A Rosetta Stone for Human Capital*

Dev Patel Justin Sandefur

December 9, 2019

Abstract

International comparisons of human capital figure prominently in economists' explanations of poverty, economic growth, and trade patterns. But how can we accurately measure the global distribution of skills when countries do not take the same tests? We develop a new methodology to non-parametrically link scores from distinct populations. By administering an exam combining items from different assessments to 2,300 primary students in India, we estimate conversion functions among four of the world's largest standardized tests spanning 80 countries. Armed with this learning "Rosetta Stone", we revisit various well-known results, showing, *inter alia*, that learning differences between most- and least-developed countries are larger than existing estimates suggest. Applying our translations to microdata, we match pupils' socio-economic status to moments of the global income distribution and document several novel facts: (i) students with the same household income score significantly higher if they live in richer countries; (ii) the income-test score gradient is steeper in countries with greater income inequality; (iii) girls read better than boys at all incomes but only outperform them in mathematics at the lowest deciles of the global income distribution; and (iv) the test-score gap between public and private schools increases with inequality, partially due to a rise in socio-economic sorting across school types.

Keywords: Human Capital, Education Quality, Learning Assessments, India
JEL Classification Numbers: I25, J24, O15, O53

*Patel: Harvard University, Department of Economics (devpatel@fas.harvard.edu). Sandefur: Center for Global Development (jsandefur@cgdev.org). This project could not have happened without the help of our CGD colleague Anit Mukherjee. Maryam Akmal, Ben Crisman, and Jen Richmond provided excellent research assistance. The ideas here benefited from discussions with Robert Barro, Rukmini Banerji, Luis Crouch, Claudia Goldin, Emma Harrington, Andrew Ho, Beth King, Michael Kremer, Nathan Nunn, Lant Pritchett, Evan Soltas, Liesbet Steer, and seminar participants at the Harvard Graduate Student Development Workshop and the Association for Public Policy Analysis and Management Fall Research Conference. This research was supported by the International Commission on Financing Global Education Opportunity, and Patel acknowledges support from a National Science Foundation Graduate Research Fellowship under grant DGE1745303. The views expressed are those of the authors and should not be attributed to the board of directors or funders of the Center for Global Development. All errors are our own.

1 Introduction

How much do differences in human capital explain patterns of economic growth, trade, and inequality around the world? Empirical research attempting to answer this question has struggled with a fundamental fact: globally comparable measures of skill do not exist. Even the largest international standardized tests cover less than a third of school-age children and currently exclude all students in low-income countries. Most of the developing world participates in regional assessments lacking a comparable scale, as shown in Figure 1. Economists consequently often rely on measures of educational enrollment (Barro, 1991; Mankiw et al., 1992; Hendricks and Schoellman, 2018), yet the quality of a school year varies tremendously across countries (Behrman and Birdsall, 1983; Singh, 2019). This is a major impediment to academic research as well as basic policy tasks like tracking progress on the United Nation's Sustainable Development Goals.

This paper develops a new method to compare human capital across different populations that take different tests. The intuition of our approach is simple. We take a single sample of students and give them questions from each major exam. By grading each child's responses on the original test scales, we calculate scores on *different* exams for the *same* child on the *same* day. Applying these relationships out of sample gives us a flexible way to convert scores for any other student who only took a single test.

A key input of this method is a hybrid test we create to link exams having no overlapping questions and that were administered to disjoint populations. This new assessment is comprised of all publicly available items from the four of the world's largest primary school standardized tests.[1] We administer this combined exam to 2,300 primary school students across 51 schools in Bihar, India, in 2016. Using statistical methods from item response theory (IRT), we grade these questions on the original scales. The resulting data set provides a bridge to link the difficulty and discrimination of items on disparate tests together onto a

[1]We link the following tests: Trends in International Mathematics and Science (TIMSS), Progress in International Reading Literacy Study (PIRLS), Latin American Laboratory for Assessment of the Quality of Education (LLECE), and Analysis Program of the CONFEMEN Education Systems (PASEC).

common scale: a "Rosetta Stone" for regional learning assessments.

To estimate our conversion functions, we extend a method from psychometrics for linking test scales from separate populations who take different tests called the non-equivalent groups with anchor items framework (Braun and Holland, 1982; Kolen and Brennan, 2014; Steinmann et al., 2014). By this approach, we retrospectively link existing learning assessments that lack any overlapping questions. In our application, there are no students who sit both, say, the grade-three Latin American test and the grade-six West African test. But the students in our sample sit an exam that is anchored to both scales. By repeated application of this logic, we can create a conversion function between any two independent tests. Bootstrapping our procedure allows us to estimate confidence intervals on the translated scores.

The most demanding statistical assumptions required for linking are already embedded in the scoring approach of the original tests. Similar to college applicants taking the Standardized Aptitude Test (SAT), individual pupils in the assessments we examine see booklets with distinct but overlapping sets of questions (items). To grade performance on a common scale across booklets, these tests use an IRT framework that already imposes strong requirements on items and the underlying latent ability to be measured. Specifically, the exams assume that individual student scores are parameters in a conditional logit model where correct answers are solely a function of pupil- and item-specific effects. These same assumptions that make it possible to compare student scores within a single IRT-based test also make it possible to link scores across multiple tests or subsequent rounds of a test so long as they contain overlapping items. For example, we must assume that administering questions in different languages does not impact their relative difficulty, which is already taken into consideration in the design of the TIMSS and PIRLS questions which were administered in 58 languages. We explore the robustness of our results to concerns raised in the psychometric literature about these assumptions and provide diagnostic tests of the validity of the links we estimate.

Our paper extends existing approaches to measuring learning across countries that have

focused on linking aggregate country results from separate assessments without relying on the underlying psychometric structure or often even the microdata. The seminal work of Hanushek and Kimko (2000) developed the dominant method to combine scores from two different tests by exploiting "doubloon cells": countries which administered both assessments to a representative sample of pupils within a reasonable time frame. Hanushek and Woessmann (2012) extend this technique by adding an assumption that the cross-country variance in national test-score means among a select group of high-income OECD countries is constant over time. Barro and Lee (2001), Altinok and Murseli (2007), and especially Angrist et al. (2019) link using other countries to dramatically expand coverage, and Altinok et al. (2018) adopt a similar method while also using equipercentile matching for some tests.

We build on these results by taking a microdata approach, essentially turning every student who sits our combined exam into a "doubloon cell." Previous work has been constrained by the few countries which sit multiple exams at approximately the same time, so assessments without any overlap cannot be linked. Our methodology, by contrast, can be used to link *any* tests together, regardless of the coverage or timing of the original exams.[2] Traditional linking functions often bridge assessments administered in different years, to different cohorts, at different grade levels. Since we rely exclusively on pupil-level comparisons, our comparisons across tests are based on data from the same pupil on the same day. The standard doubloon approach also relies on the comparison of country averages, requiring strong functional form assumptions beyond those embedded in the assessments themselves about the underlying distribution of each test. Our method imposes no such assumptions about the conversion functions, allowing for a flexible, non-parametric relationship between test scales across the full ability distribution.

Equipped with this "Rosetta Stone," we convert the test scores of 628,587 students from

[2]Using similar psychometric techniques as our approach among tests that share common items, Das and Zajonc (2010) places the test scores for two Indian states onto the TIMSS math scale. Sandefur (2018) exploits overlap between the Southern African exam and TIMSS items to link scores in Southern Africa. We build on these by administering a new hybrid exam in which we do not have to rely on tests sharing questions *ex ante*, allowing us to link many more countries on both math and reading scales.

80 countries onto common math and reading scales. We first present new estimates of cross-country differences in test scores. Existing human capital measures based on the quantity of schooling fail to predict education quality, explaining less than a third of variation in test scores. Comparing our estimates to the dominant approach to linking test scores in the literature, we estimate significantly different proportions of students meeting international learning benchmarks, particularly for pupils in the poorest countries who perform worse by our measure.

To examine the macroeconomic implications of test scores, we study how exports vary by learning achievement. Aggregating goods to the industry level using the approach in Autor et al. (2013) and Autor et al. (Forthcoming) and measuring skill intensity by industry in U.S. census data, we find that countries with higher test scores export more in skill-intensive industries. This is particularly true for math scores and industries whose occupations disproportionately use quantitative skills.

Extending the sample of countries also sheds new light on the relationship between government education spending per pupil and test scores around the world. This relationship is positive and significant with and without controlling for the log of per capita GDP but also highly concave: steep for low- and lower-middle income countries and relatively flat for upper-middle and high-income countries. While causal inference must rest on better identified microeconometric studies (Jackson et al., 2015; Muralidharan et al., 2019), the cross-country pattern suggests the external validity of education expenditure studies in rich countries may be limited when inferring lessons for the developing world.

Our methodology allows us to delve much deeper than the country averages that have dominated existing work. To compare socio-economic gradients on a meaningful scale, we match the moments of the distribution of pupils' socio-economic status from surveys conducted alongside international learning assessments to the global distribution of per capita income and consumption as reported by Lakner and Milanovic (2016).

At the pupil-level, test scores increase fairly linearly with log household income. This re-

lationship masks considerable heterogeneity within countries, in both the slope and intercept. National income matters. Students with similar household income in real purchasing-power parity dollars post dramatically different scores across countries. For example, an Argentine pupil from a household with an income of $8,000 (PPP) per capita is predicted to score nearly 100 points lower in mathematics on the TIMSS scale than an Italian student with equivalent purchasing power. Inequality also matters. While learning is positively associated with parental income in almost all countries, this learning-income gradient is significantly steeper in countries with a higher Gini coefficient of income inequality, consistent with the literature pointing to the importance of horizontal inequality in explaining intergenerational mobility Krueger (2012); Corak (2013).

Combining microdata on a common global scale also allows us to examine how other dimensions of inequality in test scores varies by country, including gender disparities and the public-private school gap. Lower test scores for girls are strongly correlated with higher rates of child marriage and desired fertility. Girls read better than boys on average across the global income distribution but only score better in math in the world's poorer households. Turning to school type, the test-score difference between private and government schools is significantly larger in more unequal countries when measured on a comparable scale. Plausible corrections for selection into private institutions suggest much of this gap is due to segregation along socio-economic status between pupils in government and private schools.

This paper proceeds as follows. Section 2 describes our empirical approach and psychometric estimation exploiting features of item-response theory. Section 3 describes the data collection for our hybrid anchor test and presents a series of diagnostic tests to assess the reliability of the key set of assessment scale conversion functions. Section 4 presents the main empirical results on human capital around the world, and section 5 concludes.

2 Psychometric Approach

2.1 Converting Scores to a Common Scale

To construct a link among separate international and regional assessments, we adopt a non-equivalent groups with anchor test (NEAT) design (Kolen and Brennan, 2014), combining questions from multiple source tests and administering them to a common set of examinees. The 2,314 students in the our sample took hybrid exams that pulled all of the publicly available questions from several major assessments: TIMSS, PIRLS, PASEC, and LLECE. For each student in our sample, we estimate the score that pupil would earn on each reference test, allowing us to create a common link. To ensure equivalence between the individual scores on our hybrid exam and the reference scale, we exploit the fact that all major international tests use item response theory (IRT) to grade student responses.

The key assumption used in grading these assessments is that for a given student i, the probability of answering question j correctly depends only on the student's ability θ_i and a set of item-specific parameters $a_j, b_j, c_j, d_{j,n} \in p_j$. Throughout our analysis, we estimate math and reading aptitude separately. Implicit in all models is a monotonicity assumption that as ability increases, the probability of answering a given question correctly increases as well. Let x_{ij} be the response of student i to item j. Most items are dichotomously scored ($x_{ij} \in \{0, 1\}$) multiple-choice questions and use either a one-, two-, or three-parameter logistic model, the latter of which is shown in equation 1.

$$P(x_{ij} = 1|\theta_i, a_j, b_j, c_j) = c_j + \frac{(1 - c_j)}{1 + \exp\left(a_j * (\theta_i - b_j)\right)} \tag{1}$$

One-parameter (1PL) or two-parameter (2PL) models can be easily derived from the three-parameter (3PL) formula in equation 1 by setting $c_j = 0$ for the 2PL items and additionally fixing $a_j = -1$ for 1PL items. The PIRLS and TIMSS public items also include some polytomous questions with partial credit options ($x_{ij} \in \{0, 1, 2\}$). These items require a general partial credit model (GPCM). For each partial credit item j, the probability of a

student i scoring in the l^{th} level ($x_{ij} = l$) of m_j ordered score categories is shown in equation 2.

$$P(x_{ij} = l|\theta_i, a_j, b_j, d_{j,1}, ..., d_{j,m_j-1}) = \frac{\exp\left(\sum_{v=0}^{l} a_j\left(\theta_i - b_j + d_{j,v}\right)\right)}{\sum_{g=0}^{m_i-1}\exp\left(\sum_{v=0}^{g} a_j\left(\theta_i - b_j + d_{j,v}\right)\right)} \tag{2}$$

Given these models, known as item characteristic curves, and the corresponding item parameters $a_j, b_j, c_j, d_{j,n} \in p_j$, the joint probability of a vector of responses X_i to items $j \in J$ is given by equation 3.

$$P(\vec{x_{ij}}|\theta_i, p_j) = \prod_{j=1}^{J}\prod_{l=0}^{m_j-1} P_{i,l}\left(\theta_i\right)^{u_{l_j}} \tag{3}$$

In constructing this vector of student responses X_i, we grade questions that students leave blank as incorrect.[3] $P_{j,l}\left(\theta_i\right)$ follows the corresponding model specification from equation 1 and 2, with $m_j = 2$ for dichotomously scored items. The function $u_{l_j} = 1$ if $l_j \in X_i$ and 0 otherwise. The item-level parameters p_j are calculated by the creators of the original tests using maximum likelihood estimation of the probabilities in equation 3. We take these same parameters as fixed when estimating the score for a given test, and estimate the parameters for other items of the same type (math or reading). In doing so, we use information on student ability from non-source questions to inform the score on the source scale, all the while preserving the integrity of the link. We estimate a marginal maximum likelihood model that maps a vector of responses X_i from our new sample to the item characteristic curves in equations 1 and 2, outputting a θ_i for each student. Repeating this exercise holding fixed the item parameters from each source test provides a score mapping from each test to another. The intuition underlying this mapping is that since the same student is answering each question on the same day, their underlying ability is held constant. We flexibly estimate this function non-parametrically using a local linear regression.

[3] We drop 94 students who did not answer any of the questions in their exam.

2.2 Assumptions

Our empirical strategy rests on several well-known assumptions (Kolen and Brennan, 2014), each of which is at least partially testable in our data. First, the ability measured by the various international and regional exams must be constant across tests. This unidimensionality assumption is slightly stronger than that already built into the source questions. While each individual test relies on unidimensionality within the set of their items (that is, each math question in TIMSS is assumed to measure the same underlying mathematical ability θ_{TIMSS}), we must further have that these θs are consistent across tests (that is, θ_{TIMSS} measures the same latent ability as θ_{PASEC}, and so on.) In essence, this assumption says that an accurate conversion scale between these international exams exists. This is particularly plausible in our setting given the rudimentary level of these exams. To test for unidimensionality, we conduct factor analysis separately for math and reading.[4] The distribution of eigenvalues and resulting scree plot shown in Appendix Figure A.1 clearly satisfy the unidimensionality standards from the psychometric literature (Drasgow and Lissak, 1983).

Second, we assume conditional independence for all items, requiring that the probability of student i correctly answering item j depend only on aptitude θ_i and the logistic function parameters p_j of item j that come from the source exams. As for unidimensionality, the core of this assumption is already built into the items we pull from the source tests. In our context, the conditional independence assumption further requires that the item parameters which we pull from the source tests are population invariant. Therefore we assume that data collection conditions, demographic characteristics, and even language of administration do not impact the item characteristic curves specified in equations 1 and 2. From a theoretical perspective, these assumptions seem plausible because these items are already administered across many different contexts within each source test: for instance, the 2011 TIMSS and

[4]Since items were randomized across booklets, we calculate an expectation-maximization algorithm to the sparse item-wise covariance matrices for math and reading, and then conduct factor analysis on the resulting matrices. The nominal sample size in the maximization was set to the average number of students who sat each item.

PIRLS was administered in 58 languages. Thus these questions have already been designed such that they can suitably measure the same ability across heterogeneous settings.[5]

To empirically test this assumption, we construct differential item functioning (DIF) plots, as shown in Appendix Figures A.2 through A.5.[6] We cannot calculate DIF according to standard psychometric procedures because the micro-data for the reference populations is not available for all of our items. However, we can visually compare the DIF plots for the Bihar students versus the reference test populations, as modeled by the item characteristic functions for each item. While there are some items for which the pupils in our sample seem to over- and under-perform, in general, the items seem suited to the Indian students.

Finally, we must assume that the common-item sets adequately represent the content of the reference test. Hastedt and Desa (2015) argue that at least 30 items are needed to sufficiently minimize linking error, and even at that level of overlap, they highlight challenges. The possible number of common items we use is limited by the number of publicly released items—particularly with the LLECE tests—and overall test length. To test the robustness of our results to this concern, we bootstrap our θ estimates using 100 draws of the data in which each item is given a 75 percent chance of being included in the item response theory estimation of a given draw. The results of this intensive-margin check are encouraging: the distribution of test scores for a given student is relatively tight across the 1000 draws. The average standard deviation in test scores within student is about 20 points on each exam (relative to a 100-point standard deviation in the underlying score distribution.)

[5] One hypothetical approach to capturing conditional independence would be to estimate the item characteristic curve parameters on our sample and compare them to the original. This is unfortunately not feasible given the far fewer number of publicly available items and students in our sample, and thus the three parameter models do not even converge. Thus we defer to the parameters from the source exams which were estimated off of many tens—or even hundreds—of thousands of student responses.

[6] A similar analysis is conducted by the College Board to detect racial and gender bias in the SAT. By comparing the θs and observed probability of answering a question correctly for the relevant subgroup (in our case, the Bihar students) against the predicted item characteristic curve, one can identify items for which certain students deviate from expected performance, presumably due to some misunderstanding in question meaning.

2.3 Grade Adjustments

One important challenge in converting among international tests is that the official exams are administered to students in different grades in different countries. For the exams in our linking sample, PASEC is given to sixth-grade pupils, LLECE to third graders, and TIMSS/PIRLS primarily to fourth-grade students. This does not impact the validity of our psychometric method above but poses a practical issue because even after converting scores to a common scale, observed differences across countries could be due to true variation in learning quality or simply to differences in years of schooling. The existing approach in the literature has been to assume away these grade effects, treating all scores collected during primary school as equivalent across countries regardless of the grade of the students (Angrist et al., 2019).

Armed with our learning Rosetta Stone, we develop a novel approach to adjust for grade bias by taking advantage of the microdata from the doubloon countries in our sample that sit both LLECE and TIMSS/PIRLS. For Chile and Honduras in math and Colombia and Honduras in reading, we can apply our conversions to obtain learning outcomes on a common scale for different grades in the same country. We make a simple equipercentile assumption that the n^{th}-ranked student in grade A is also the n^{th}-ranked student in grade B. We then construct a score-per-grade adjustment for each ventile along the score distribution of these countries. Appendix Figure A.7 shows these score differences across the range of abilities. We find larger effects for third to fourth than fourth to sixth, which is consistent with education evidence showing that learning trajectories are steepest for earlier years. We also find that especially at the lowest ability levels, essentially no learning takes place between grades. This is consistent with experimental evidence on the importance of "teaching at the right level" (Banerjee et al., 2016). To make the grade adjustments to the converted TIMSS scores, we apply the corresponding adjustment based on the relevant range of scores from Colombia and Chile for the third-grade and Honduras for sixth-grade. This gives us a customized grade-adjustment based on converted score.

In order for these grade adjustments to be valid, we must first assume that for the doubloon countries, the scores across grades are comparable over time—that is, even though the tests were administered in different years to different students, the reading level of Chilean fourth-grade students in 2011 is the same as it would have been for the Chilean pupils who took LLECE in 2013 but did not enter fourth grade until 2014. Second, we must assume that conditional on baseline score, the change in test scores between grades is constant across countries. For instance, the difference in scores between a third- and fourth-grader in Chile who scores 500 in third grade must be the same as the amount learned by a third-grader in Peru who scores 500. While this is a strong assumption, because we are conditioning on baseline score, any omitted variable that would bias our estimates must differ by country conditional on grade and ability.

3 Building a Test Score "Rosetta Stone"

3.1 Hybrid assessment and data collection

Due to limitations in the publicly available items and test length considerations, it was not feasible to administer a complete exam of each assessment to every student in our sample. Thus, in order to maximize the number of items administered, the pool of public items was randomized across six tests. Each test consisted of roughly 57 items that were translated into Hindi. Appendix Figure A.6 shows some examples of these items from each source exam.

When pooled, the six randomized tests included 60 unique TIMSS items, 53 PIRLS items, 12 math and 16 reading questions from the West African exam, and 4 math and 4 reading questions from the Latin American test. Given the psychometric evidence on item order, the sequence of source exam questions was maintained across versions. A potential tradeoff with this structure is that pupils may tire as the test progresses. Students had two hours to complete the exam, and conditions were kept constant across all tests. Every student was also given the Annual Status of Education Report (ASER) mathematics and

literacy assessments. The tests were administered in 2016 to students from 51 schools across 6 districts in the Indian state of Bihar: Bhojpur, East Champaran, Gaya, Jehanabad, Nalanda, and Patna. Students ranged from grades four through eight with a median age of 11. Slightly over half (52.1 percent) of students in the survey are female.

We exclude 41 items which fewer than 30 students in our sample answered correctly (earned full-credit for partial-credit items.) These questions—largely open response (non-multiple choice) questions all of which come from PIRLS or TIMSS—do not give us enough observations to precisely estimate an item characteristic function within our sample, resulting in large DIF. Since these few dropped items with very low correct response rates tend to be those same items with very few non-missing answers, they could potentially bias our estimates of ability θ_i if included since they reflect particularly odd response patterns X_i. Among the remaining questions, an average of 236 students in our sample gave correct answers (a mean of 129 among the remaining TIMSS and PIRLS items.)

Applying our psychometric methodology yields test scores for each student on each scale. The tests were difficult for most students in our sample: the average item was answered correctly by just 39.3 percent of students who attempted that question and 9.8 percent of students including those who did not respond. Table 1 shows the portion of correct responses within each source assessment. The correct answer percentages are informative about the relative difficulty of the publicly released items from the different tests. For mathematics, for instance, TIMSS is more difficult than the West African or Latin American counterpart. The mean score column shows the IRT linking results of the predicted equivalent test scores. The results show students from the Bihar sample are at or below average on every source test, particularly for TIMSS and PIRLS which are primarily administered in more developed countries. In general, boys performed better than girls in the Bihar sample. Comparing our results to the ASER instrument, we find that in mathematics, students who score a 5 (the highest level) on ASER exhibit fairly wide variation in TIMSS scores, with an interquartile range of over 100 points and an average of 405.9. The ASER literacy instrument has a

slightly more precise correspondence among the top scorers, with 74 points on the PIRLS scale separating the 25^{th} and 75^{th} percentiles around a mean of 418.0.

