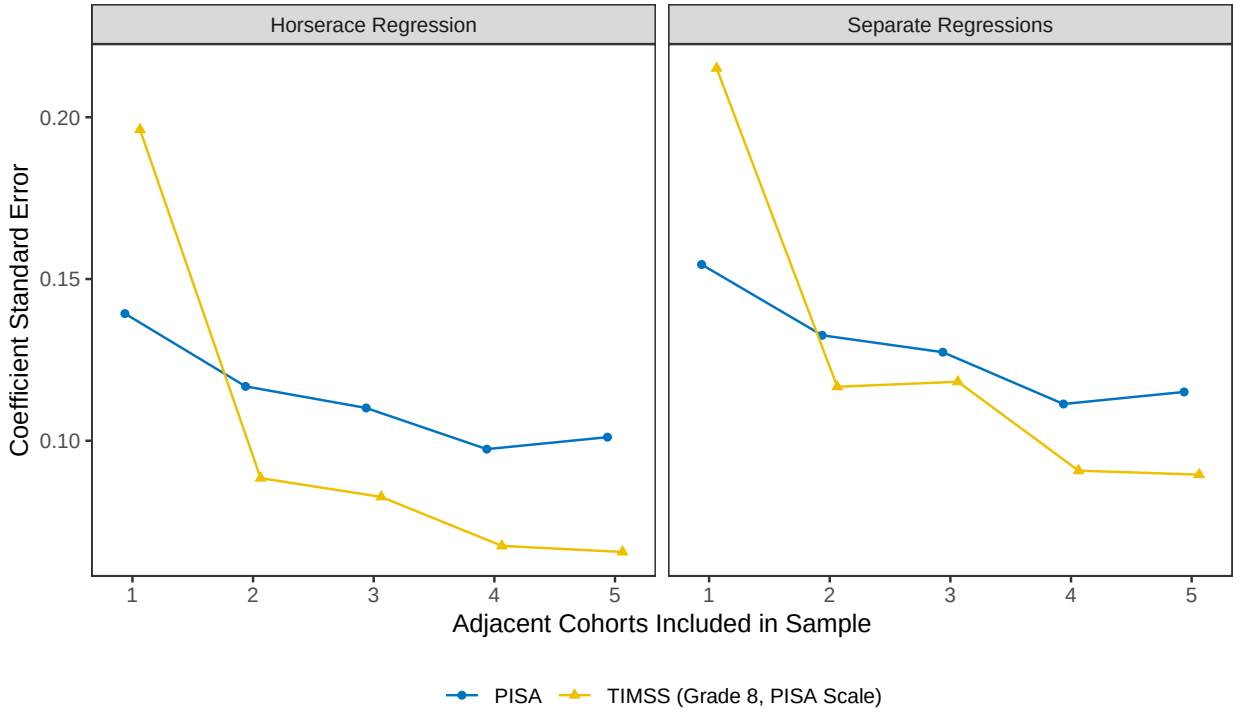


Figure B.3: Relationship Between Adulthood Numeracy and Adolescent Test Scores by Country Birth Cohort: Coefficient Standard Errors and Adjacent Cohorts Included in Sample

Note: Figure displays standard errors for coefficients estimating the relationship between average PISA and TIMSS math scores and PIAAC numeracy scores. PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Plotted standard errors show the standard errors associated with the estimated effect of PISA and TIMSS scores in data that includes the specified number of adjacent cohorts. The "Horserace Regression" panel summarizes the results of 5 separate regressions; each regression includes PISA and TIMSS as independent variables. The "Separate Regressions" panel summarizes the results of 10 separate regressions; each regression includes either PISA or TIMSS as an independent variable. All regressions include controls for age and age-squared, fixed effects for gender, test year, and region-age and region-age-squared interactions. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Observations are weighted by $w_{ict} / \sum_{i \in ct} w_{ict}$, where w_{ict} is individual i in country c in year t's final sampling weight, and $\sum_{i \in ct} w_{ict}$ denotes the sum of sampling weights in country c in year t.