3.2 Estimating scale conversion functions

To non-parametrically link the international tests, we estimate local linear regressions between student scores from each pair of assessments.[7] The Indian students' performance more closely matches the ability distribution of their African and Latin American peers, so we focus on linking test scores from poor countries (in West Africa and Latin America) on to the scales for rich countries (TIMSS and PIRLS.) This support allows us to build valid links for the biggest domain of input scores. Figure 2 thus shows the equating functions for these primary correspondence mappings, and Appendix Figure A.9 presents the full set of dyads.

There are several different approaches to assessing the reliability of these equating functions. First, the 95-percent confidence intervals of the local linear regressions shown in the graphs highlight throughout most of the distribution, the relationship is tight. At the highest levels of achievement, the low number of observations among the Bihar students does yield a less precise relationship between test scores.

Second, a simple rule-of-thumb for evaluating these relationships is the derivative of these curves should always be positive. Our functions exhibit this property. A third test is the symmetry of equating functions between dyads. In an ideal mapping for practical use in policy, the linking correspondence would be equivalent regardless of the direction of translation. Consider a function $f(A) \in B$ that maps scores from test A to those in test B and the corresponding function $g(B) \in A$ that equates the other direction. If the two functions were symmetric, then $g(f(A)) = A$ and $g(f(B)) = B$. Symmetry does not follow mechanically from our approach because the local linear regressions use different

[7] We use Epanechnikov kernels using the rule-of-thumb bandwidth (Fan and Gijbels, 1996), which fits the data well upon visual inspection and is more conservative with respect to overfitting than bandwidths estimated via leave-one-out cross-validation. See Appendix Figure A.8.

domains (and therefore different observations) when estimating conversions from A to B than B to A. In practice, we find strong visual evidence that this symmetry holds between linking functions, as shown in Appendix Figure A.9. Appendix Table A.1 presents t-tests of differences between A and $g(f(A))$ on the data from Bihar, which are insignificant both statistically and in magnitude.

To calculate confidence intervals on specific estimates, we bootstrap our data by sampling with replacement. We resample both the students in Bihar to re-estimate the equating functions as well as the official microdata from Colombia, Chile, and Honduras for the grade adjustments. For each relevant statistic, we calculate 500 draws of that measure and report the 2.5^{th} and 97.5^{th} percentiles of the resulting distribution. Due to the omission of students with certain response patterns in certain draws, the bootstrapped distribution is not symmetric, motivating the quantile bounds we report.[8]

4 The State of Learning Around the World

4.1 Country rankings

A first application of these results is to provide a more inclusive ranking of countries, pooling results from multiple tests to bring in low-income countries onto the same scale as high-income ones. Our capacity to convert among scores at the tails of the distribution is slightly limited by the support of the score distribution among the students in Bihar.[9] Functional form assumptions on the equating functions or similar extrapolations are thus needed to convert some extreme scores. To minimize the importance of such assumptions, we focus on two statistics that are less sensitive to these omitted outliers. First, we convert the me-

[8]The item response theory model on the bootstrapped sample failed to converge on some of the 500 draws, which we ignore for the purposes of reporting the confidence intervals: once for TIMSS, nine times for PASEC math, six times for PASEC reading, seven times for LLECE reading, and 10 times for LLECE math.

[9]Among the Latin American students in our sample, 2.91 percent of math scores and 6.49 percent of reading scores are outside the range of scores from the Bihari sample. Similarly, 1.70 percent of math scores and 3.20 percent of reading scores among the West African students lie beyond the linking function support.

dian scores for each country that took any one of the international or regional assessments onto the TIMSS and PIRLS scales. Second, we calculate the portion of students above the "low international benchmark" of 400, as defined by TIMSS and PIRLS. This score is comfortably within the range of our conversion functions. This requires a modest monotonicity assumption: those who score below the worst-performing student in our Bihar sample would not pass the benchmark, whereas those better than the best-performing student would.

The results of these converted country scores are presented in Figures 3 and 4. Ninety-five percent confidence intervals based on the bootstrap approach described in the previous section are also shown. The right panels show 2015 GDP per capita adjusted for purchasing power parity in 2011 U.S. dollars, according to the World Bank's World Development Indicators. Overall, low- and middle-income countries score well below OECD countries on, e.g., the TIMSS and PIRLS scales. Test scores are strongly correlated with GDP per capita. Oil-rich countries score relatively low given their income levels. We can visualize this relationship in Figure 5, which shows the relationship between log per capita GDP (PPP) in 2015 and our measure of median test score. The relationship between education quality and income does not change statistically when the sample is expanded according to a Wald Test of the seemingly unrelated regressions.

How do these expanded country rankings compare to other measures human capital? Figure 6 shows the relationship between our "quality" measure of median test scores in 4^{th} grade and the "quantity" measure of average years of schooling for those older than 25 for each country in 2010 from Barro and Lee (2013). Both measures have been residualized on log income. Schooling quantity has relatively low predictive power of education quality, explaining less than a third of variation in test scores.[10]

We can also compare our estimates to previous attempts to pool learning assessments

[10] As an additional robustness check of the validity of our estimates, we reassess a recent literature that has examined the role of preferences in economic development, particularly highlighting patience as an important determinant (Falk et al., 2018; Dohmen et al., 2018). We confirm that relationship with our expanded sample of countries' test scores, as shown in Appendix Figure A.10. Countries with higher average levels of patience as measured by Falk et al.'s (2018) Global Preferences Survey have higher test scores.

without using item response theory or microdata. To illustrate this point, we calculate the number of percentiles on each country's original score distribution that exceeds the "low international benchmark" of 400 on the TIMSS and PIRLS scale. Figure 7 compares this value to the comparable estimate produced by the World Bank based on the approach of Altinok et al. (2018). There are important caveats to this comparison, however. First, our scores are specifically for fourth-grade students, while the World Bank's measure is intended to capture primary schooling in general. Second, Altinok et al.'s (2018) country coverage is much larger than ours given their linking approach, so we can only compare our estimates among a subsample of their scores. Third, our scores are based solely on data since 2011, while their approach pools across more rounds of testing. Despite these differences, the relative country rankings are very similar: the Spearman correlation coefficients between the expanded TIMSS and PIRLS scales and the existing World Bank measures are 0.90 for math and 0.91 for reading.

This ordinal consistency masks some important differences in magnitude along the test score distribution, however. For instance, we estimate that among fourth-grade students in Chad, just 17 percent meet the low international benchmark for math and only five percent in reading, as compared to 44 percent according to the World Bank. Figure 8 plots the absolute percentage point difference in the two estimates against the log of each country's per capita income. There is a strong, negative relationship particularly in math suggesting that the our linking techniques differ most among the poorest countries.

4.2 Skill levels and international trade patterns

To examine the macroeconomic implications of test scores, we study how exports vary by learning achievement.[11] Several papers have documented the important role that human

[11] Our measure is poorly suited to estimate human capital's role in long-run endogenous growth models because our test scores are all measured within the past decade, unlike, for instance, those in Altinok et al. (2018). However, our methodology could be applied to link any exams from any time period retrospectively if the item-level data are available and could trivially be expanded to previous iterations of the exams we link that use a common scale across rounds.

capital can play in structural transformation (Ciccone and Papaioannou, 2009; Bombardini et al., 2012). We test a standard prediction of the Heckscher-Ohlin model by examining whether countries with higher test scores export relatively more in skill-intensive industries (Findlay and Kierzkowski, 1983; Romalis, 2004). To do so, we use export value data in 2017 at the Harmonized System six-digit level from Gaulier and Zignago (2010). We aggregate goods to the industry level using the approach in Autor et al. (2013) and Autor et al. (Forthcoming). To measure skill intensity, we calculate the portion of employees with at least a high school and college degree by industry in the five percent sample of the 2000 United States Census (Ruggles et al., 2019). We estimate the role of test scores according to equation 4, where V_{ci} is the log export value for country c in industry i, γ_c is a country fixed effect, and ω_i is an industry fixed effect.

$$V_{ci} = \beta Score_c \times SkillIntensity_i + \gamma_c + \omega_i + \epsilon_{ci} \tag{4}$$

The results of this regression are shown in Table 2. The results are consistent whether we use math or reading scores or the portion with at least high school or college degrees: countries export more in skill-intensive industries when they have higher test scores.

To further explore this channel, we construct a measure of math and reading intensity separately by industry. The Occupational Information Network (O*NET) is a survey administered by the U.S. Department of Labor to a random sample of U.S. workers in each occupation. We create an index for math using the abilities for number facility and mathematical reasoning and an index for reading using the abilities of written comprehension, written expression, oral expression, and oral comprehension. We follow Autor et al. (2003) to convert average scores by occupation to their weighted percentile rank in the distribution of abilities. O*NET reports two measures for each skill: *importance*, which captures how necessary each ability is for the occupation, and *level*, which indicates how high that level of skill needs to be. We show both measures in our regressions. We use the distribution of

occupations across industries in the 2000 census to link these measures to the corresponding industries. Table 3 shows the results of these regressions where we include interactions both for the converted TIMSS scale and converted PIRLS scale. Interestingly, for both math- and reading-intensive industries, we find that having higher math test scores is associated with relatively more exports but not higher reading scores, though the standard errors are large.

4.3 Education spending and test scores

We next turn to the relationship between education spending and test scores. This topic has received detailed attention by a rich literature in economics using much more credibly causal techniques than the ones presented here (see, for instance, Hanushek (1989); Jackson et al. (2015); de Ree et al. (2017); Muralidharan et al. (2019).) Nevertheless, the cross-country association is important in informing potential explanations for why learning in low-income countries lags so significantly behind rich ones. Using data on government primary expenditure per pupil according to the World Bank for the latest available year, we find a significant positive correlation between education spending and median test score in both reading and math among poor countries but not rich ones. Figure 9 plots this relationship using a quadratic fit. The association is steepest at the lowest levels of expenditure, but at approximately $5,000 per student in purchasing power parity dollars, the relationship levels off. This suggests that external validity of education expenditure studies in rich countries may be limited when inferring lessons for the developing world. This concave relationship is robust to controlling for per capita income, as shown in Table 4.

4.4 Global income inequality and inter-generational transmission of human capital

A key innovation of our approach is our ability to expand the scope of comparable student microdata. We apply these links to create a pooled, cross-national, microdata set with

comparable test scores and incomes at the pupil level.[12] The conversion functions allow us to link exam performance to a common reading and math scale. Like our earlier estimates, we make no distributional assumptions in the linking, so within each country, we restrict our sample to those students in the domain of the Bihar linking function. We convert student scores to the TIMSS and PIRLS scales for math and reading, respectively, and apply the corresponding grade adjustments.[13] This allows us to link 80 countries and 628,587 unique students from the West African, Latin American, TIMSS, and PIRLS exams to their incomes.

The various learning assessments in our database collect information on household asset ownership, but not consumption or income, and no monetary value is reported for assets. We calculate wealth percentiles using the first principal component of asset vectors within each country, a la Filmer and Pritchett (2001). To convert these assets onto a common income scale, we assume an equipercentile relationship between wealth and income. In contrast to, e.g., Young (2012), we require no assumptions about the cardinality of the household wealth index to make this link, only that the ordinal ranking of the household wealth index provides a valid approximation of the ordinal ranking of household income within each country. We draw the world income distribution from the most recent available year for each country using the Lakner-Milanovic World Panel Income Distribution (Lakner and Milanovic, 2016), which includes average incomes in 2005 Purchasing power Parity dollars by country decile.[14]

[12] Although the analysis here focuses only in the years from which we pool the publicly available items, our linking functions could be applied to expand the pool of students to include all past and future rounds of these exams which are themselves internally linked.

[13] We exclude the countries from TIMSS and PIRLS in which only a specific city or state is tested instead of a representative sample of the entire nation, with the exception of the United Kingdom for which we consider scores from England but not Northern Ireland.

[14] To combat attenuation bias, we apply a cubic spline interpolation to the values at the midpoint of each decile and linearly extrapolate for the bottom and top five percentiles. Embedded in this conversion is a further assumption that the sample of tested students is nationally representative of the income distribution data. There are two potential violations. First, the student samples that constitute the test scores are representative of the enrolled population but not of those students who do not go to school. Second, even under full enrollment and a monotonic relationship between wealth and income, the test score data is only representative of the population with primary-age children. While we do not directly address these biases, we expect their overall effect to be small, particularly given that the average net enrollment rate for our linked countries is 93.10 according to the latest available primary net enrollment rate from the World Bank. We exclude seven countries from the test score data that are not available in the income distribution data: Bahrain, Kuwait, Malta, Oman, Qatar, Saudi Arabia, and United Arab Emirates.

The resulting joint distribution of income and test scores around the world is presented in the binned scatter plot shown in Figure 10. The relationship is quite linear in both math and reading. On average, pupils at all incomes below the World Bank low-income classification score below the low international standard. This overall relationship, however, masks considerable heterogeneity by country. Figure 11 shows ordinary-least-squares lines of best fit for each nation in the mathematics sample in gray. The pooled estimates for countries by World Bank income classification are overlaid in colors. The most striking result from this image is the importance of the country effect. At $1,000 in per capita income, for instance, test scores range more than two standard deviations. It should be noted however that within-country measurement error in the student-level income would create attenuation bias exacerbating this range. With this caveat in mind, we can examine the degree to which country incomes matter for test scores *conditional* on student's own household income. Table 5 presents these regressions for math and reading. A bivariate regression of test scores on country income explains 30 percent of the variation in test scores across students. Controlling flexibly for fixed effects of 300 income bins, the R^2 increases by approximately 6 percentage points, and the coefficient on country income falls by roughly 35 percent yet remains substantial and statistically significant. The results are qualitatively similar when controlling for country-specific slopes in student household income.

We can formally decompose the role of household income and country of residence. In simple OLS regressions, income alone explains 25 percent of the variation in performance, while country fixed effects explain 46 percent. In models combining detailed income bin dummies and country fixed effects, we calculate Shorrocks-Shapley decompositions of the R^2 (Shorrocks, 1982). Setting the number of income bins quantiles equal to the number of countries, country fixed effects contribute 68.9 percent of the model's explanatory power for math and 66.2 percent for reading. This share is relatively stable across increasingly flexible specifications for income, as shown in Appendix Figure A.12. This evidence supports the claim that it is not simply that richer pupils perform better but rather that students with the

same level of household resources have radically different educational outcomes depending on their country of residence. This is consistent with existing research on the importance of country of birth on economic outcomes (Milanovic, 2015; Clemens et al., 2019).

The slopes of these gray lines can be interpreted as a measure of intergenerational mobility within each country. Consider a country with a relatively flat slope, so income does not predict test scores. In such a society, we might expect relatively higher convergence of incomes because there is no poverty disadvantage in learning. By contrast, a country with a steep slope would have more persistent inequality as the rich amass more human capital than the poor. We plot the relationship between the correlation of math scores and income and each country's Gini coefficient in Figure 12. This result is reminiscent of the so-called "Great Gatsby Curve" documented by Krueger (2012) and Corak (2013). In mathematics, for instance, Nordic countries like Norway and Finland with low inequality also have low score-income correlations (0.017 and -0.0017, respectively), while the opposite is true of high-inequality Latin American countries like Colombia, Brazil, and Guatemala (0.41, 0.39, and 0.34). Within countries, our converted microdata also allows us to measure test-score inequality. Figure 13 plots the relationship between the ratio of 90^{th} percentile student and 10^{th} percentile student against the median test score. There is a clear negative relationship, with countries with more unequal test score distributions having significantly lower median performances.

4.5 Gender gaps in reading and math

Without a common learning scale, it is possible to compare the sign, but not the magnitude of gender gaps in learning outcomes across regions of the world. We use our converted scores to re-examine gender gaps and find large heterogeneity across countries. It is important to note that the test score data are only representative of students who are enrolled in school, and thus we cannot make statements about gender gaps in the population as a whole without making assumptions about selection in countries below full-enrollment. Pooling to-

gether students across countries and converting their incomes to a common scale via the method above, a clear difference emerges for gender equality in math and reading, as shown in Figure 14. Across all household income levels on average, girls score higher in reading than boys. In math, however, this relative gain exists only for poorer students. At approximately $3,000 per year in annual income, the average boy scores higher, corresponding to roughly the World Bank's threshold for designating upper-middle income countries. Appendix Figure A.11 presents coefficient plots of the interaction between female and household income percentile, showing a notable negative association among richer households in math yet not reading.

Across countries, the relative performance of girls correlates positively with important gender-related outcomes. For instance, Figure 15 plots the relationship between child marriage rates for women age 20 to 24 and the difference in boys' and girls' math scores on the TIMSS scale and in reading on the PIRLS scale. Controlling for income, a 10 percentage point increase in the number of girls married before they turn 18 is associated with a 2.5 point fall in girls' math scores relative to boys. Countries where women on average desire one more child have female students who perform 5.6 points lower in math than their male peers, as shown in Table 6. These results suggest that countries with more conservative social institutions for women also have much lower relative female test scores.

4.6 Private school premiums around the world

The growth of private schooling is one of the most striking trends in global education over recent decades, with the share of primary pupils in private schools rising from 8% in 1980 to 18% in 2018 worldwide (World Bank, 2019). Yet experimental evidence on the causal returns to private schooling has been mixed, suggesting an important role of selection in student sorting (Muralidharan and Sundararaman, 2015). We examine the scope for this story using our linked microdata and information on school type. We begin by estimating the private school premium in each country j using a bivariate regression of each student

i's test score on a dummy $Private_{ij}$ equal to one if that child attends a private school.[15] This coefficient, denoted β_j^{Base}, captures both the average causal effect of attending a private institution and the selection effect driven by endogenous student sorting. To help account for this latter margin, we apply the approach of Oster (2019) to estimate the potential bias from the omitted variables. We first estimate β_j^{Obs} from equation 5 which expands the model from β_j^{Base} to now include student i's gender, country-specific dummies for asset ownership, and the first principal component of this wealth vector.

$$T_{ij} = \alpha_j + \beta_j^{Obs} * Private_{ij} + \omega_{\mathbf{j}}\mathbf{X_{ij}} + \epsilon_{ij} \ \forall j \tag{5}$$

Second, we estimate $\beta_j^{Adj.}$ which adjusts this coefficient based on movement between β_j^{Base} and β_j^{Obs} and the change in R^2 in these two models. We make several assumptions in implementing this procedure. We assume that each of the observables are meaningfully related to the test score T_{ij} at least in part through an impact on the selection into private schools. In terms of the Oster (2019) model, we assume that there are no significant shifts in R^2 by including $\mathbf{X_i}$ that are completely independent of an impact on β. Next, we make assumptions about the maximum R^2 that could be achieved by a regression including all observable and unobservable variables as well as the relative degree of selection between these two sets. We follow Oster (2019) who suggests using 1.3 times the R^2 used in the regression with observables and a relative selection ratio of 1.

Figure 16 plots these three coefficients for each country in our sample. In nearly every country, the private school test score premium falls by half to two-thirds under these simple selection adjustments. These results suggest an important role of student sorting in explaining the cross-type differences in aptitude. To further explore this dynamic, we estimate the bivariate relationships between private school premium coefficients β_j^{Basic} and $\beta_j^{Adj.}$ against country income inequality as measured by a Gini coefficient in a seemingly unrelated regressions framework. We can reject equality between the two slopes with a two-sided p-value

[15] Private school status is only available for a subset of countries in the TIMSS and PIRLS samples.

of less than 0.001. Figure 17 shows that the unadjusted β_j^{Basic} is positively correlated with inequality, but the relationship between the adjusted $\beta_j^{Adj.}$ is much weaker. This evidence is consistent with a model in which private schools compete on socio-economic exclusivity rather than learning gains in highly unequal countries.

5 Conclusion

A key component of understanding human capital's role in development is a globally comparable measure of skill. We develop a new methodology combining original data collection with statistical methods from item response theory to link scores across any set of assessments. We apply this approach to convert the reading and math aptitude of students from Latin America and West Africa to the TIMSS and PIRLS scales commonly used in rich countries. Our technique allows us to relax the demanding distributional assumptions of existing linking methods. This method is portable to convert between any variety of assessments across different contexts, places, and periods, sidestepping many of the political economy challenges of international standardized testing.

Because our approach links individual test scores as opposed to country averages, we can delve into how the distribution of scores changes around the world. We document that learning gaps between rich and poor students widen with a country's income inequality and that more unequal score distributions are found in countries with lower median performances. The role of selection in explaining the private school premium is greatest in the most unequal nations. The STEM gender gap is reversed among the world's poorest, and girls score higher than boys across the whole global income distribution.

The results of our conversion demonstrate significant gaps in learning between pupils in the world's richest and poorest households. But perhaps more strikingly, even for students whose families earn the same amount, there is enormous heterogeneity in what skills they gain in primary school. Further research is needed to understand why some school systems

work so much better than others in building human capital for the next generation.

6 Bibliography

Altinok, Nadir and Hatidje Murseli, "International Database on Human Capital Quality," *Economics Letters*, 2007, *96* (2), 237–244.

_ , Noam Angrist, and Harry Anthony Patrinos, "Global Data Set on Education Quality (1965-2015)," *World Bank Policy Research Working Paper*, 2018.

Angrist, Noam, Simeon Djankov, Pinelopi Goldberg, and Harry A. Patrinos, "Measuring Human Capital," *World Bank Policy Research Working Paper Seris*, 2019, (8742).

Autor, David, David Dorn, and Gordon Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, 2013, *103* (6), 2121—2168.

_ , _ , and _ , "When Work Disappears: Manufacturing Decline and the Falling Marriage-Market Value of Young Men," *American Economic Review: Insights*, Forthcoming.

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics*, 2003, *118* (4), 1279–1333.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton, "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India," *Working Paper*, 2016.

Barro, Robert J., "Economic Growth in a Cross Section of Countries," *Quarterly Journal of Economics*, 1991, *106* (2), 407–443.

_ and Jong-Wha Lee, "Schooling Quality in a Cross-Section of Countries," *Economica*, 2001, *68* (272), 465–488.

_ and _ , "A New Data Set of Educational Attainment in the World, 1950-2010," *Journal of Development Economics*, 2013, *104*, 194–198.

Behrman, Jere R. and Nancy Birdsall, "The Quality of Schooling: Quantity Alone is Misleading," *American Economic Review*, 1983, *73* (5), 928–946.

Bombardini, Matilde, Giovanni Gallipoli, and Germán Pupato, "Skill Dispersion and Trade Flows," *American Economic Review*, 2012.

Braun, H. I. and P. W. Holland, "Observed-Score Test Equating: A Mathematical Analysis of Some ETS Equating Procedures," in P. W. Holland and D. B. Rubin, eds., *Test Equating*, New York: Academic, 1982, pp. 9—49.

Ciccone, Antonio and Elias Papaioannou, "Human Capital, the Structure of Production, and Growth," *Review of Economics and Statistics*, 2009.

Clemens, Michael A., Claudio E. Montenegro, and Lant Pritchett, "The Place Premium: Bounding the Price Equivalent of Migration Barriers," *Review of Economics and Statistics*, 2019, *101* (2), 201—213.

Corak, Miles, "Income Inequality, Equality of Opportunity, and Intergenerational Mobility," *Journal of Economic Perspectives*, 2013, *27* (3).

Das, Jishnu and Tristan Zajonc, "India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement," *Journal of Development Economics*, 2010, *92* (2), 175–187.

de Ree, Joppe, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia," *Quarterly Journal of Economics*, 2017, *133* (2), 993–1039.

Dohmen, Thomas, Benjamin Enke, Armin Falk, David Huffman, and Uwe Sunde, "Patience and Comparative Development," *Quarterly Journal of Economics*, 2018, *133* (4), 1645—1692.

Drasgow, Fritz and Robin Lissak, "Modified Parallel Analysis: A Procedure for Examining the Latent Dimensionality of Dichotomously Scored Item Responses," *Journal of Applied Psychology*, 1983, *68*, 363–373.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde, "Global Evidence on Economic Preferences," *Quarterly Journal of Economics*, 2018, *133* (4), 1645—1692.

Fan, J. and I. Gijbels, *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall, 1996.

Filmer, Deon and Lant Pritchett, "Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India*," *Demography*, 2001, *38* (1), 115–132.

Findlay, Ronald and Henryk Kierzkowski, "International Trade and Human Capital: A Simple General Equilibrium Model," *Journal of Political Economy*, 1983, *91* (6), 957–978.

Gaulier, Guillaume and Soledad Zignago, "BACI: International Trade Database at the Product-Level. The 1994—2007 Version," *CEPII Working Paper, No. 2010-23*, 2010.

Hanushek, Eric A. and Dennis D. Kimko, "Schooling, Labor-Force Quality, and the Growth of Nations," *American Economic Review*, 2000, pp. 1184–1208.

_ **and Ludger Woessmann**, "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation," *Journal of Economic Growth*, 2012, *17* (4), 267–321.

Hanushek, Erik A., "The Impact of Differential Expenditures on School Performance," *Educational Researcher*, 1989, *18* (4), 45–51.

Hastedt, Dirk and Deana Desa, "Linking Errors Between Two Populations and Tests: A Case Study in International Surveys in Education," *Practical Assessment, Research & Evaluation*, 2015, *20* (14), 2.

Hendricks, Lutz and Todd Schoellman, "Human Capital and Development Accounting: New Evidence from Wage Gains at Migration," *Quarterly Journal of Economics*, 2018, *133* (2), 665–700.

Jackson, C. Kirabo, C. Rucker Johnson, and Claudia Persico, "The Effects of School Spending on Educational and Economic Outcomes: Evidence from School Finance Reforms," *Quarterly Journal of Economics*, 2015, *131* (1), 157—218.

Kolen, Michael J. and Robert L. Brennan, *Test Equating, Scaling, and Linking: Methods and Practices, Third Edition*, Springer, 2014.

Krueger, Alan, "The Rise and Consequences of Inequality," *Presentation made to the Center for American Progress, January 12th*, 2012.

Lakner, Christoph and Branko Milanovic, "Global Income Distribution: From the Fall of the Berlin Wall to the Great Recession," *World Bank Economic Review*, 2016, *30* (2).

Mankiw, N. Gregory, David Romer, and David N. Weil, "A Contribution to the Empirics of Economic Growth," *Quarterly Journal of Economics*, 1992, *107* (2), 407–437.

Milanovic, Branko, "Global Inequality of Opportunity: How Much of Our Income is Determined by Where We Live?," *Review of Economics and Statistics*, 2015, *97* (2), 452—460.

Muralidharan, Karthik and Venkatesh Sundararaman, "The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India," *Quarterly Journal of Economics*, 2015, *130* (3), 1011–1066.

_ , Isaac Mbiti, Youdi Schipper, Constantine Mandak, and Rakesh Rajani, "Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania," *Quarterly Journal of Economics*, 2019, *134* (3), 1627–1673.

Oster, Emily, "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 2019, *37* (2), 187–204.

Romalis, John, "Factor Proportions and the Structure of Commodity Trade," *American Economic Review*, 2004, *94* (1), 67–97.

Ruggles, Steven, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas, and Matthew Sobek, "IPUMS USA: Version 9.0 [dataset]," *IPUMS*, 2019.

Sandefur, Justin, "Internationally Comparable Mathematics Scores for Fourteen African Countries," *Economics of Education Review*, 2018, *62*, 267–286.

Shorrocks, A. F., "Inequality Decomposition by Factor Components," *Econometrica*, 1982, *50* (1).

Singh, Abhijeet, "Learning More with Every Year: School Year Productivity and International Learning Gaps," *Journal of the European Economic Association*, 2019.

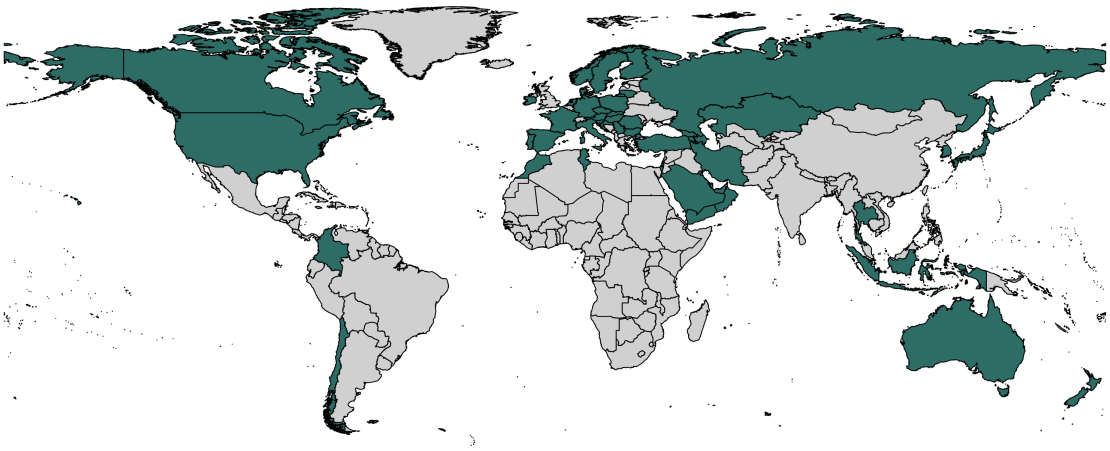
Steinmann, Isa, Rolf Strietholt, and Wilfried Bos, "Linking International Comparative Student Assessment," *LINCS Technical Report*, 2014.

World Bank, "World Development Indicators," 2019.

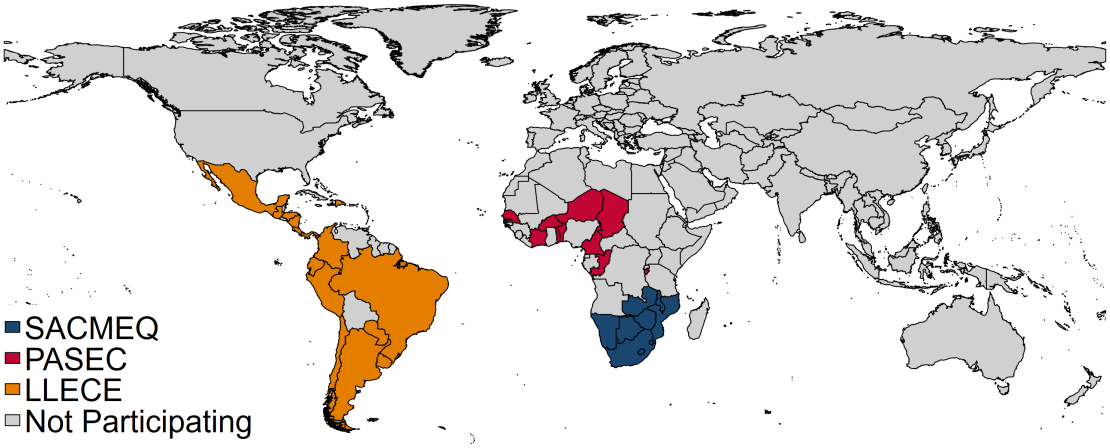
Young, Alwyn, "The African growth miracle," *Journal of Political Economy*, 2012, *120* (4), 696–739.

Figure 1: Coverage of regional and international learning assessments

(a) TIMSS or PIRLS



(b) Regional Assessments: PASEC, LLECE, and SACMEQ



Note: Figure 1 maps the coverage of some of the world's largest standardized tests. Panel 1a shows countries which participated in Trends in International Mathematics and Science (TIMSS) and Progress in International Reading Literacy Study (PIRLS) in 2011, and panel 1b shows countries which participated in Analysis Program of the CONFEMEN Education Systems (PASEC) in 2014, Latin American Laboratory for Assessment of the Quality of Education (LLECE) in 2013, and the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) in 2007. Data from the World Bank's Education Statistics.

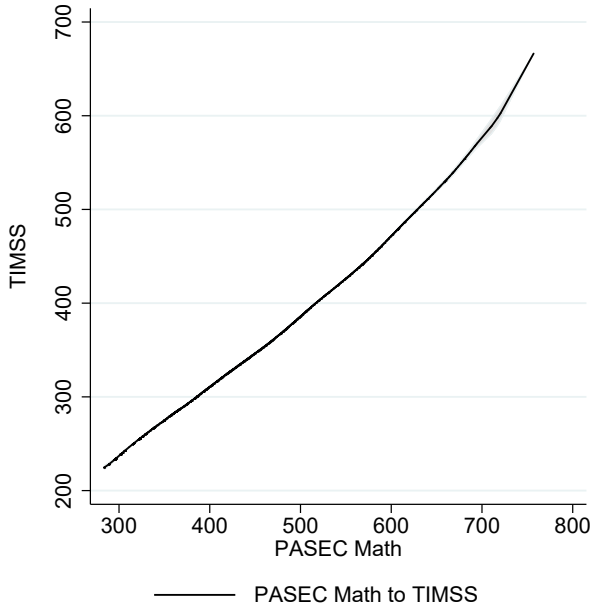
Table 1: Summary Statistics by International Assessment

Source Test	Number of Overlap Items	Average Score in Bihar Sample	% Correct in Ref. Pop.	% Correct in Bihar Sample	% Correct Bihar— Proper Grade	Average # Non-Missing Answers per Item
LLECE Math	4	688.90	51.00	56.99	42.32	1027
LLECE Read	4	652.11	51.00	33.51	23.57	745
PASEC Math	12	468.49	22.67	46.39	44.79	915
PASEC Read	16	480.38	22.38	49.37	47.96	908
PIRLS	31	402.61	63.23	31.05	23.56	330
TIMSS	42	389.14	54.38	38.04	28.61	366

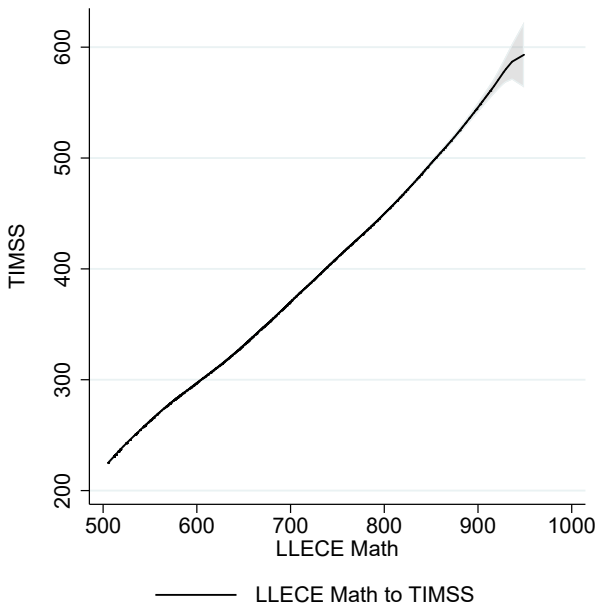
Note: Table 1 shows summary statistics for each source exam. The second column gives the number of public items used in the aggregated tests (excluding the 37 items that were answered correctly by fewer than 20 .) The average scores of students in the Bihar sample on the original reference scales is given in column three. By design, the international mean is 500 for PASEC, TIMSS, and PIRLS exams and 700 for LLECE. The average portion correct by item in the reference population is in column four, and the corresponding number for the Bihar sample is in column five. Column six gives this same calculation conditional on the appropriate grade level corresponding to the reference population: grade six for PASEC, and grade four for PIRLS and TIMSS. Since no third graders sat the Bihar exam, column six shows the average for fourth grade students for the LLECE assessments. The last column shows the average number of non-missing student responses per item.

Figure 2: Test Equating Functions

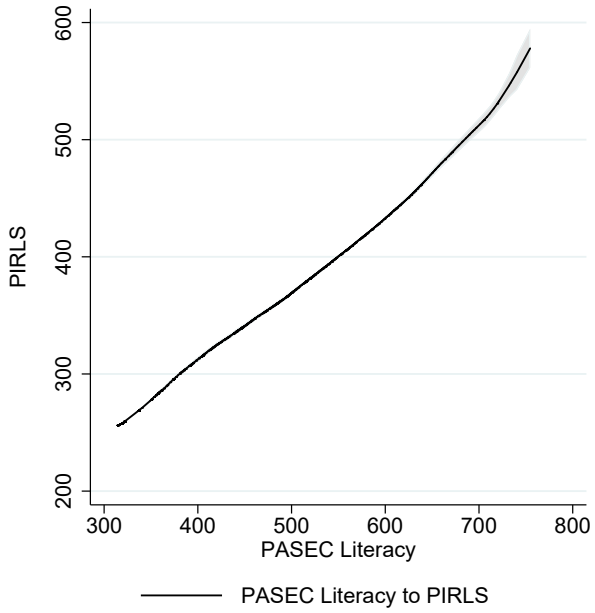
(a) PASEC to TIMSS



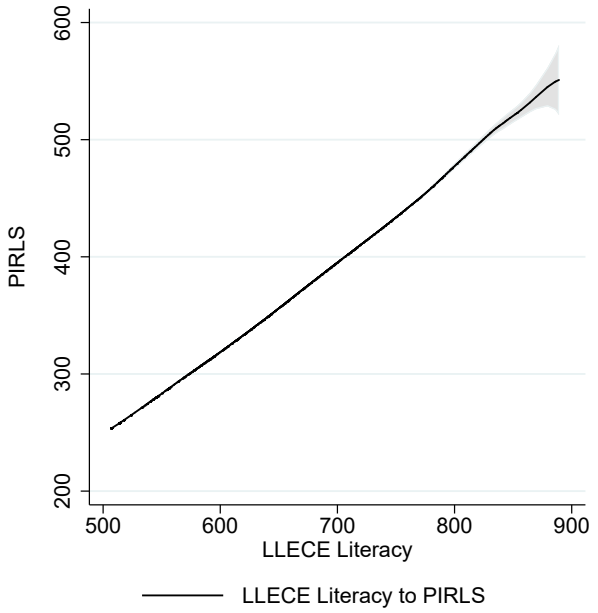
(b) LLECE to TIMSS



(c) PASEC to PIRLS

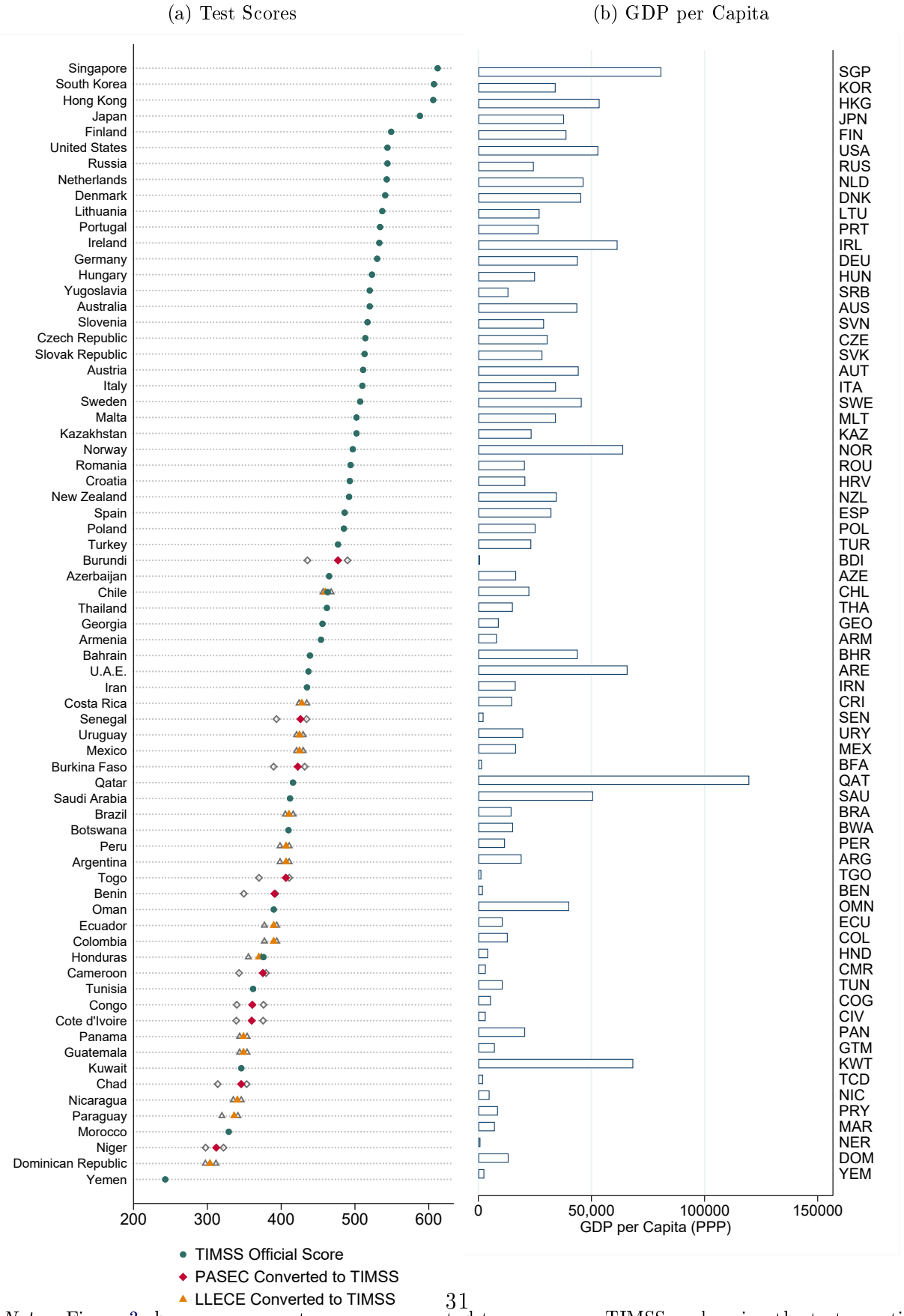


(d) LLECE to PIRLS



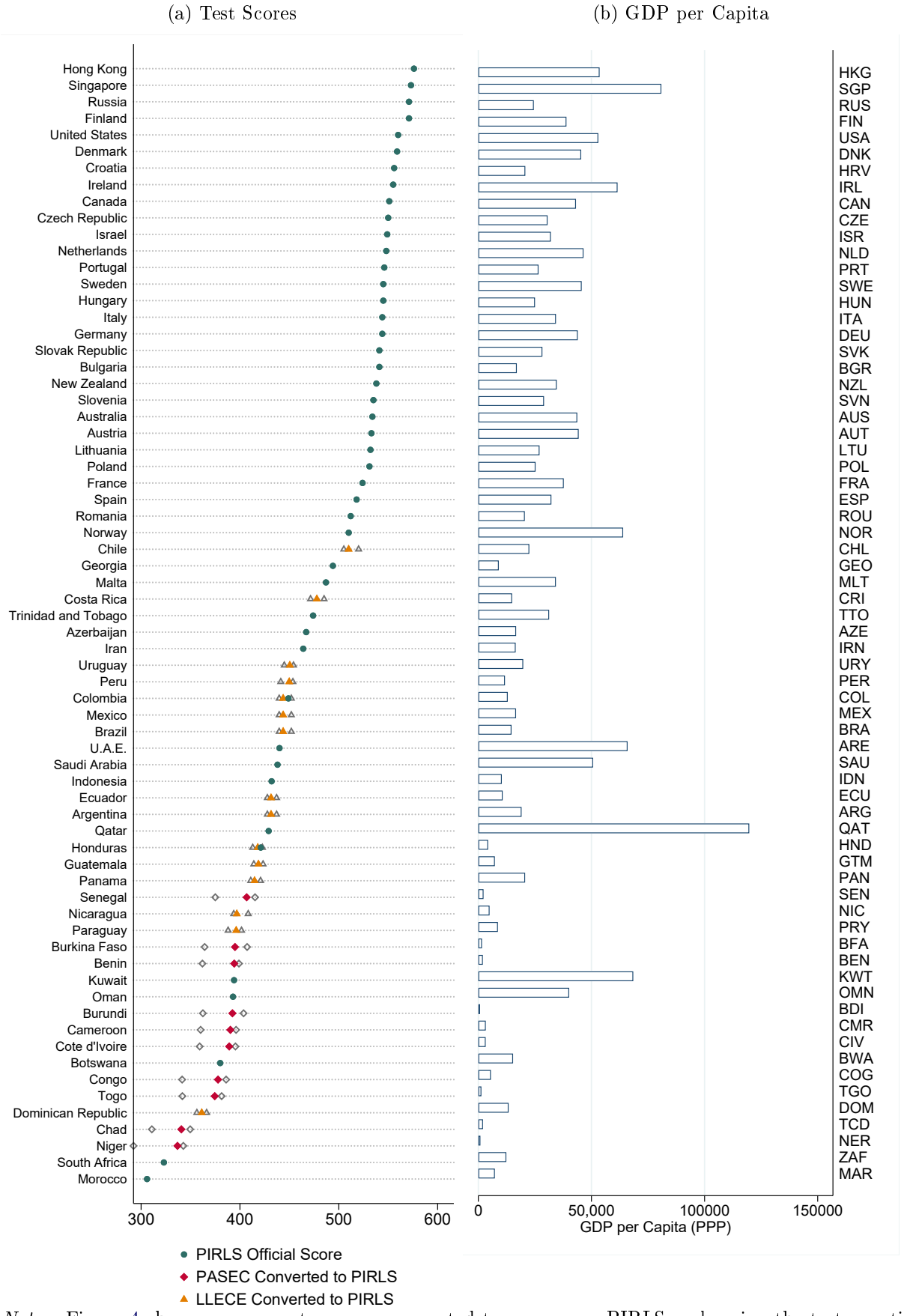
Note: Figure 2 shows test equating functions moving from PASEC and LLECE grade 3 to TIMSS and PIRLS, with 95 percent confidence intervals denoted by the shaded area. Local linear regressions using Epanechnikov kernels are used with a bandwidth equal to one-fifth of the standard deviation of the reference tests (20 points.)