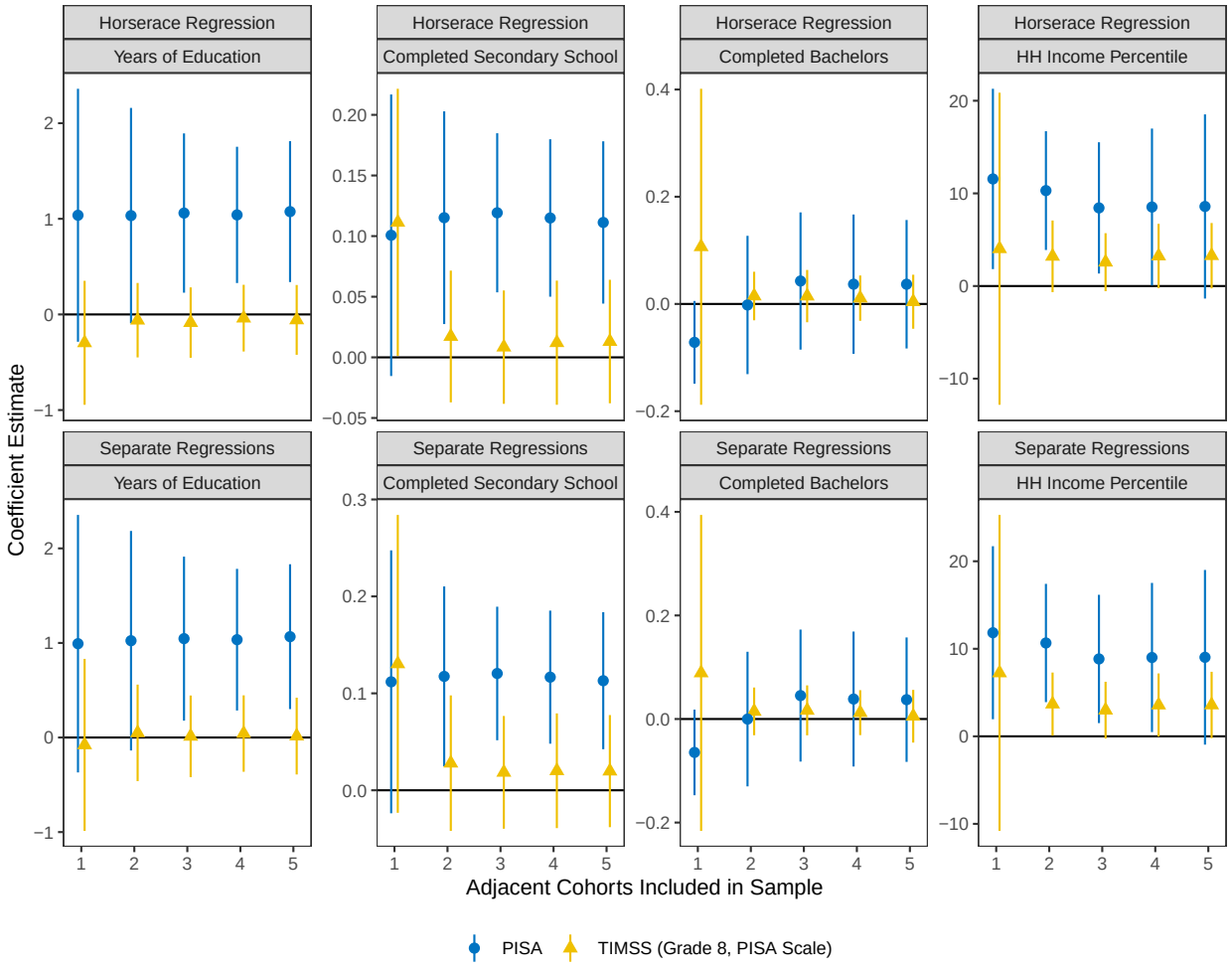


Figure B.4: Relationship Between Adulthood Outcomes and Adolescent Test Scores by Country Birth Cohort: Coefficient Estimates and Adjacent Cohorts Included in Sample