Figure 3: Median Math Score on TIMSS Scale



Note: Figure 3 shows mean country scores converted to a common TIMSS scale using the test equating functions in figure A.9. The right panel shows 2015 GDP per capita PPP in 2011 dollars (World Bank, 2019). For further details, see section 3.2.

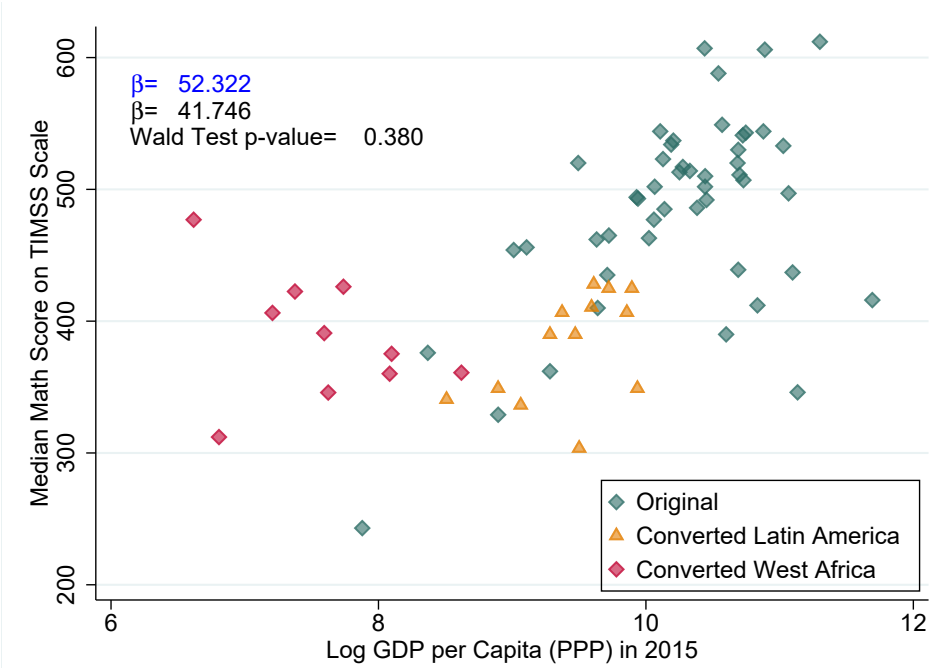
Figure 4: Median Literacy Score on PIRLS Scale



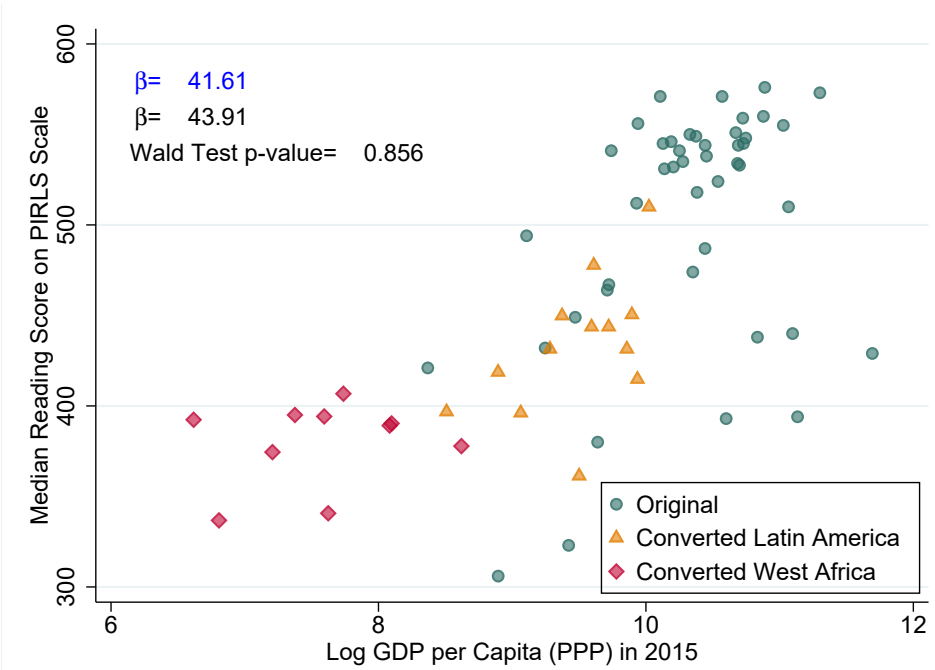
Note: Figure 4 shows mean country scores converted to a common PIRLS scale using the test equating functions in figure A.9. The right panel shows 2015 GDP per capita PPP in 2011 dollars (World Bank, 2019). For further details, see section 3.2.

Figure 5: Test Scores and Income Across Countries

(a) Math



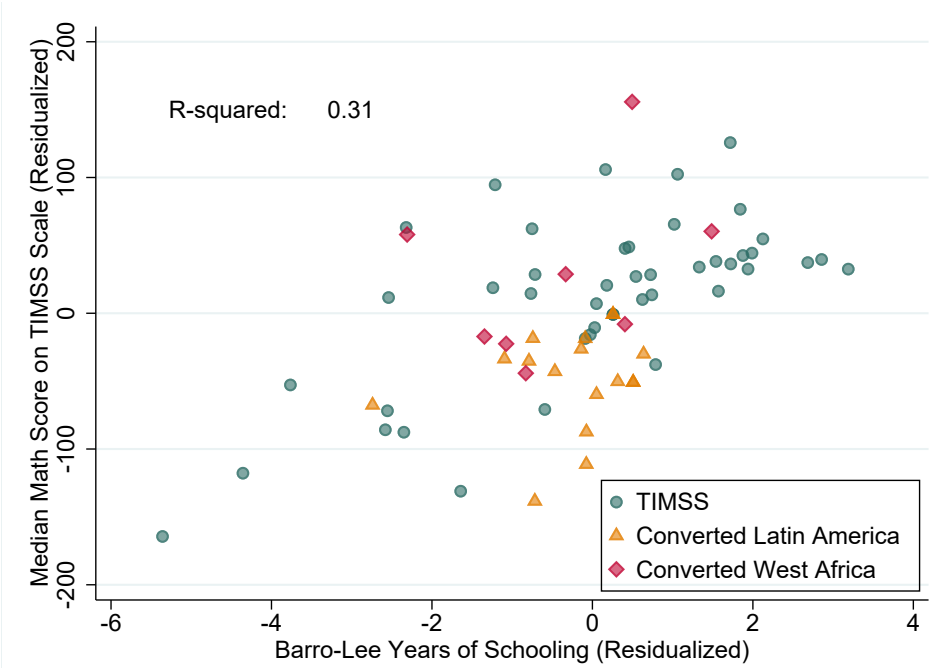
(b) Reading



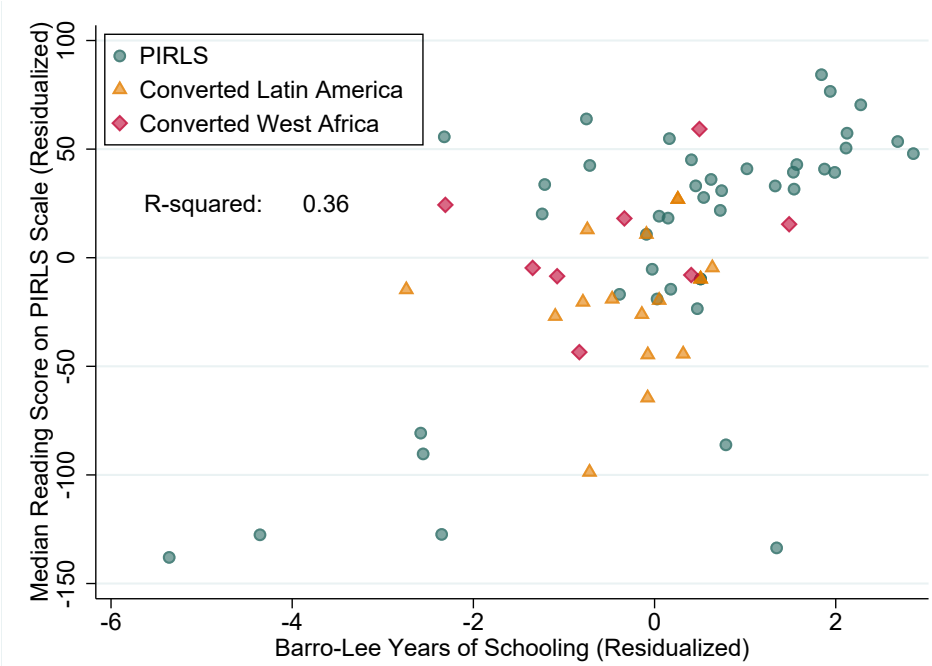
Note: Figure 5 shows scatter plots for median test scores and log GDP per capita (PPP) in 2015 for math and reading, respectively. Income data comes from the World Bank's World Development Indicators (World Bank, 2019). Math and reading scores have been converted to TIMSS and PIRLS scales, respectively. The first β-coefficient in blue is the coefficient on income in a bivariate regression with score limiting to the original sample of countries who take the test. The second β-coefficient corresponds to the same regression, expanding the sample to the converted country scores from Latin America and West Africa. The Wald Test p-value tests whether these two coefficients are different using seemingly unrelated regressions.

Figure 6: Does Quantity Proxy for Quality?

(a) Math



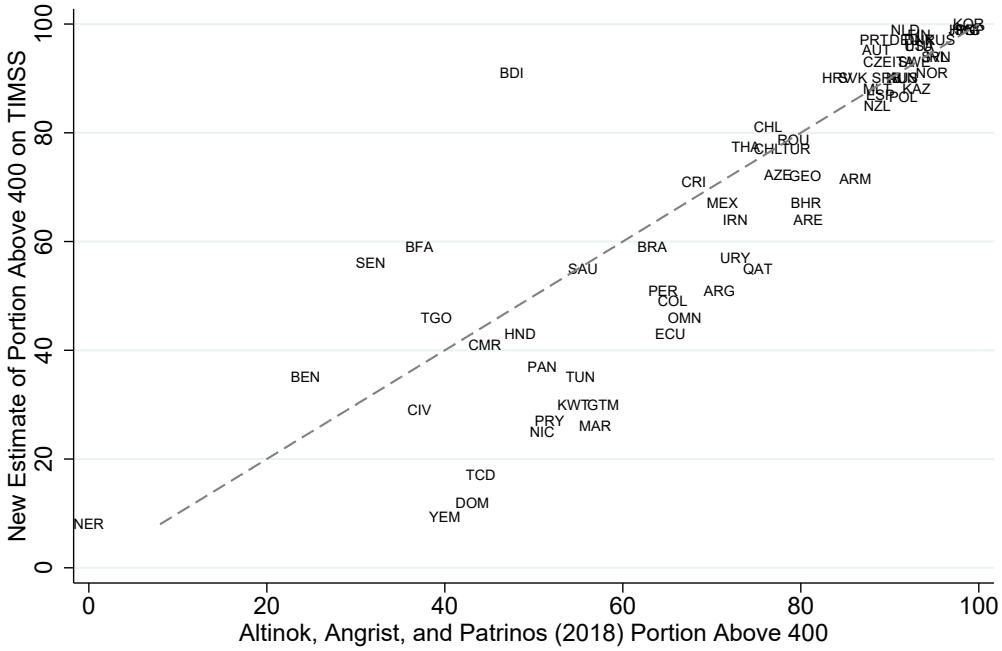
(b) Reading



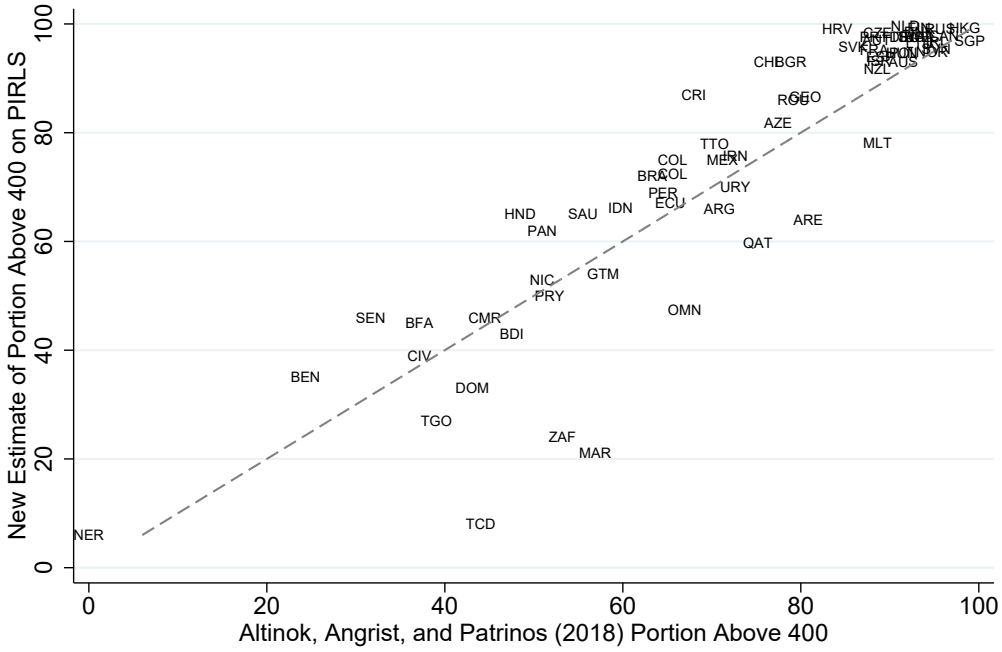
Note: Figure 6 shows the Barro and Lee (2013) measure of average years of schooling against our measure of median test score by country, both of which have been residualized on log GDP per capita in 2015 (PPP) from the World Bank's World Development Indicators (World Bank, 2019). Figure 6a shows math scores on the TIMSS scale, and figure 6b shows reading scores on the PIRLS scale. The reported R^2 is from a regression of residualized test score on residualized years of schooling.

Figure 7: Comparison Between New Estimates and Altinok et al. (2018)

(a) Math



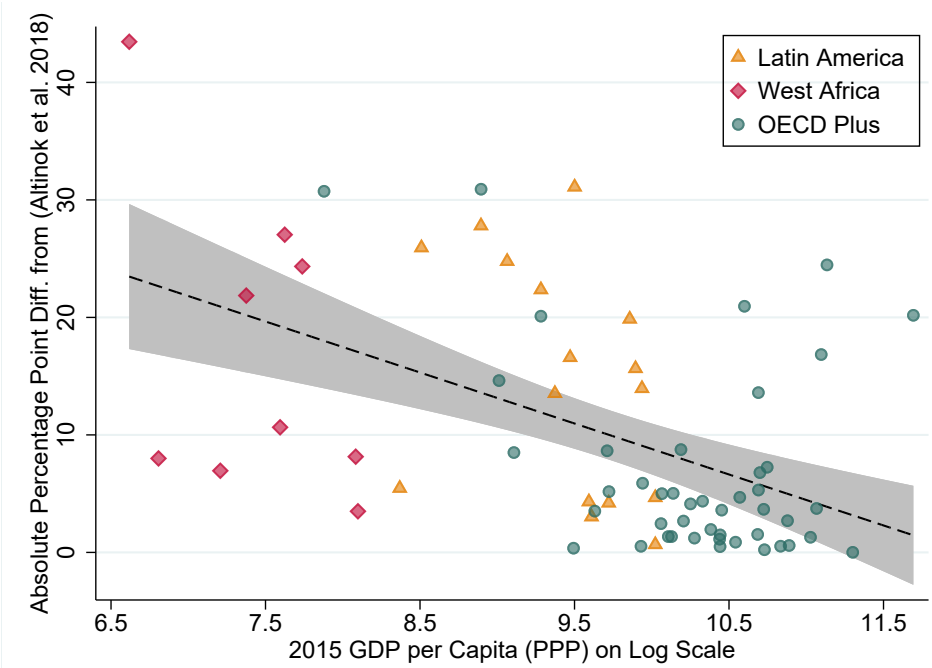
(b) Reading



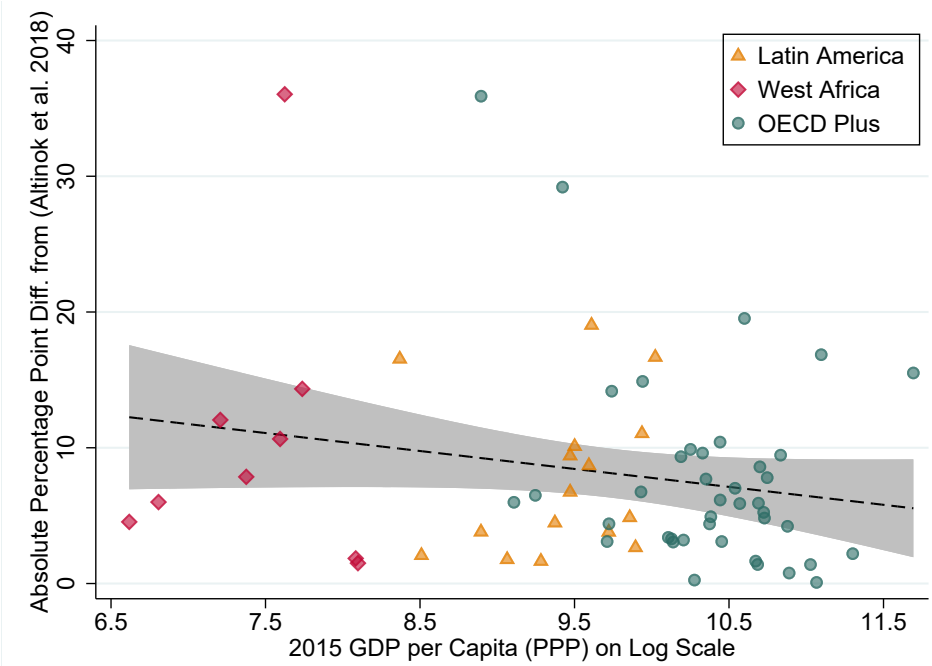
Note: Figures 7 shows comparisons between our new estimated country scores and the existing World Bank standard based on Altinok et al. (2018). The plotted statistic for both measures is the portion of students scoring above 400 points. Figure 7a shows this relationship for math on the TIMSS scale, and figure 7b does the same for reading on the PIRLS scale. The dashed gray line denotes where the two estimates give the same portion meeting the "low international benchmark." Our sample is based on countries that took the 2011 TIMSS, 2013 LLECE, and 2014 PASEC exams, while the Altinok et al. (2018) numbers are the most recent years available for those countries. Only countries that have values from both scales are shown.

Figure 8: Deviation from Altinok et al. (2018) by GDP per Capita (PPP)

(a) Math



(b) Reading



Note: Figure 8 plots the average percentage point threshold difference from Altinok et al. (2018) against country income. The threshold difference is calculate by taking the average absolute percentage point difference in the portion of students who pass the "low international benchmark" in reading and math from our new estimates and the World Bank's measure. Income data is measured in log GDP per capita in 2015 (PPP) from the World Bank's World Development Indicators (World Bank, 2019). Ordinary least squares lines of best fit are plotted in dashed lines, and 95 percent confidence intervals are shown in shaded areas.

Table 2: Test Scores and Exports in Skill-Intensive Industries

	Math		Reading	
	College	High School	College	High School
Score × Skill Intensity	9.02791	2.45827	6.98260	1.91748
SE	(2.74229)	(0.55908)	(2.60647)	(0.60487)
Country F.E.	Yes	Yes	Yes	Yes
Industry F.E.	Yes	Yes	Yes	Yes
Observations	29,549	29,549	28,585	28,585
R^2	0.58	0.58	0.59	0.59

Note: Table 2 presents results on the relationship between test scores and the value of exports in skill-intensive industries. Scores are converted to the common TIMSS and PIRLS scales for math and reading, respectively and have been divided by 1000 for readability. Skill intensity is measured by the portion of employees in a given industry with at least a high school or college degree according to the 2000 United States Census five percent microdata sample from Ruggles et al. (2019). Export data from Gaulier and Zignago (2010) has been aggregated from the HS six digit code to the industry level. All regressions include country and industry fixed effects. Standard errors in parentheses are clustered at the country-level.

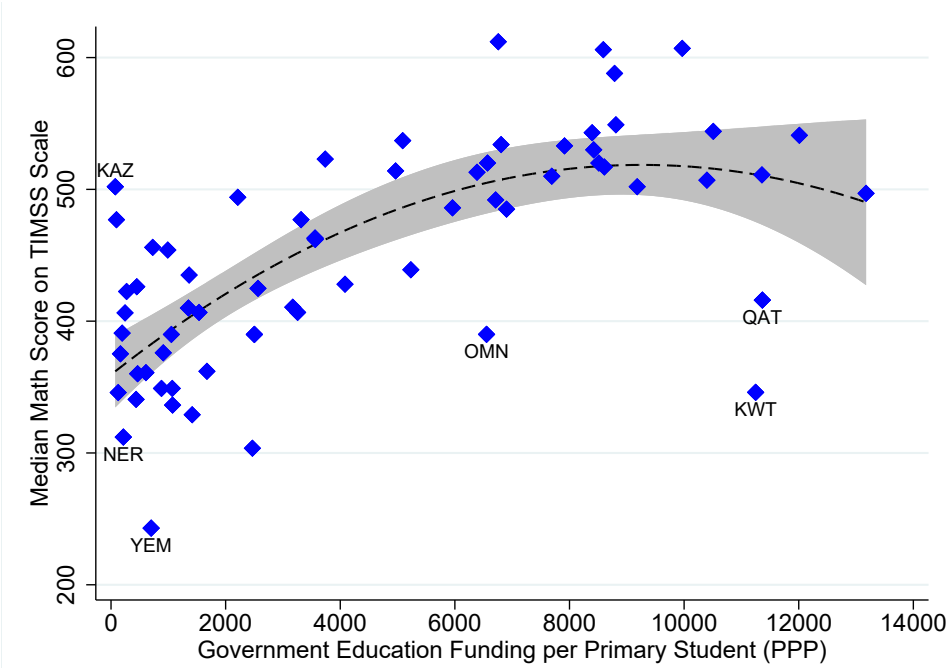
Table 3: Test Scores and Exports by Skill-Type

	Math		Reading	
	Level	Importance	Level	Importance
Math Score × Skill Intensity	0.45062	0.47942	0.61629	0.37363
SE	(0.31060)	(0.33555)	(0.30851)	(0.22617)
Reading Score × Skill Intensity	-0.09071	-0.10407	-0.25291	-0.14234
SE	(0.34108)	(0.37046)	(0.34561)	(0.25071)
Country F.E.	Yes	Yes	Yes	Yes
Industry F.E.	Yes	Yes	Yes	Yes
Observations	29,549	29,549	28,585	28,585
R^2	0.58	0.58	0.59	0.59

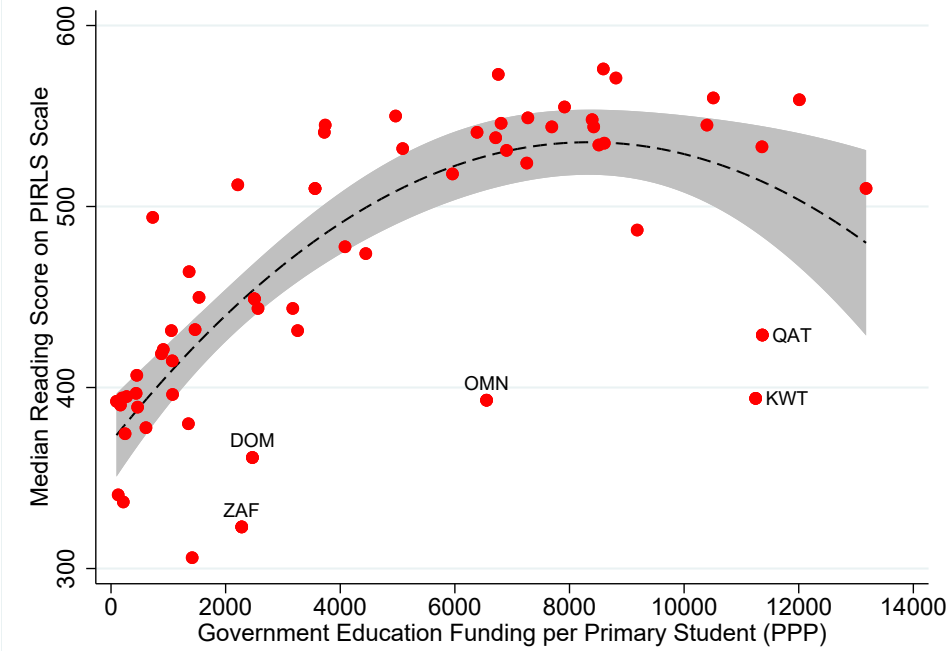
Note: Table 3 presents results on the relationship between test scores and the value of exports in skill-intensive industries. Scores are converted to the common TIMSS and PIRLS scales for math and reading, respectively and have been divided by 100 for readability. Skill intensity is measured using O*NET linked to industries from occupations using the 2000 United States Census five percent microdata sample from Ruggles et al. (2019). Export data from Gaulier and Zignago (2010) has been aggregated from the HS six digit code to the industry level. All regressions include country and industry fixed effects. Standard errors in parentheses are clustered at the country-level.

Figure 9: Government Education Spending and Test Scores

(a) Math



(b) Reading



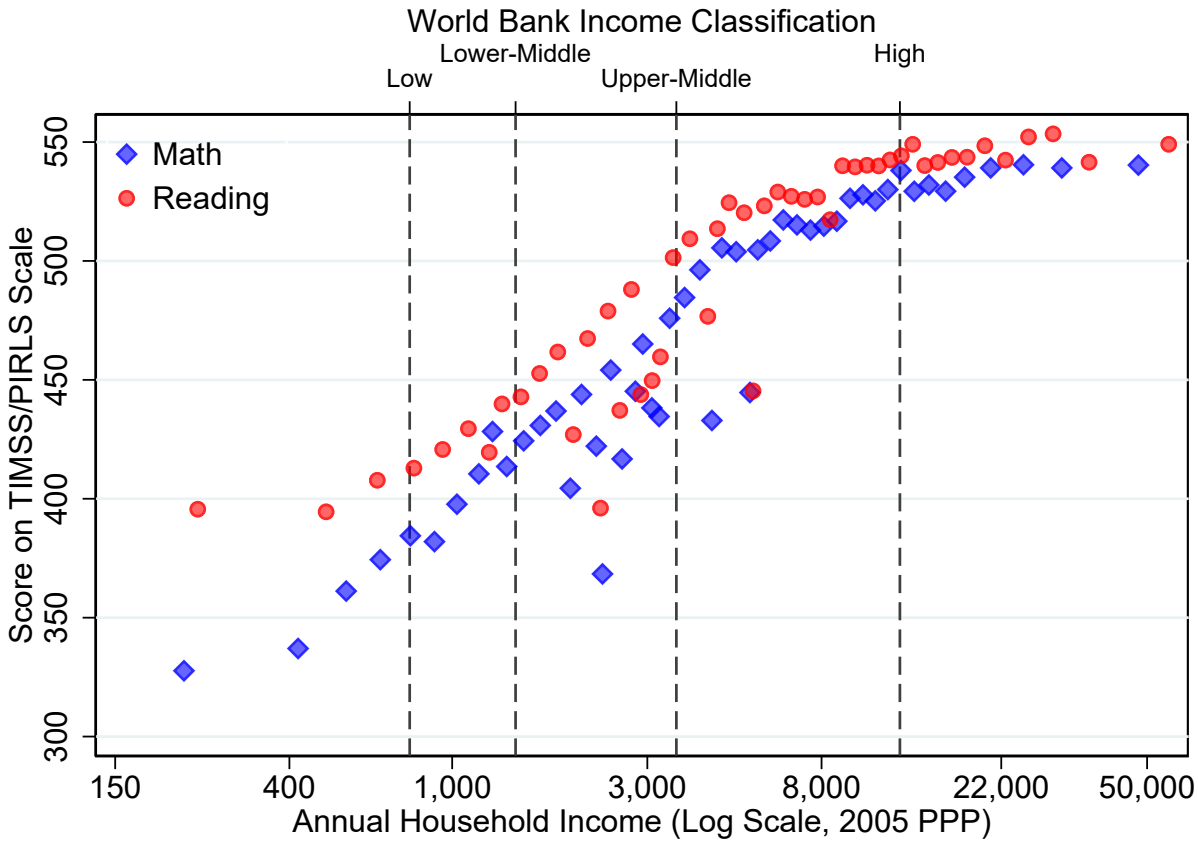
Note: Figure 9 shows the relationship between education spending and test scores. Figure 9a plots the median math score converted onto the TIMSS scale for each country against the average government spending on primary education per student for the latest available year, according to the World Bank. Figure 9b does the same using median reading score on the PIRLS scale. A quadratic least squares line of best fit is denoted by the dashed line, and the associated 95 percent confidence interval are shown in gray. Selected country codes are also displayed.

Table 4: Government Spending and Test Scores

	Math Score			Reading Score		
	(1)	(2)	(3)	(4)	(5)	(6)
Log GDP per Capita 2015 (PPP)	43.79 (8.938)		0.472 (18.23)	45.62 (5.768)		3.284 (9.890)
Govt. Funding per Primary Pupil		3437.0 (665.9)	3410.7 (1244.2)		3969.9 (482.2)	3772.4 (851.9)
Govt. Funding per Primary Pupil Squared		-0.186 (0.0590)	-0.184 (0.0821)		-0.238 (0.0483)	-0.228 (0.0642)
Observations	67	67	67	63	63	63
R^2	0.380	0.505	0.505	0.504	0.621	0.621

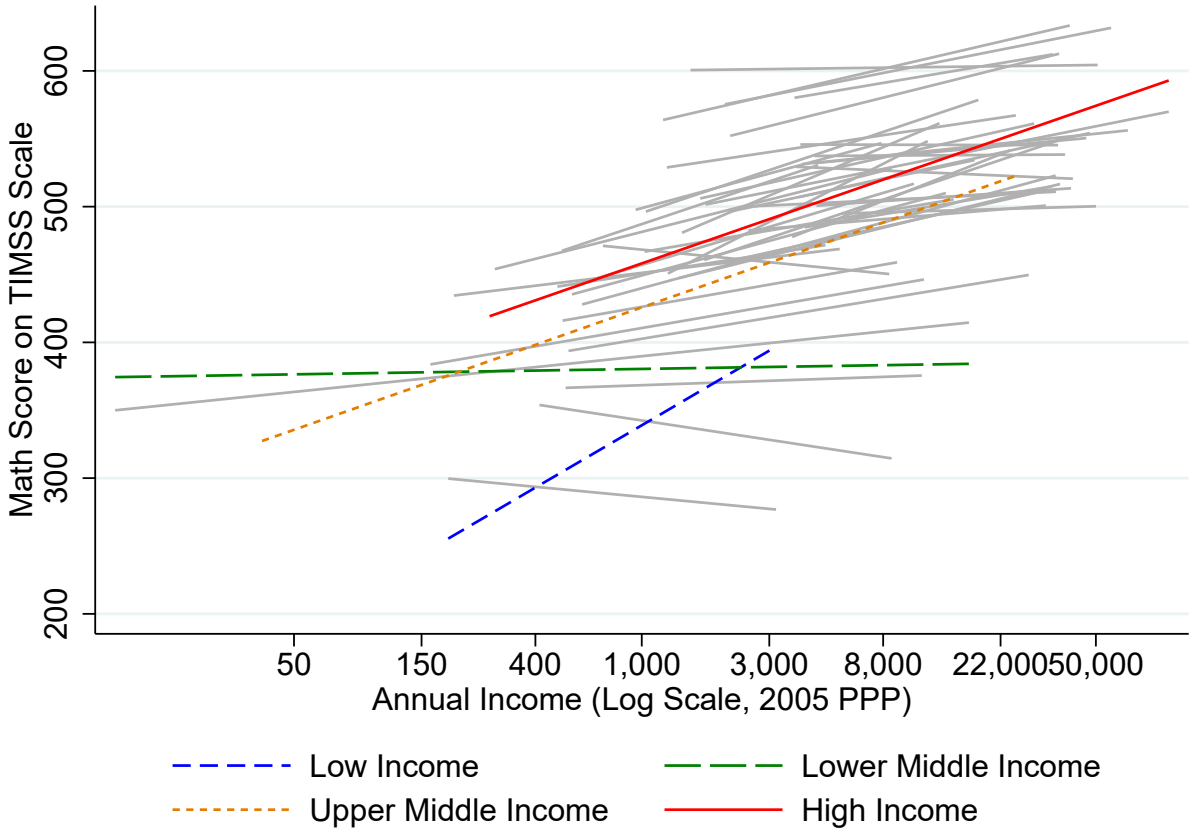
Note: Table 4 shows cross-country regressions between our expanded median test scores, government education spending per pupil, and log per capita income. Income data is measured in log GDP per capita in 2015 (PPP) from the World Bank's World Development Indicators. Government education spending per pupil is also in PPP dollars and comes from the last available year from the World Bank (World Bank, 2019). The expanded TIMSS and PIRLS scores add PASEC and LLECE countries onto the corresponding scales using our linking functions and grade adjustments. Robust standard errors are shown in parentheses. The coefficients on spending are scaled by 100,000 for readability.

Figure 10: The Global Relationship Between Learning and Income



Note: Figure 10 shows a binned scatter plot with 50 buckets of learning scores against per capita income, combining test score data from LLECE, PASEC, TIMSS, and PIRLS. Test scores are converted to the TIMSS and PIRLS scale. Wealth percentiles by country were first calculated from the exam data using the first principal component of household assets. These percentiles were then linked to a spline-interpolation of the global income distribution from Lakner and Milanovic (2016). Income is plotted on a log scale.

Figure 11: Test Scores and Income Within Countries



Note: The gray lines in figure 11 show ordinary least squares estimates of lines of best fit between test scores and income at the percentile-level within each country. The colored lines show the same relationships but pooling countries according to their World Bank income classification. Test scores are all converted to the TIMSS scale. Income is plotted on a log scale.

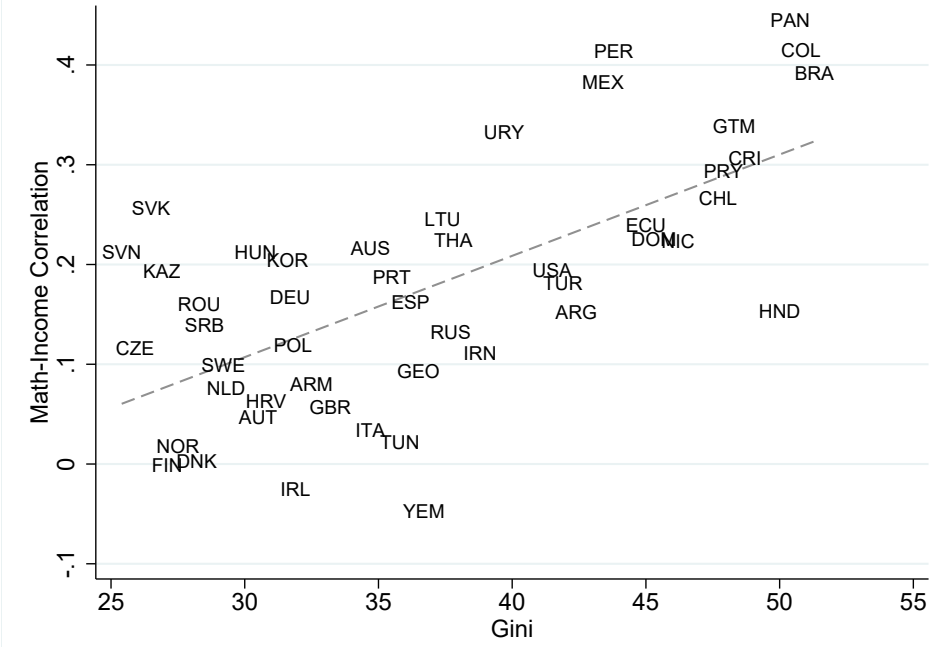
Table 5: Test Scores, Country Income, and Household Income

	Math (TIMSS Scale)			Reading (PIRLS Scale)		
	(1)	(2)	(3)	(4)	(5)	(6)
Log GDP per Capita 2015 (PPP)	58.12 (8.098)	56.18 (5.636)	27.28 (17.75)	54.66 (5.234)	45.05 (7.175)	37.61 (21.25)
Income Bin F.E.s	No	Yes	No	No	Yes	No
Country-Specific Slopes	No	No	Yes	No	No	Yes
Observations	290,831	290,831	290,831	296,600	296,600	296,600
Clusters	65	65	65	63	63	63
R^2	0.308	0.396	0.490	0.326	0.392	0.464

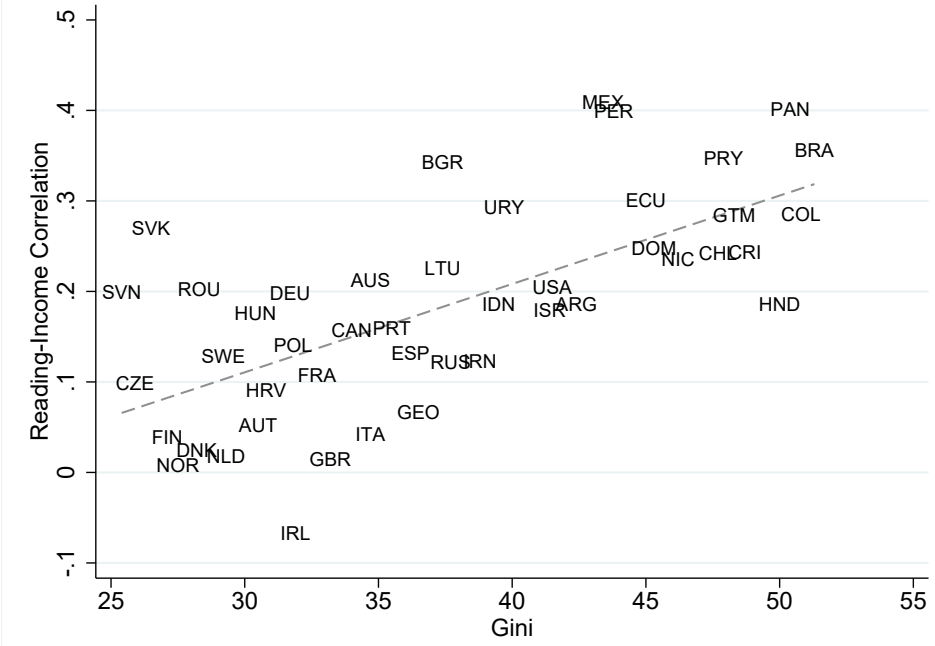
Note: Table 5 regresses individual-level test scores on the country's log GDP per capita separately for math and reading. The first column of each subject shows the simple bivariate OLS regression. The second column flexibly controls for fixed effects of 300 quantiles of student-level log household income. The third column allows for country-specific linear trends in log household income. Income data is measured in log GDP per capita in 2015 (PPP) from the World Bank's World Development Indicators (World Bank, 2019), and household income is also in PPP following the procedure described in the text. The expanded TIMSS and PIRLS scores add PASEC and LLECE countries onto the corresponding scales using our linking functions and grade adjustments. Standard errors clustered at the country level are shown in parentheses.

Figure 12: Gatsby Curve

(a) Math



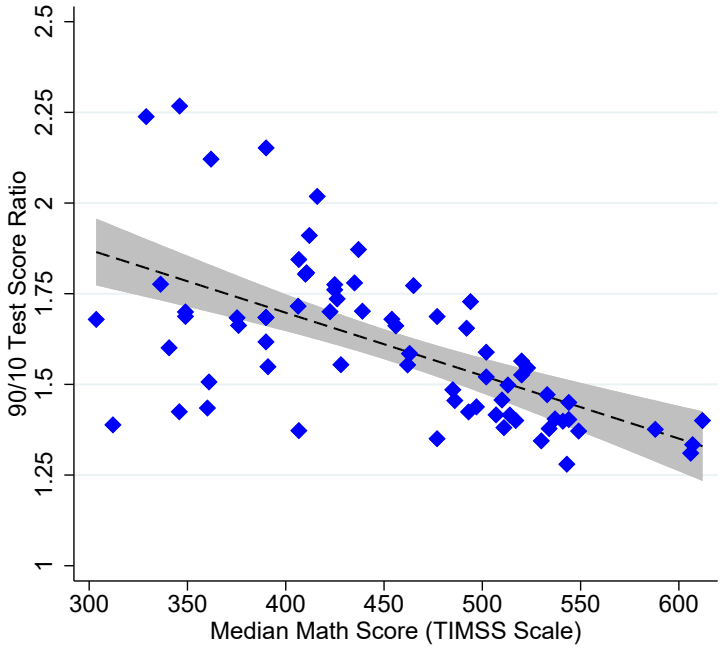
(b) Reading



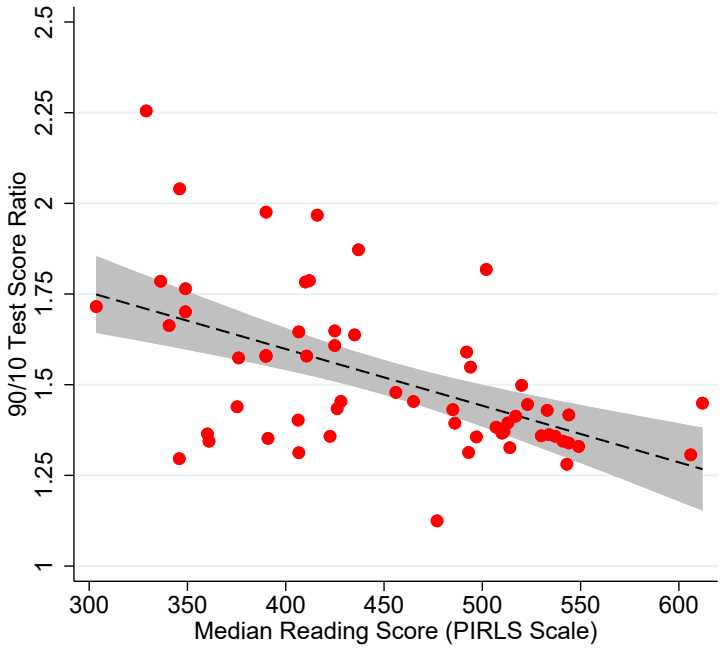
Note: Figure 12 plots each country's Gini coefficient against the correlation between income at the decile-level and test score. Figure 12a shows math scores converted to the TIMSS scale, and figure 12b shows reading scores on the PIRLS scale. The ordinary least squares line of best fit is plotted in the gray dashed lines.