Note: Figure displays regression coefficients estimating the relationship between average PISA and TIMSS math scores and adulthood outcomes in SDR data. PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Plotted coefficients show the estimated effect of PISA and TIMSS scores in data that includes the specified number of adjacent cohorts. Each "Horserace Regression" panel summarizes the results of 5 separate regressions; each regression includes PISA and TIMSS as independent variables. Each "Separate Regressions" panel summarizes the results of 10 separate regressions; each regression includes either PISA or TIMSS as an independent variable. All regressions include fixed effects for age, gender, survey year, survey wave, region-by-age, and country-by-survey year. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Error bars show 95% confidence intervals. Observations are weighted by SDR-provided weights.

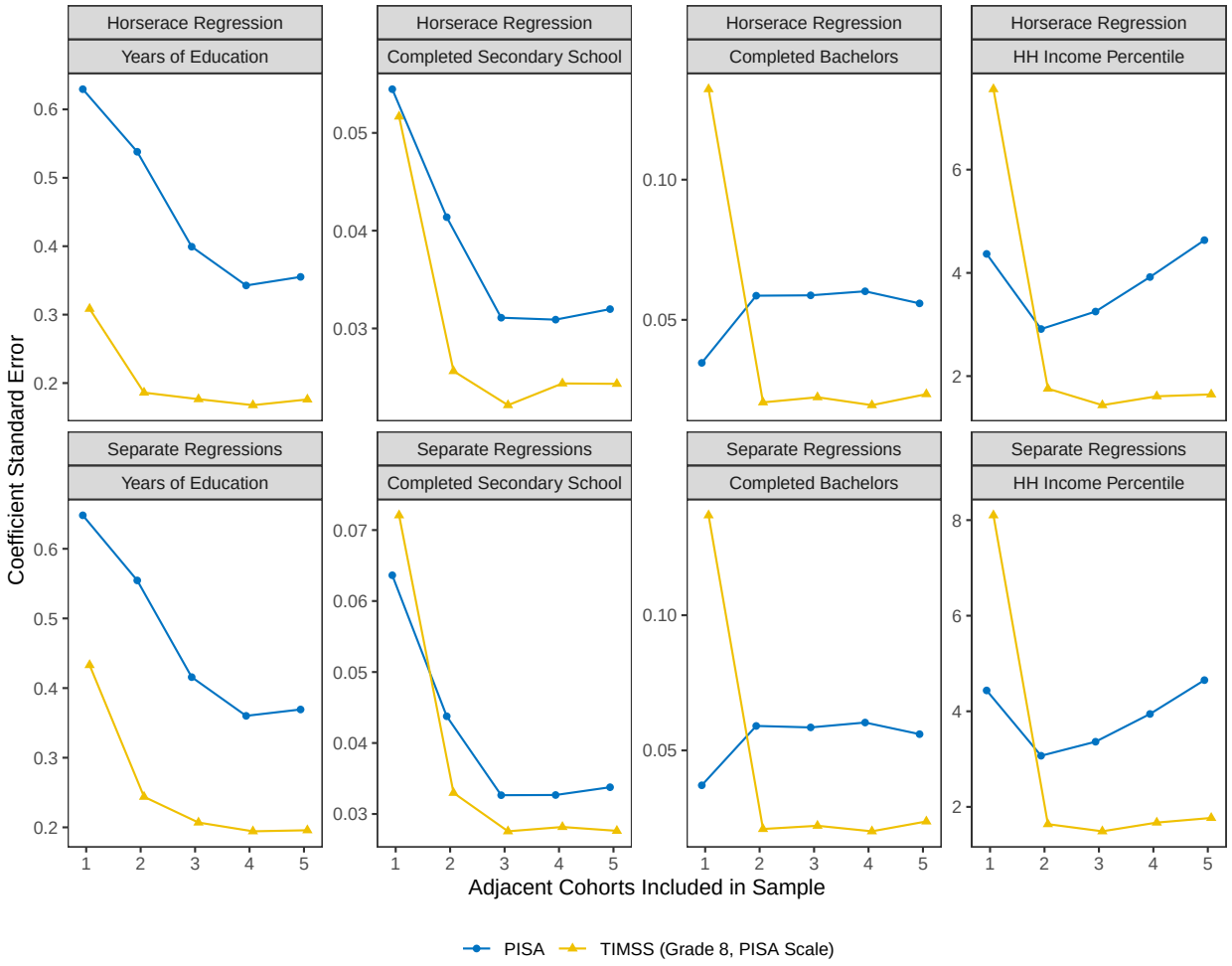


Figure B.5: Relationship Between Adulthood Outcomes and Adolescent Test Scores by Country Birth Cohort: Coefficient Standard Errors and Adjacent Cohorts Included in Sample

Note: Figure displays standard errors for coefficients estimating the relationship between average PISA and TIMSS math scores and adulthood outcomes in SDR data. PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Plotted standard errors show the standard errors associated with the estimated effect of PISA and TIMSS scores in data that includes the specified number of adjacent cohorts. Each "Horserace Regression" panel summarizes the results of 5 separate regressions; each regression includes PISA and TIMSS as independent variables. Each "Separate Regressions" panel summarizes the results of 10 separate regressions; each regression includes either PISA or TIMSS as an independent variable. All regressions include fixed effects for age, gender, survey year, survey wave, region-by-age, and country-by-survey year. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Observations are weighted by SDR-provided weights.