Figure 13: Median Test Score and 90/10 Learning Inequality

(a) Math

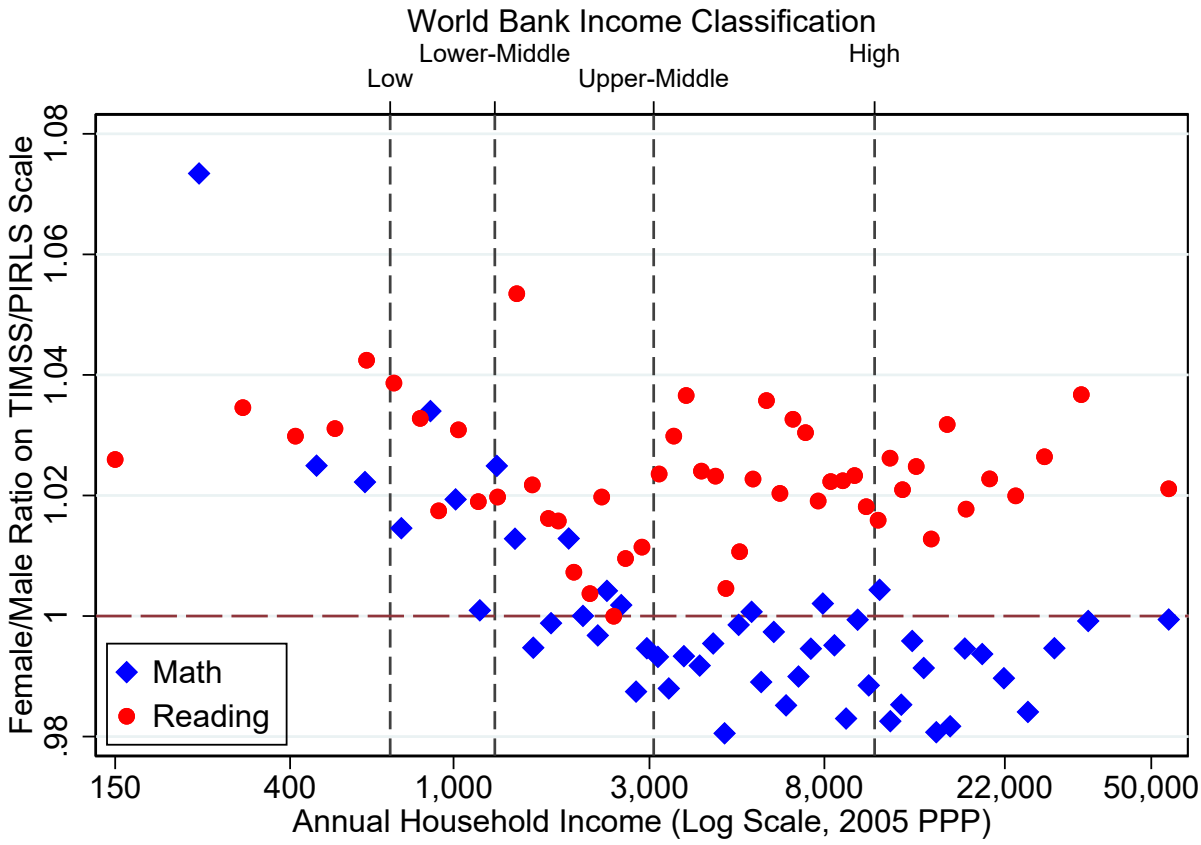


(b) Reading



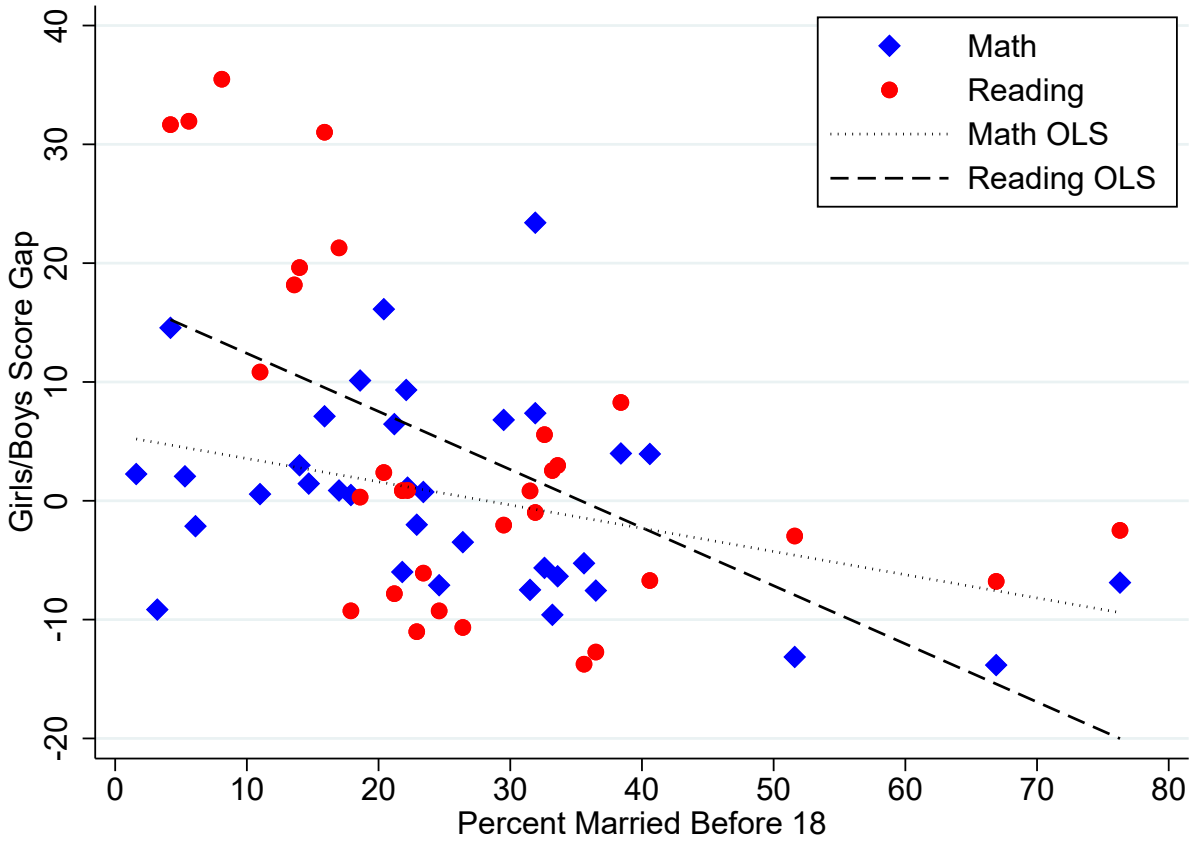
Note: Figure 13 shows the relationship between the median test score and the ratio of the 90^{th} and 10^{th} percentiles. The relationship for math scores converted to the TIMSS scale is shown in figure 13a, and that for reading on the PIRLS scale is shown in figure 13b. The outlier Yemen, which has both the lowest median math score (243.00) and the highest 90/10 ratio (3.04) is omitted from the graph for readability.

Figure 14: Gender Differences in Test Scores Across the Global Income Distribution



Note: Figure 14 shows the ratio of female to male scores by 50 quantiles of the per capita income distribution, combining test score data from LLECE, PASEC, TIMSS, and PIRLS. Test scores are converted to the TIMSS and PIRLS scale. Wealth percentiles by country were first calculated from the exam data using the first principal component of household assets. These percentiles were then linked to a spline-interpolation of the global income distribution from Lakner and Milanovic (2016). Each point is the ratio of the average test score by gender within each global income percentile. Income is plotted on a log scale.

Figure 15: Gender Score Gap and Child Marriage



Note: Figure 15 plots the coefficient on female in a bivariate regression with converted test scores for each country against the portion of women ages 20-24 who were first married by age 18. The marriage data come from a variety of household surveys compiled together by the World Bank (World Bank, 2019). The statistics for the latest available year are used. The circles denote the gender gap in reading on the PIRLS scale, and the diamonds denote the same for math on the TIMSS scale. The ordinary least squares lines of best fit are plotted in the dashed lines.

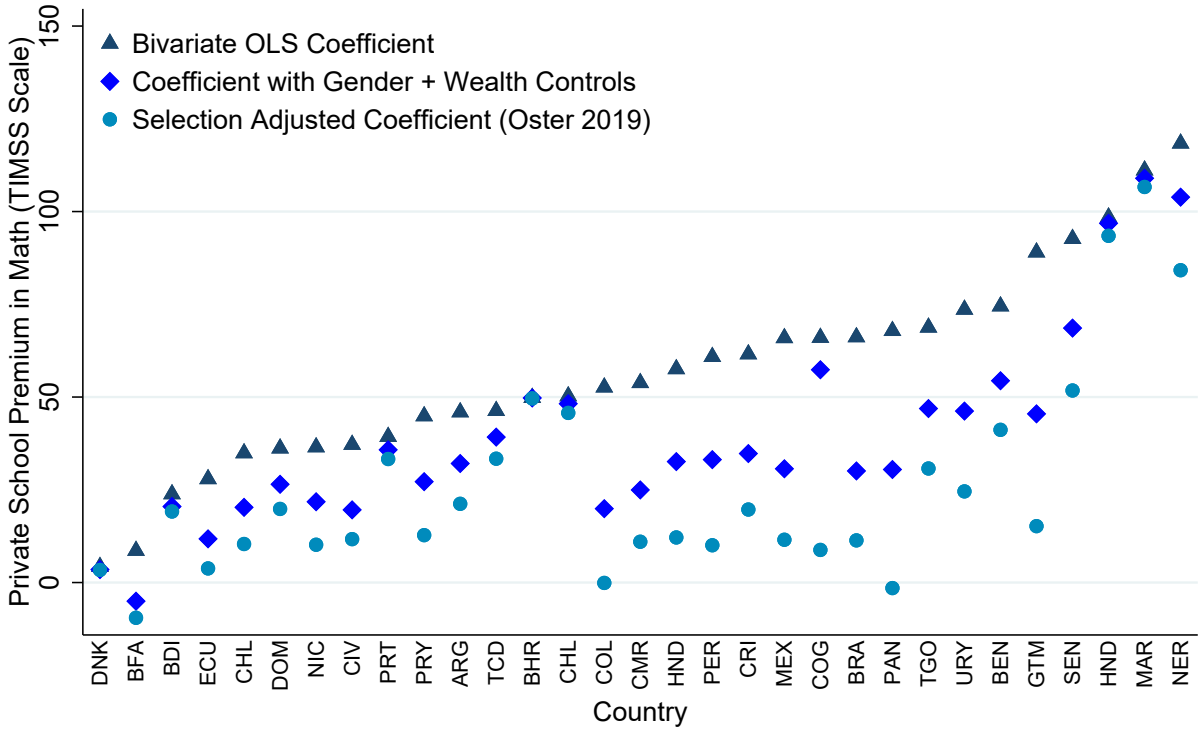
Table 6: Gender Test Score Gap and Female Empowerment

	Girls - Boys Math Gap (TIMSS Scale)		Girls - Boys Reading Gap (PIRLS Scale)	
	(1)	(2)	(3)	(4)
Log GDP per Capita 2015 (PPP)	-1.239 (2.047)	-6.975 (3.632)	-1.132 (2.569)	0.185 (2.889)
% Married by 18	-0.247 (0.100)		-0.539 (0.191)	
Desired Fertility		-5.603 (1.900)		-1.230 (1.959)
Observations	35	24	31	22
R^2	0.163	0.331	0.306	0.044

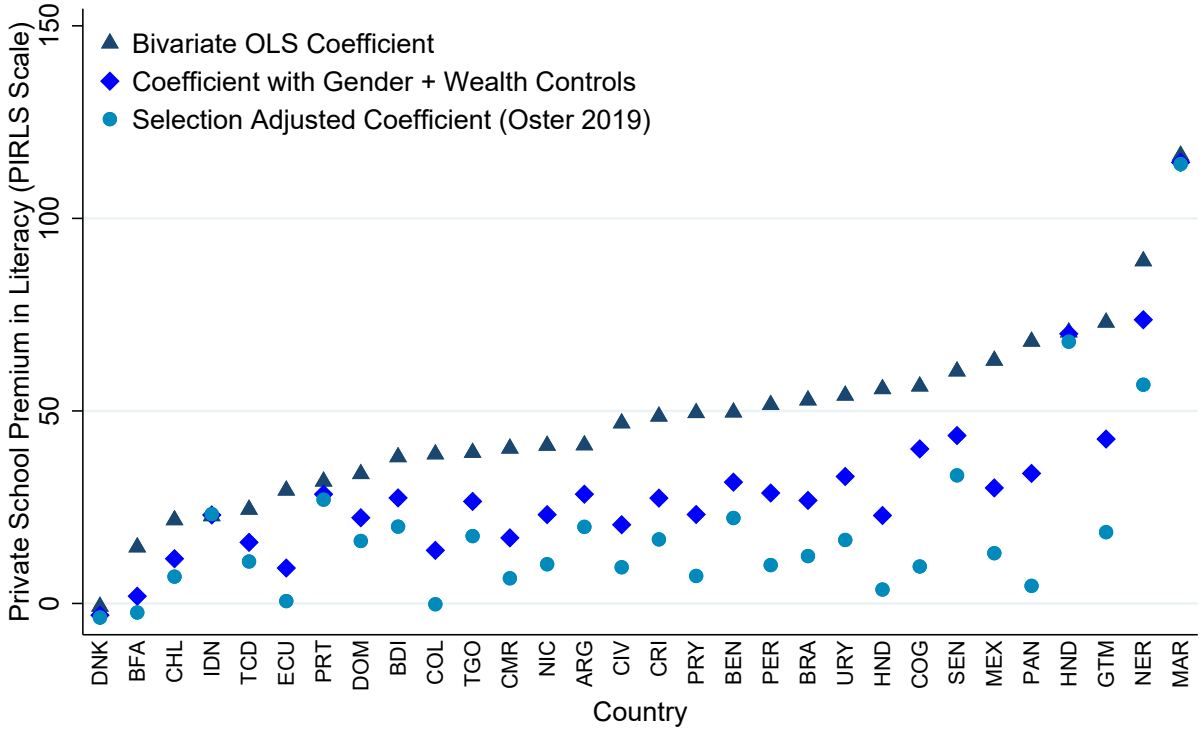
Note: Table 6 shows cross-country regressions between our expanded median test scores, child marriage rates, women's desired fertility, and income. Income data is measured in log GDP per capita in 2015 (PPP) from the World Bank's World Development Indicators. The test score gender gap is the coefficient on female in a bivariate regression with converted test scores for each country. The portion of women ages 20-24 who were first married by age 18 and the desired fertility data come from a variety of household surveys compiled together by the World Bank (World Bank, 2019). The expanded TIMSS and PIRLS scores add PASEC and LLECE countries onto the corresponding scales using our linking functions and grade adjustments. Robust standard errors are shown in parentheses.

Figure 16: The Private School Premium Across Countries

(a) Math

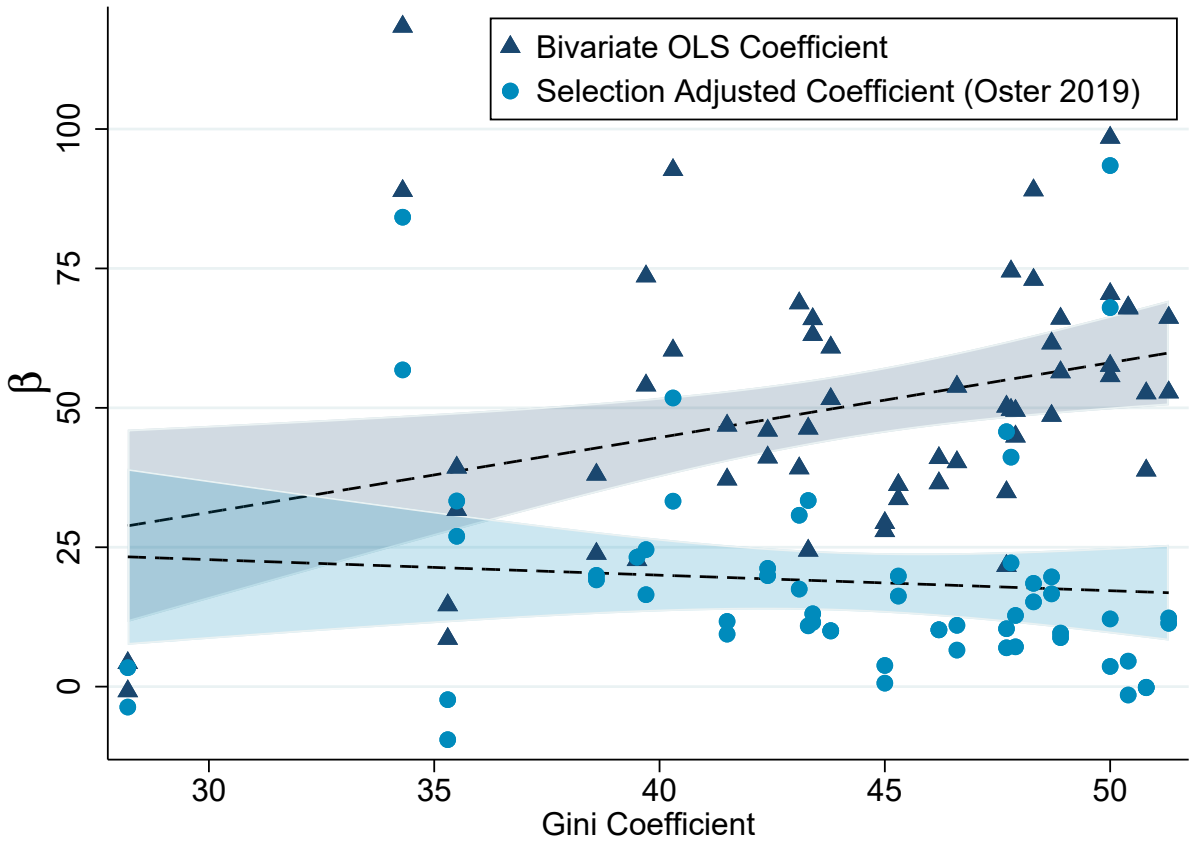


(b) Reading



Note: Figure 16 plots β_j^{Base}, β_j^{Obs}, and $\beta_j^{Adj.}$ for each country in our sample. Panel (a) shows these coefficients for math, and panel (b) displays the same for reading.

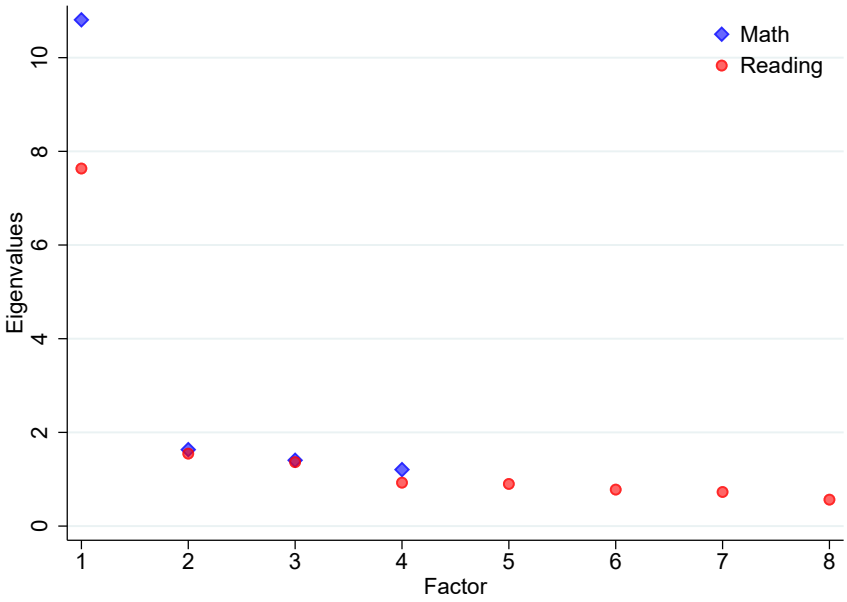
Figure 17: The Private School Premium and Inequality



Note: Figure 17 plots the private school premium in math and reading against the country's gini coefficient from the World Bank's World Development Indicators (World Bank, 2019). The triangles denote the private school premium from a bivariate regression of test scores on an indicator for private schools. The circles denote the coefficient adjusted for selection following the procedure from Oster (2019).

A Appendix

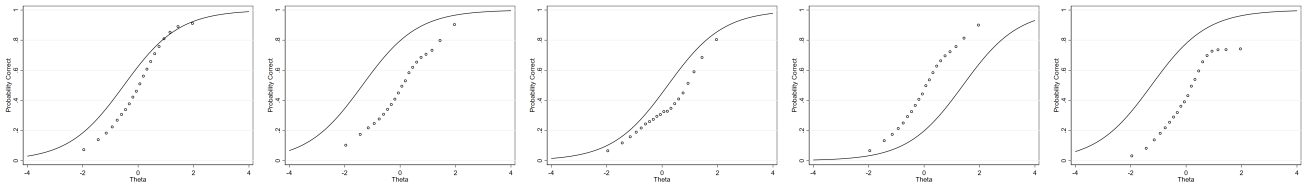
Figure A.1: Unidimensionality Test



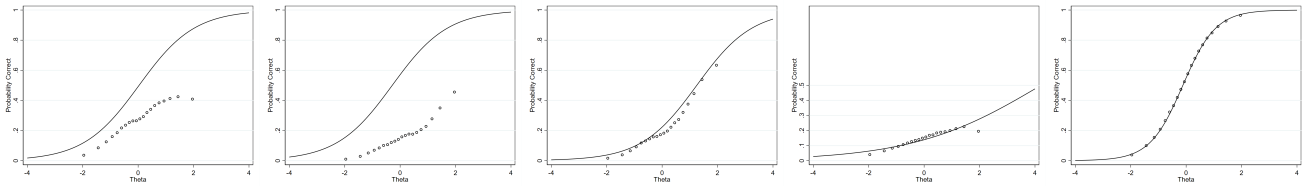
Note: Figure A.1 shows the Scree Plot for each principal component in a factor analysis of all math items and reading items. To estimate a single set of eigenvalues across all test booklets, an expectation-maximization algorithm was applied to the sparse item-wise covariance matrix. The nominal sample size was set to the average number of students who sat each item.

Figure A.2: Differential Item Functioning Plots

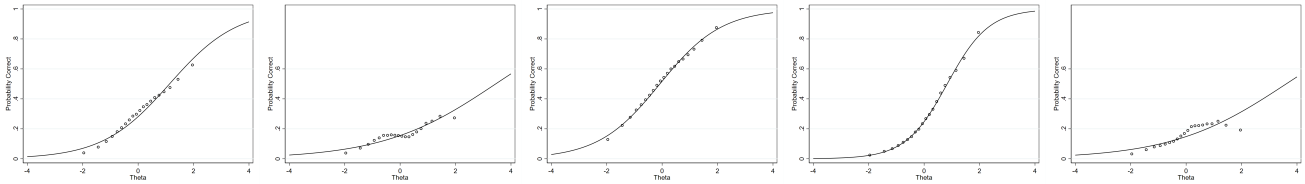
(1) LLECE Math: M3E2607C | (2) LLECE Math: M3G1603C | (3) LLECE Math: M3M2069C | (4) LLECE Math: M3N3236C | (5) LLECE Read: L31A01041L



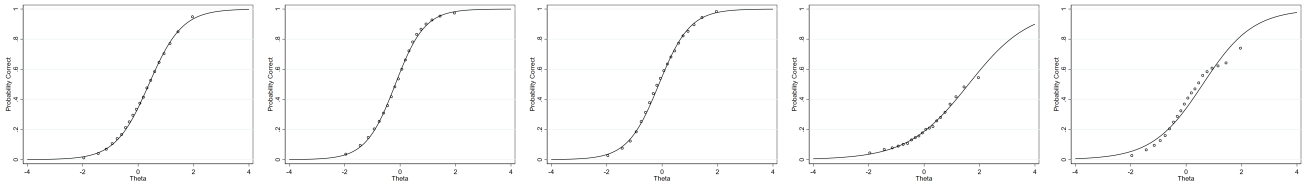
(6) LLECE Read: L31A01074T | (7) LLECE Read: L31B04064R | (8) LLECE Read: L32B09502Za | (9) PASEC Math: M2 | (10) PASEC Math: M22



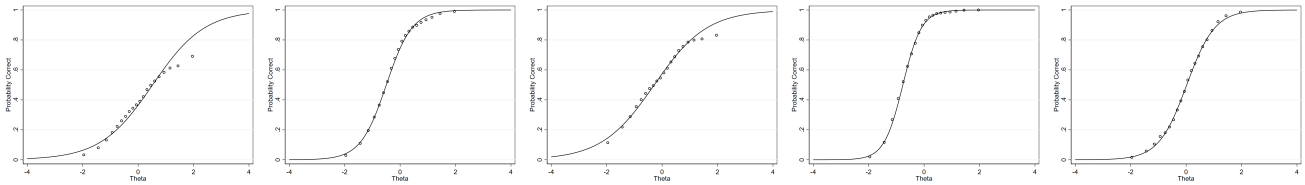
(11) PASEC Math: M4 | (12) PASEC Math: M45 | (13) PASEC Math: M49 | (14) PASEC Math: M54 | (15) PASEC Math: M59



(16) PASEC Math: M6 | (17) PASEC Math: M62 | (18) PASEC Math: M68 | (19) PASEC Math: M69 | (20) PASEC Math: M76



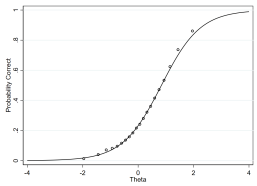
(21) PASEC Read: F36 | (22) PASEC Read: F37 | (23) PASEC Read: F49 | (24) PASEC Read: F54 | (25) PASEC Read: F55



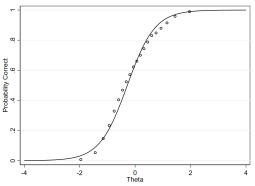
Note: Figures A.2 through A.5 show differential item functioning plots of the item characteristic curves (solid line) based on the reference test item parameters and the observed probability of portion correct (dots). The latter is estimated using five plausible values to take into account the uncertainty around the latent variable. Each student is included simultaneously into multiple bins of θ with probability proportional to the density of the student's latent trait distribution. For polytomously scored items, the general partial credit models are shown with a separate item characteristic curve for each of the possible scores. Lines and corresponding observed portions are shown in matching colors. The horizontal dashed lines for the three-parameter logistic functions denote the "guessing" line which corresponds to the expected probability of a correct random answer.

Figure A.3: DIF Plots — Continued

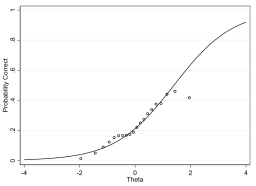
(1) PASEC Read: F56



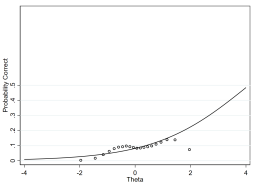
(2) PASEC Read: F57



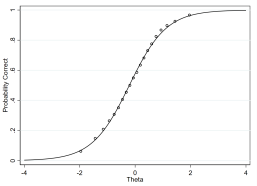
(3) PASEC Read: F58



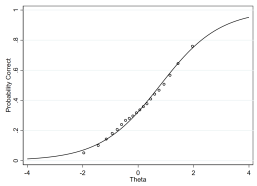
(4) PASEC Read: F59



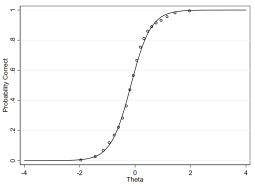
(5) PASEC Read: F82



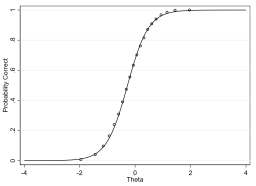
(6) PASEC Read: F84



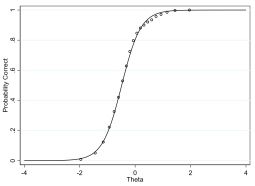
(7) PASEC Read: F85



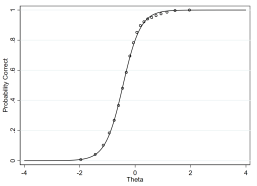
(8) PASEC Read: F86



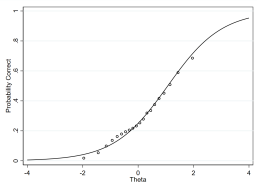
(9) PASEC Read: F87



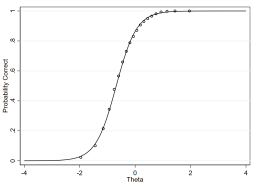
(10) PASEC Read: F88



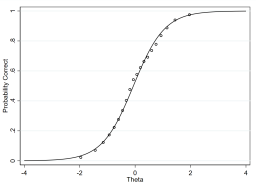
(11) PASEC Read: F89



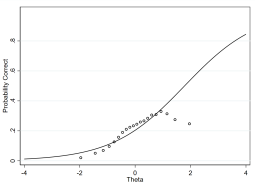
(12) PIRLS: R21E01M



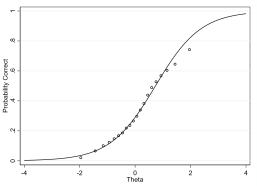
(13) PIRLS: R21E02M



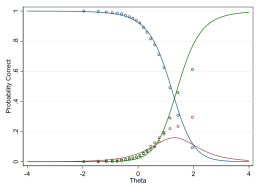
(14) PIRLS: R21E03M



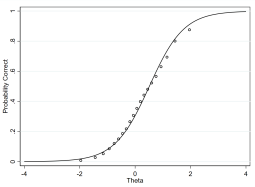
(15) PIRLS: R21E04M



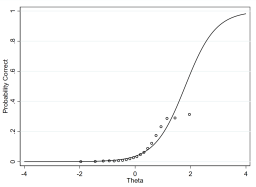
(16) PIRLS: R21E05C



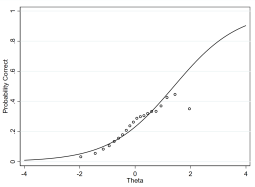
(17) PIRLS: R21E06M



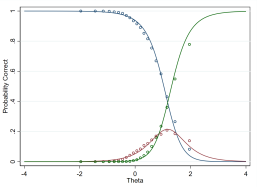
(18) PIRLS: R21E07C



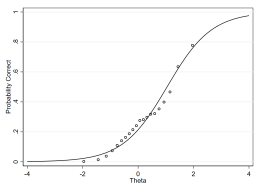
(19) PIRLS: R21E08M



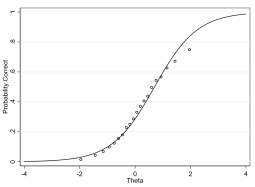
(20) PIRLS: R21E09C



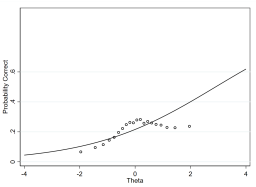
(21) PIRLS: R21E11M



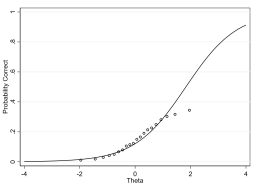
(22) PIRLS: R21N01M



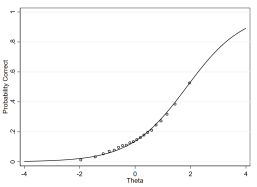
(23) PIRLS: R21N05M



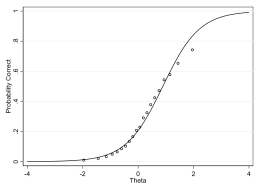
(24) PIRLS: R21N07M



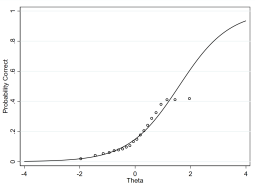
(25) PIRLS: R21N09M



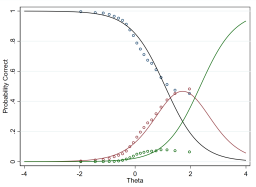
(26) PIRLS: R21N10M



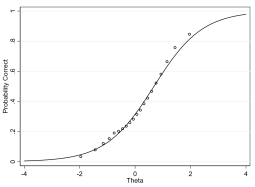
(27) PIRLS: R31G01M



(28) PIRLS: R31G02C



(29) PIRLS: R31G03M



(30) PIRLS: R31G05M

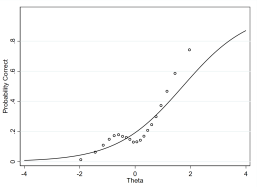
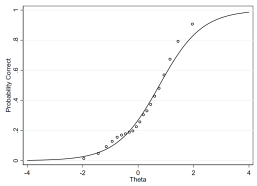
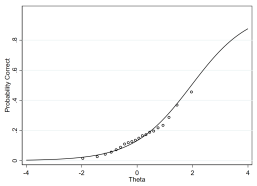


Figure A.4: DIF Plots — Continued

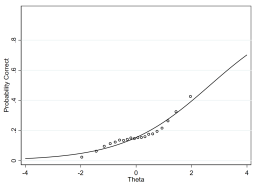
(1) PIRLS: R31G06M



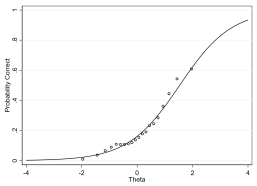
(2) PIRLS: R31G07M



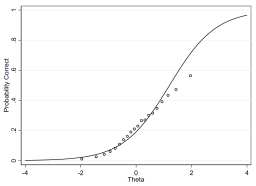
(3) PIRLS: R31G09M



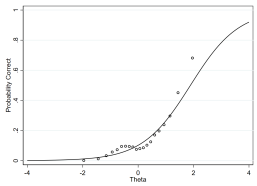
(4) PIRLS: R31G11M



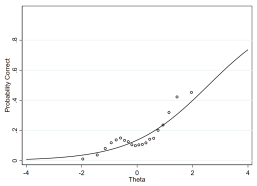
(5) PIRLS: R31G14M



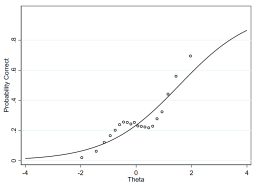
(6) PIRLS: R31P01M



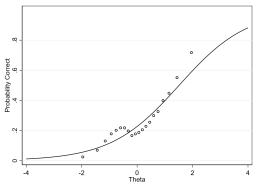
(7) PIRLS: R31P04M



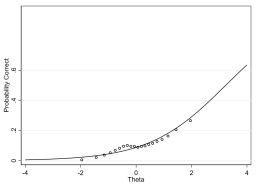
(8) PIRLS: R31P08M



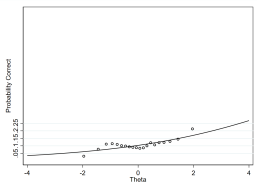
(9) PIRLS: R31P10M



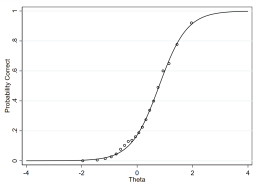
(10) PIRLS: R31P11M



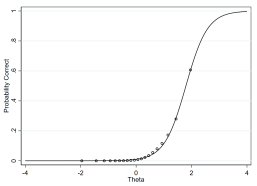
(11) PIRLS: R31P12M



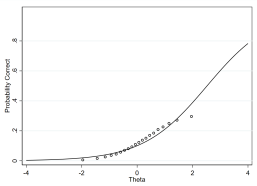
(12) PIRLS: R31P13M



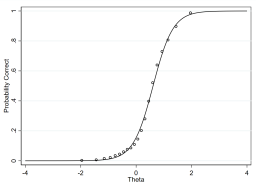
(13) TIMSS: M031009



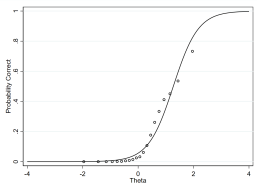
(14) TIMSS: M031071



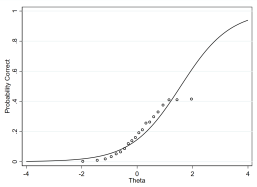
(15) TIMSS: M031079B



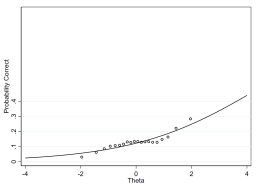
(16) TIMSS: M031079C



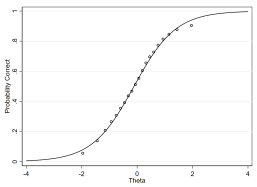
(17) TIMSS: M031088



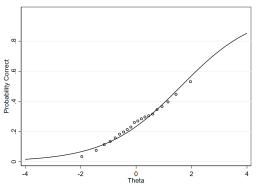
(18) TIMSS: M031093



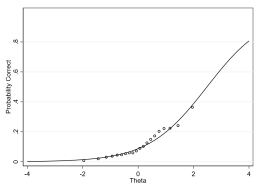
(19) TIMSS: M031128



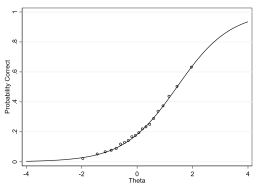
(20) TIMSS: M031155



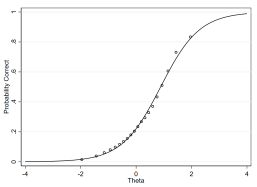
(21) TIMSS: M031159



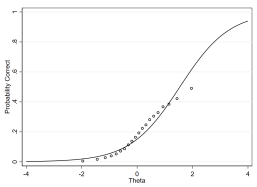
(22) TIMSS: M031185



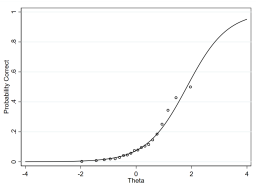
(23) TIMSS: M031187



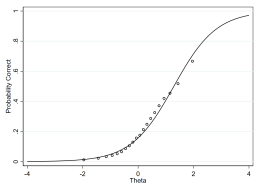
(24) TIMSS: M031210



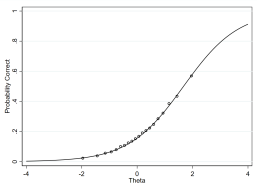
(25) TIMSS: M031218



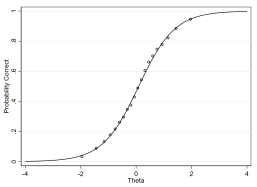
(26) TIMSS: M031252



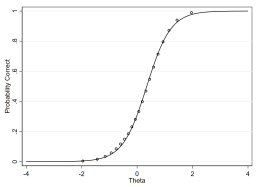
(27) TIMSS: M031294



(28) TIMSS: M031313



(29) TIMSS: M031316



(30) TIMSS: M031317

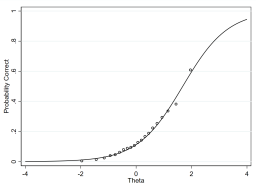
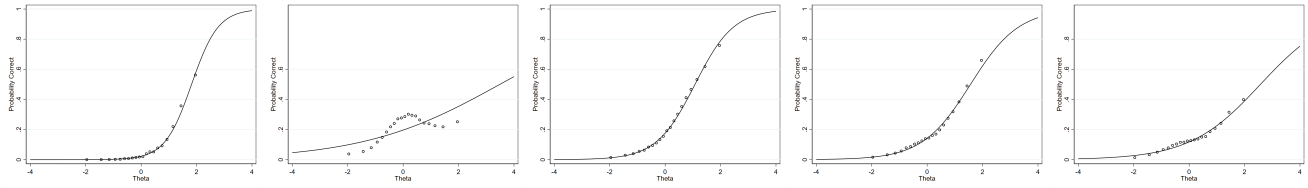
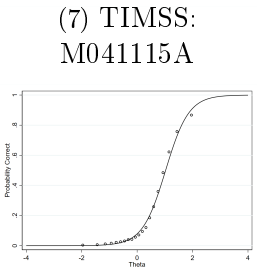


Figure A.5: DIF Plots — Continued

(1) TIMSS: M041003 (2) TIMSS: M041010 (3) TIMSS: M041011 (4) TIMSS: M041041 (5) TIMSS: M041098



(6) TIMSS: M041104



(7) TIMSS: M041115A

(8) TIMSS: M041115B

(9) TIMSS: M041122

(10) TIMSS: M041155

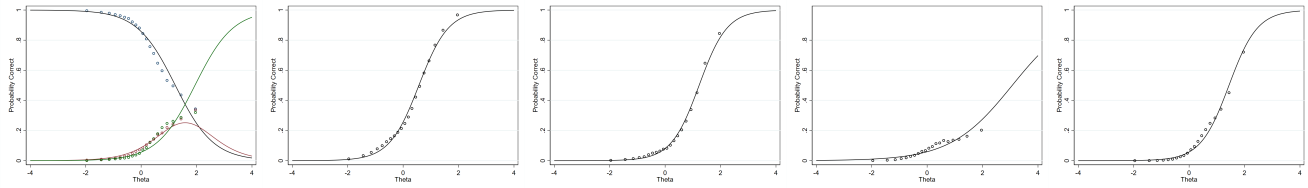
(11) TIMSS: M041160A

(12) TIMSS: M041175

(13) TIMSS: M041184

(14) TIMSS: M041265

(15) TIMSS: M041299



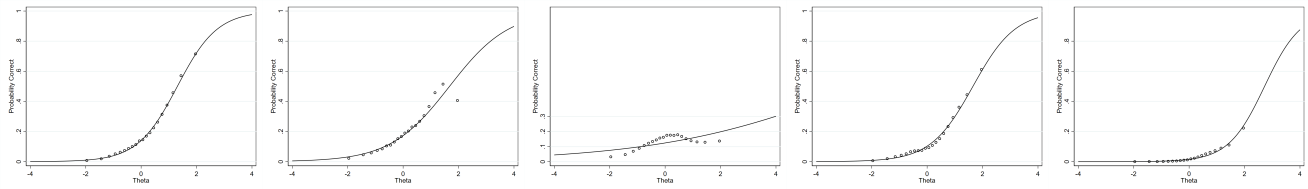
(16) TIMSS: M041320

(17) TIMSS: M041329

(18) TIMSS: M051007

(19) TIMSS: M051091

(20) TIMSS: M051109



(21) TIMSS: M051123

(22) TIMSS: M051203

(23) TIMSS: M051305

(24) TIMSS: M051601

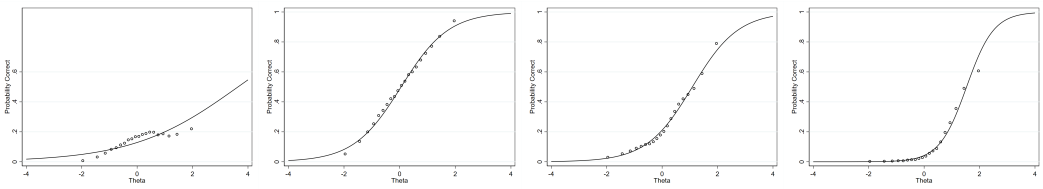
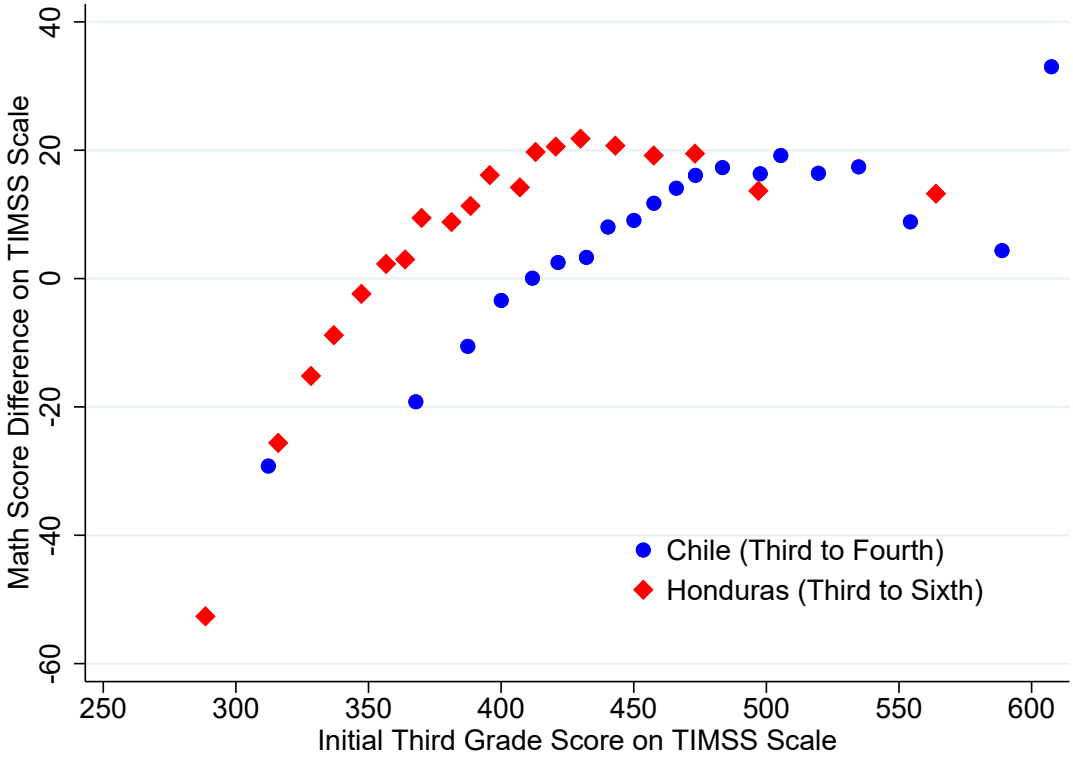
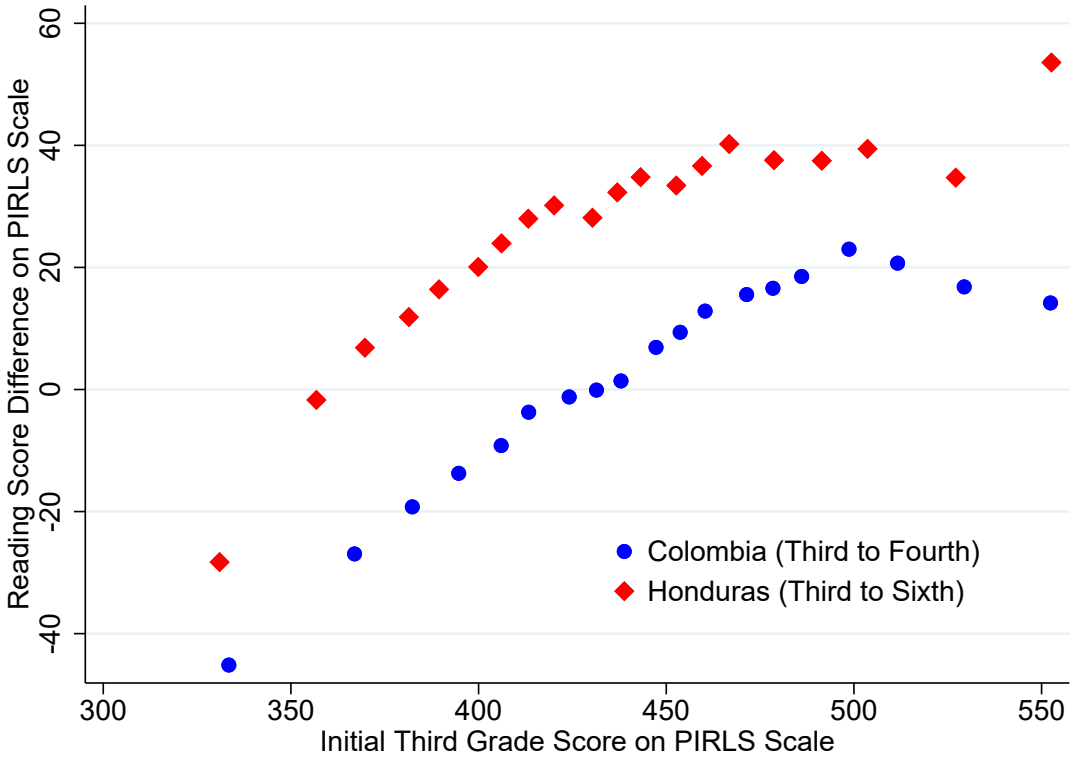


Figure A.7: Distribution of Grade Value-Added

(1) Math



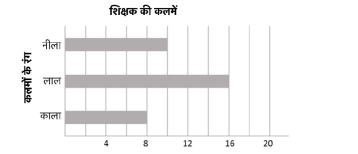
(2) Reading



Note: Figure A.7 shows the difference in average scores by ventile for the countries which sat both LLECE and either TIMSS or PIRLS. The mid-point between the highest and lowest score in each ventile is plotted on the x-axis. Scores were first converted from LLECE to either TIMSS or PIRLS using the conversion functions from the Bihar students.