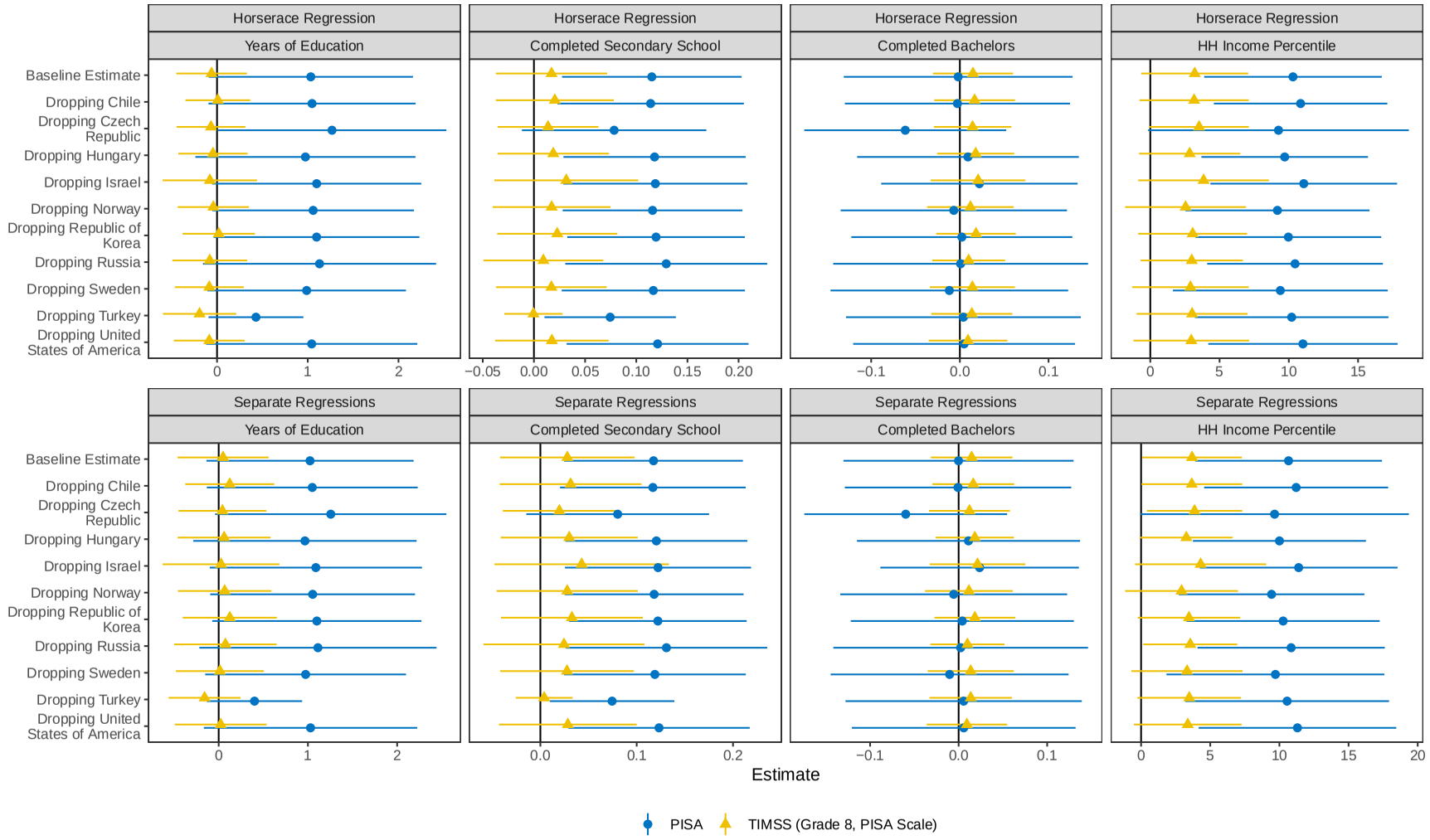


Figure B.6: Relationship Between Adulthood Outcomes and Adolescent Test Scores by Country Birth Cohort: Sensitivity to Country Omission

Note: Figure displays regression coefficients estimating the relationship between average PISA and TIMSS math scores and adulthood outcomes in SDR data. In each panel, "Baseline Estimate" reflects the estimate shown in Table 3 Columns 7 and 8 (for separate regressions) and 9 (for horserace regressions). All other rows show equivalent coefficients after dropping one country from the data. PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Each "Horserace Regression" panel summarizes the results of 11 separate regressions; each regression includes PISA and TIMSS as independent variables. Each "Separate Regressions" panel summarizes the results of 22 separate regressions; each regression includes either PISA or TIMSS as an independent variable. All regressions include fixed effects for age, gender, survey year, survey wave, region-by-age, and country-by-survey year. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Error bars show 95% confidence intervals. Observations are weighted by SDR-provided weights.

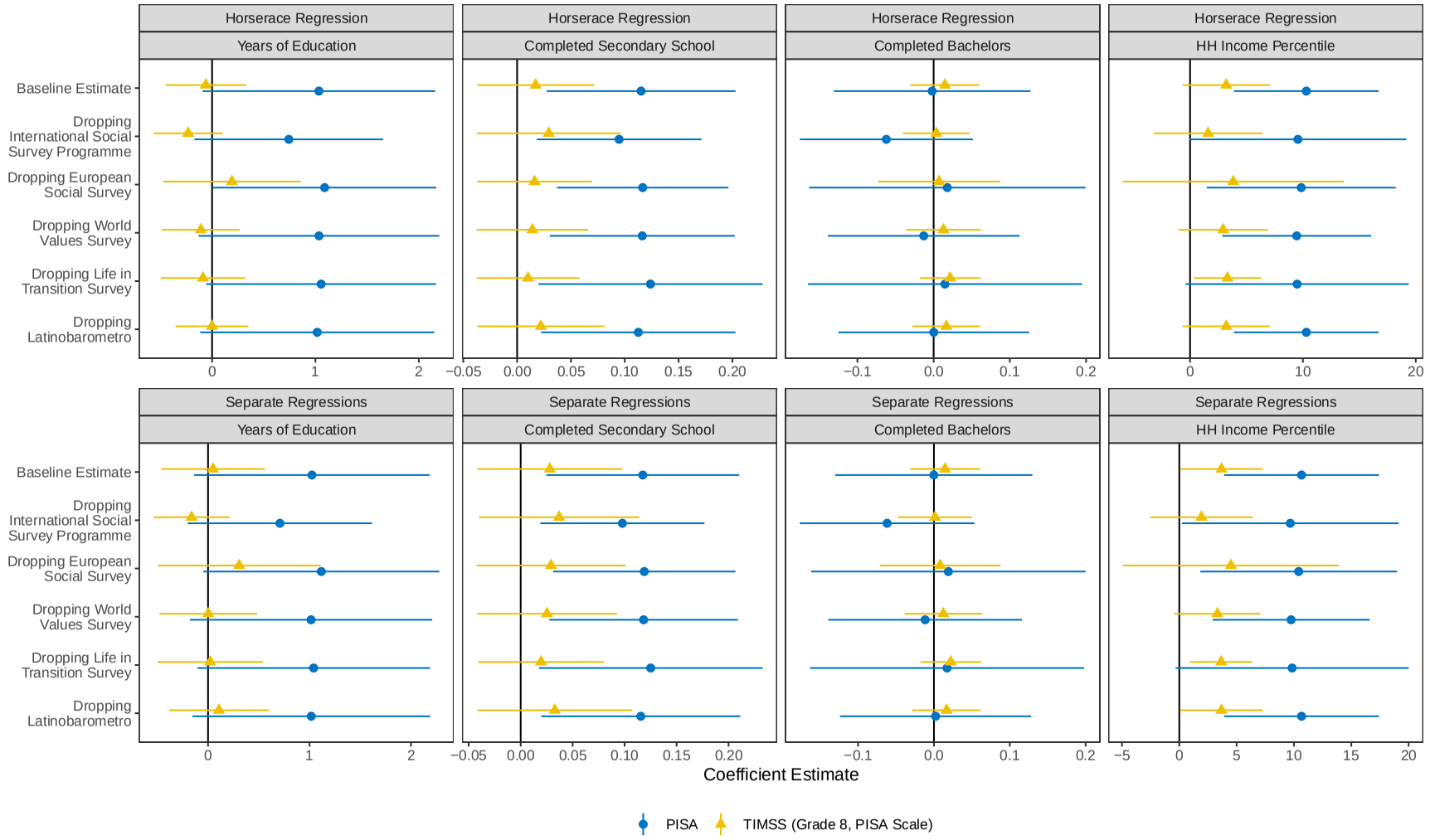


Figure B.7: Relationship Between Adulthood Outcomes and Adolescent Test Scores by Country Birth Cohort: Sensitivity to Survey Omission

Note: Figure displays regression coefficients estimating the relationship between average PISA and TIMSS math scores and adulthood outcomes in SDR data. In each panel, "Baseline Estimate" reflects the estimate shown in Table 3 Columns 7 and 8 (for separate regressions) and 9 (for horserace regressions). All other rows show equivalent coefficients after dropping one survey from the data. PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Each "Horserace Regression" panel summarizes the results of 6 separate regressions; each regression includes PISA and TIMSS as independent variables. Each "Separate Regressions" panel summarizes the results of 12 separate regressions; each regression includes either PISA or TIMSS as an independent variable. All regressions include fixed effects for age, gender, survey year, survey wave, region-by-age, and country-by-survey year. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Error bars show 95% confidence intervals. Observations are weighted by SDR-provided weights.

Table B.1: Secular Test Score Growth in PISA versus TIMSS

	(1)	(2)	(3)
Dependent Variable: PISA or TIMSS Test Score			
Year	-0.011**	-0.001	
	(0.004)	(0.002)	
Year × Test = TIMSS (Grade 8, PISA Scale)	0.012*	0.013***	0.014***
	(0.005)	(0.003)	(0.003)
Num.Obs.	3333105	3333105	3333105
R2	0.019	0.350	0.362
Test FEs	✓	✓	✓
Country FEs		✓	✓
Country Time Trend			✓

Note: Table displays regression results estimating differences in over-time growth in PISA and Grade 8 TIMSS math scores. Data is stacked student-level test results from PISA and Grade 8 TIMSS math assessments. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Robust standard errors clustered at the country level and adjusted for multiple imputation using Rubin's rule in parentheses. Observations are weighted by $w_{ict} / \sum_{i \in ct} w_{ict}$, where w_{ict} is individual i in country c for test t's sampling weight, and $\sum_{i \in c} w_{ict}$ denotes the sum of sampling weights in country c and test t. Here, test t refers to a test (PISA or Grade 8 TIMSS) and year (e.g. 1999, 2000, etc.) combination. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table B.2: Sensitivity of Income Estimates to Controls for Household Composition

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Main Income Estimates									
PISA	5.856+		6.007*	5.764*		5.668*	10.663**		10.300**
	(2.763)		(2.619)	(2.248)		(2.150)	(3.070)		(2.912)
TIMSS (Grade 8, PISA Scale)		5.016***	5.098**		4.771***	4.719**		3.683*	3.207+
		(1.126)	(1.299)		(1.067)	(1.253)		(1.638)	(1.757)
Num.Obs.	31403	31403	31403	31403	31403	31403	31403	31403	31403
R2	0.083	0.083	0.084	0.085	0.086	0.086	0.103	0.102	0.103
p-value: PISA = TIMSS	-	0.807	0.776	-	0.742	0.742	-	0.013	0.004
Panel B: Main Income Estimates with Household Size Fixed Effects									
PISA	5.229+		5.315	5.223*		5.186+	10.238*		9.986*
	(2.886)		(2.982)	(2.366)		(2.443)	(3.417)		(3.328)
TIMSS (Grade 8, PISA Scale)		4.462**	4.509**		3.876*	3.853*		2.846	2.404
		(1.291)	(1.396)		(1.474)	(1.564)		(1.744)	(1.720)
Num.Obs.	26887	26887	26887	26887	26887	26887	26887	26887	26887
R2	0.174	0.174	0.174	0.176	0.176	0.176	0.192	0.192	0.193
p-value: PISA = TIMSS	-	0.807	0.820	-	0.742	0.698	-	0.013	0.008
Country FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Age and Gender FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Survey Year FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Survey Wave FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Region-by-Age FEs				✓	✓	✓	✓	✓	✓
Country-by-Survey Year FEs							✓	✓	✓

Note: Table displays regression results estimating the relationship between average PISA and TIMSS math scores and percentile of household income in SDR data. Panel A does not include any controls for household composition. Panel B includes fixed effects for household size. In the table, PISA and TIMSS represent average PISA and TIMSS scores at the country-by-birth cohort level. Both PISA and TIMSS scores are transformed to the PISA scale based on the procedure in Gust et al. (2024) after incorporating the sample restrictions described in Section 2. Region refers to World Bank regions: East Asia and Pacific, Europe and Central Asia, Latin America and Caribbean, Middle East and North Africa, and North America. Robust standard errors clustered at the country and birth year level in parentheses. Observations are weighted by SDR-provided weights. p-values shown in Columns 2, 5, and 8 reflect the results of a test of equality of coefficients for the estimates in Columns 1 and 2, Columns 4 and 5, and Columns 7 and 8, respectively. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.