Figure A.6: Examples of Items in the Combined Test

(1) TIMSS math



उपर दिया गया चार्ट शिक्षक के मेज में रखी लाल, काली और नीली कलमों की संख्या दर्शाता है। लाल कलमों की संख्या काली कलमों की संख्या से कितनी अधिक है?

Ⓘ 2 अधिक
Ⓘⓘ 4 अधिक
Ⓘⓘⓘ 6 अधिक
Ⓘⓥ 8 अधिक

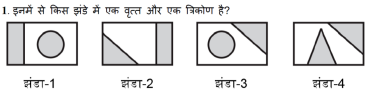
(2) PASEC math

एक अंतरिक्ष यात्री ने धरती से 15 जनवरी 2012 को सुबह 7 बजे उड़ान भरी। वह धरती पर वापिस 23 जनवरी 2012 को शाम 8 बजे लौटा। उसने अंतरिक्ष में कितना समय बिताया?

Ⓘ 7 दिन 20 घंटे
Ⓘⓘ 7 दिन 27 घंटे
Ⓘⓘⓘ 8 दिन 13 घंटे
Ⓘⓥ 8 दिन 14 घंटे

(3) LLECE Grade 3 math

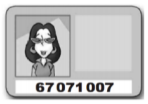
1. इनमें से किस झंडे में एक वृत्त और एक त्रिकोण है?



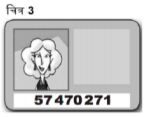
Ⓘ झंडा-1
Ⓘⓘ झंडा-2
Ⓘⓘⓘ झंडा-3
Ⓘⓥ झंडा-4

(4) LLECE Grade 6 math

मतिल्दा के पहचान पत्र की संख्या में दस लाख की 7 इकाइयाँ, हज़ार की 7 इकाइयाँ और दसियों की 7 इकाइयाँ हैं।

चित्र 1


चित्र 2


चित्र 3


चित्र 4


1. मतिल्दा से संबंधितदस्तावेज किस चित्र में है?

Ⓘ 1
Ⓘⓘ 2
Ⓘⓘⓘ 3
Ⓘⓥ 4

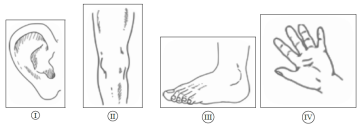
(5) PIRLS literacy

4. किसान को बाज का बच्चा कहाँ मिला?

Ⓘ उसके घोंसले में
Ⓘⓘ नदी के किनारे
Ⓘⓘⓘ पत्थर की कगार पर
Ⓘⓥ बेंत के बीच

(6) PASEC literacy

निम्न में से किस चित्र में एक पांव दिखाया गया है?



(7) LLECE Grade 3 literacy

प्रिय आद्रेस,

यह पत्र मैं तुम्हे टोक्यो, जापान की राजधानी, से लिख रही हूँ। मैं यहाँ दो महीने पहले ही आई हूँ और अपनी कक्षाओं से बहुत खुश हूँ। मेरे सहकर्मी बहुत अच्छे हैं और उन्होने मेरा स्वागत बहुत उत्साह के साथ किया। जैसे की तुम जानते ही हो, यहाँ सब कुछ अलग है। यहाँ का खाना बहुत स्वादिष्ट है पर मुझे ब्रेड और फलों के जूस की याद आती है जो हम साथ बनाया करते थे।

मेरे मुहल्ले में बहुत से पर्यटक और खिलाड़ी रहते हैं। जिस घर में मैं रहती हूँ वो काफी छोटा है और यहाँ अक्सर मौसम गरम रहता है। जिस परिवार के साथ मैं यहाँ रहती हूँ वे बहुत ही स्नेहमय हैं और उन्होने मुझसे स्पेनी भाषा सीखने की कोशिश भी की, परंतु अभी तक मैं उनसे इस भाषा में कुछ बुलवा नही पाई हूँ। मुझे यहाँ बहुत मजा आ रहा है।

अपने दोस्त की ओर से प्रेम।

लाउरा

1. कथन के अनुसार, लाउरा जापान में प्रमुख रूप से क्या करती है?

Ⓘ वह पढ रही है
Ⓘⓘ वह खेलती है
Ⓘⓘⓘ वह एक पर्यटक है
Ⓘⓥ वह स्पेनी भाषा पढाती है

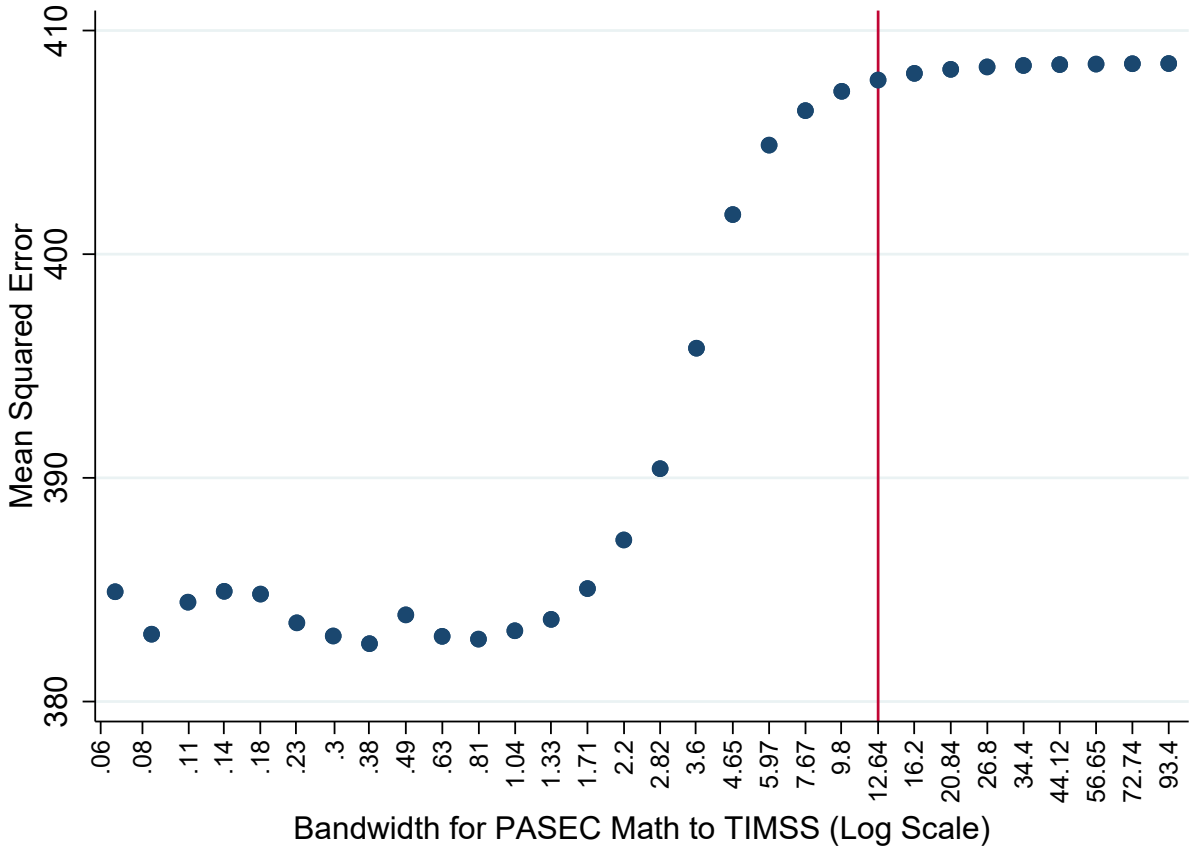
(8) LLECE Grade 6 literacy

1. उपर दिए गये मूल में, दादुर के कथन "कुछ देर बाद मुझे आकाश में एक बड़ा तारा दिखाई दिया और फिर सूर्य." का क्या अर्थ है?

Ⓘ कि उसने पूरी रात जग कर बिताई थी.
Ⓘⓘ कि एक तारे ने सूर्य को ढक लिया था.
Ⓘⓘⓘ कि उसे आकाश की ओर देखने में मजा आया था.
Ⓘⓥ कि एक अजीब घटना घाटी थी.

Note: Figure A.6 shows Hindi translations of sample items from each source test that were used in the Bihar assessment instrument.

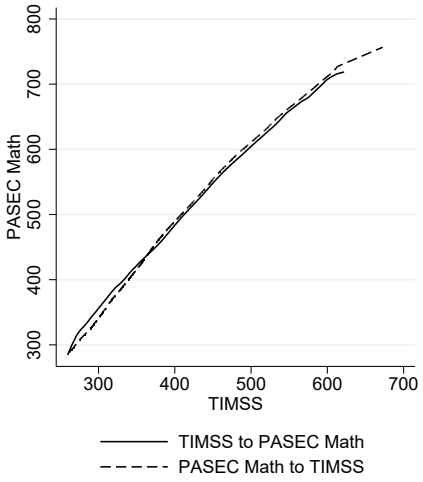
Figure A.8: Sample Bandwidth Comparison



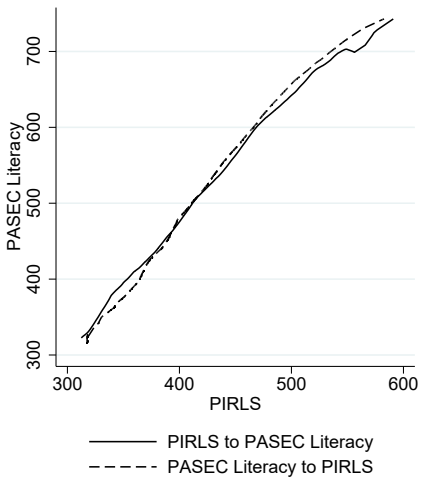
Note: Figure A.8 shows the mean squared error by bandwidth for a local linear regression of TIMSS on PASEC math scores using leave-one-out cross-validation. The points are plotted on a log scale. The red vertical line denoted the ROT bandwidth actually used, which is more conservative than the loss-minimizing choice but not by a particularly large amount.

Figure A.9: Symmetry of Test Equating Functions

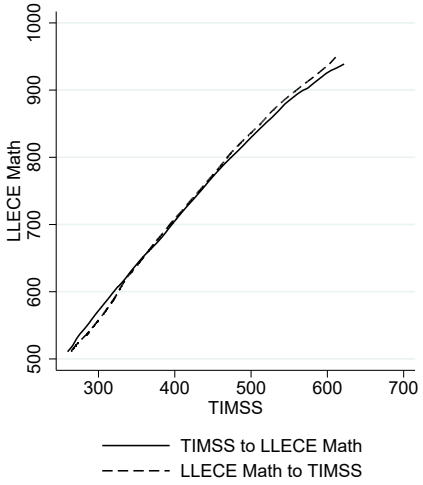
(1) TIMSS and PASEC



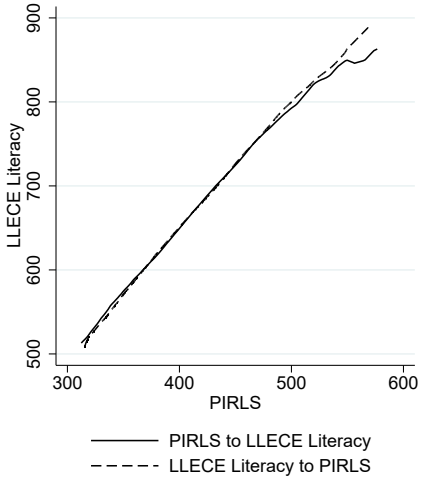
(2) PIRLS and PASEC



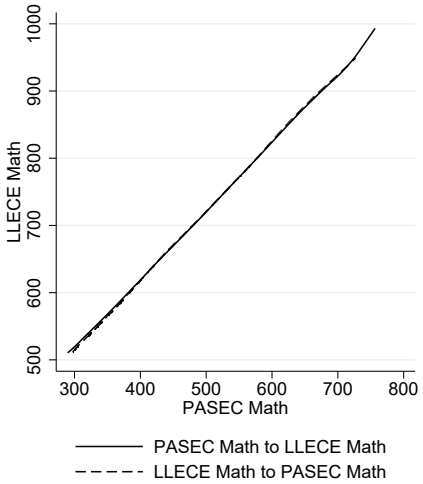
(3) TIMSS and LLECE



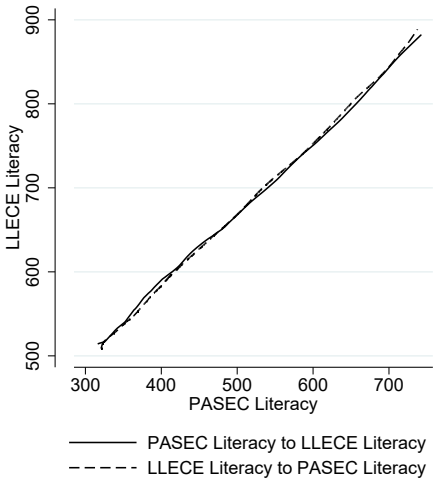
(4) PIRLS and LLECE



(5) PASEC and LLECE (Math)



(6) PASEC and LLECE (Reading)



Note: Figure A.9 shows the pairs of test equating functions for each dyad estimated with local linear regressions and a ROT bandwidth.

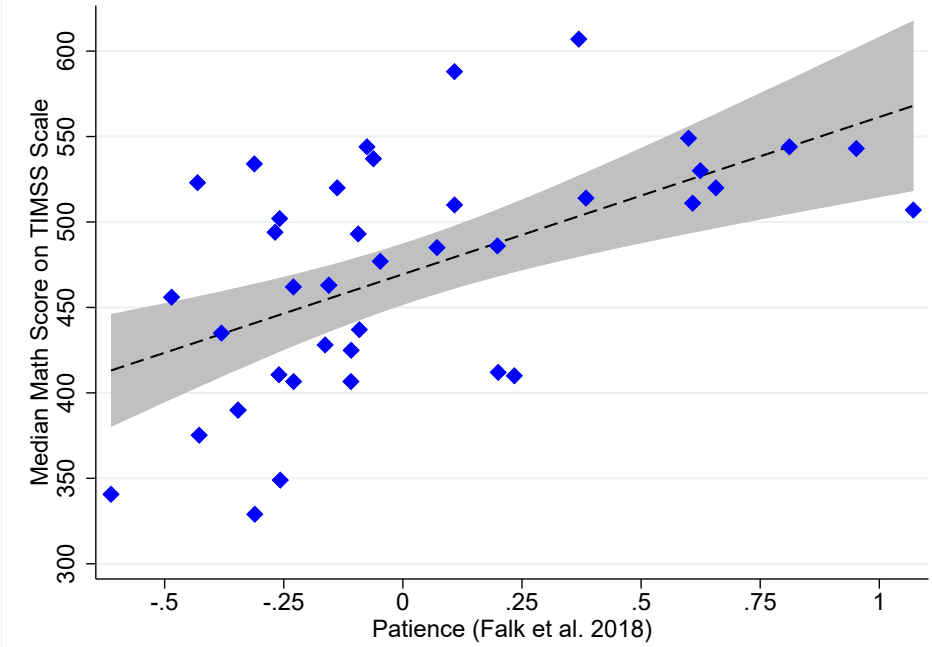
Table A.1: Symmetry Test

	(1) Original		(2) Converted Twice		T-test P-value
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)
PASEC Math	2314	468.488	2313	468.559	0.979
		(1.955)		(1.809)	
PASEC Reading	2314	480.378	2314	480.791	0.884
		(2.084)		(1.913)	
LLECE Math	2314	688.897	2313	688.901	0.999
		(2.008)		(1.886)	
LLECE Reading	2314	652.112	2314	652.290	0.943
		(1.774)		(1.719)	

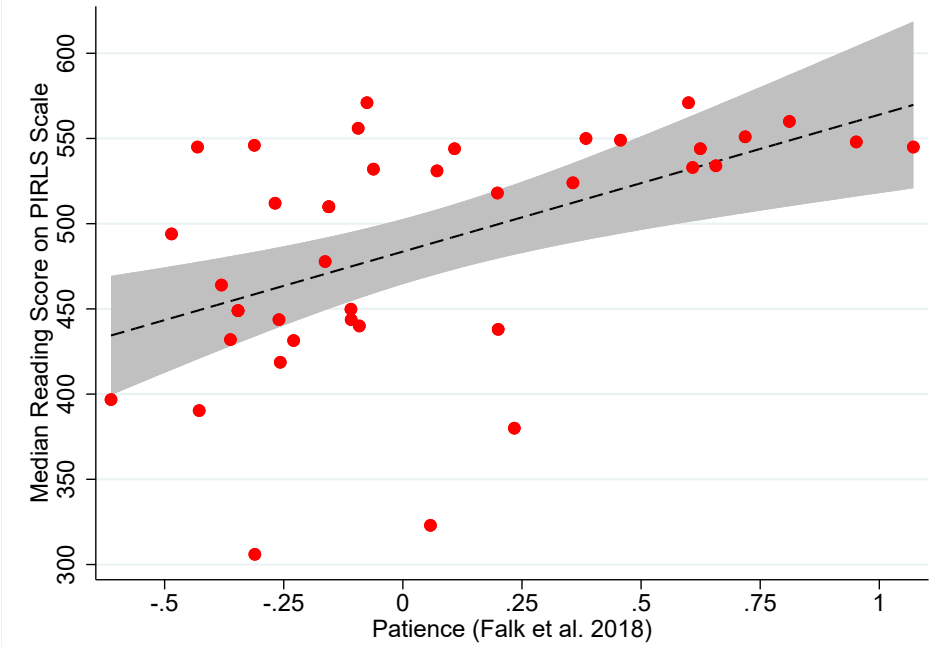
Note: Table A.1 presents a simple test of symmetry in our linking functions. Column (1) shows summary statistics for the original test scores of our students. Column (2) presents the same statistics after first converting them twice through the conversion functions. For the math outcomes, the original PASEC and LLECE scores are converted to TIMSS using the linking function in that direction, and then the the corresponding reverse linking function is applied to those estimates to recover new versions of the original scores. The same method is applied to the reading scores with PIRLS. The last column shows the p-value for a t-test on the two score distributions being different.

Figure A.10: Patience and Test Scores

(1) Math



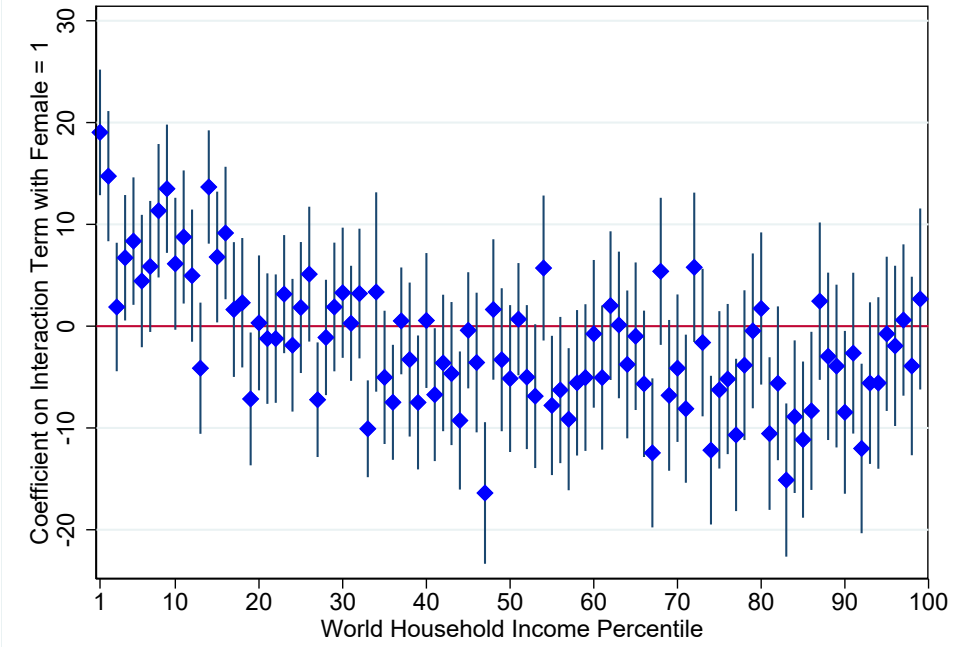
(2) Reading



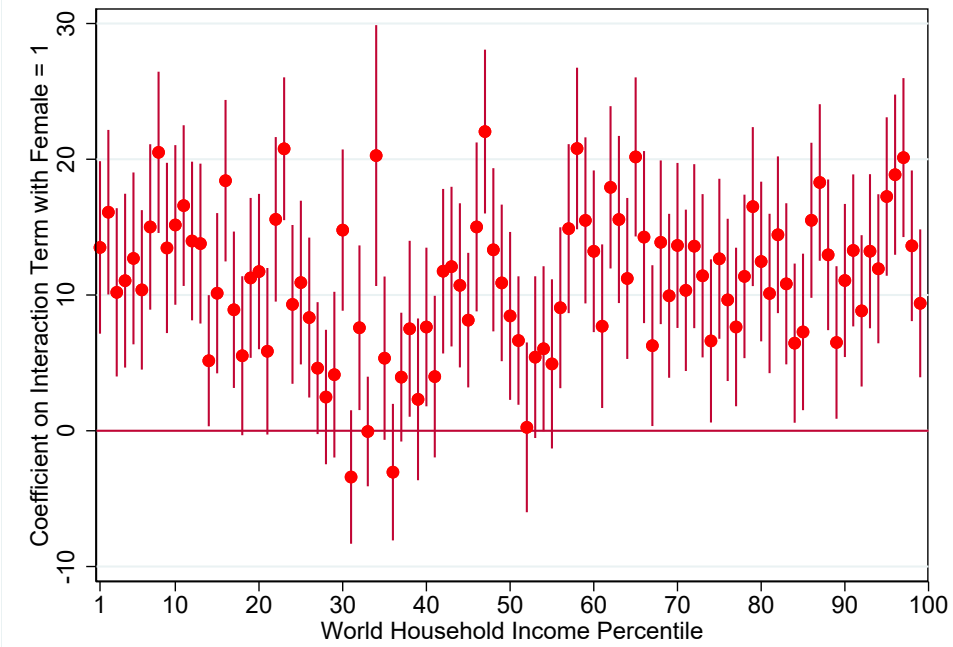
Note: Figure A.10 plots the correlation between patience, as measured by the Global Preferences Survey (Falk et al., 2018; Dohmen et al., 2018), and test scores converted to a common scale. Panel (1) shows math scores on the TIMSS scale, and panel (2) shows reading scores on the PIRLS scale. Dashed lines denote ordinary least squares lines of best fit, and the shaded areas denote 95 percent confidence interval.

Figure A.11: Gender Gaps by World Household Income Percentile

(1) Math

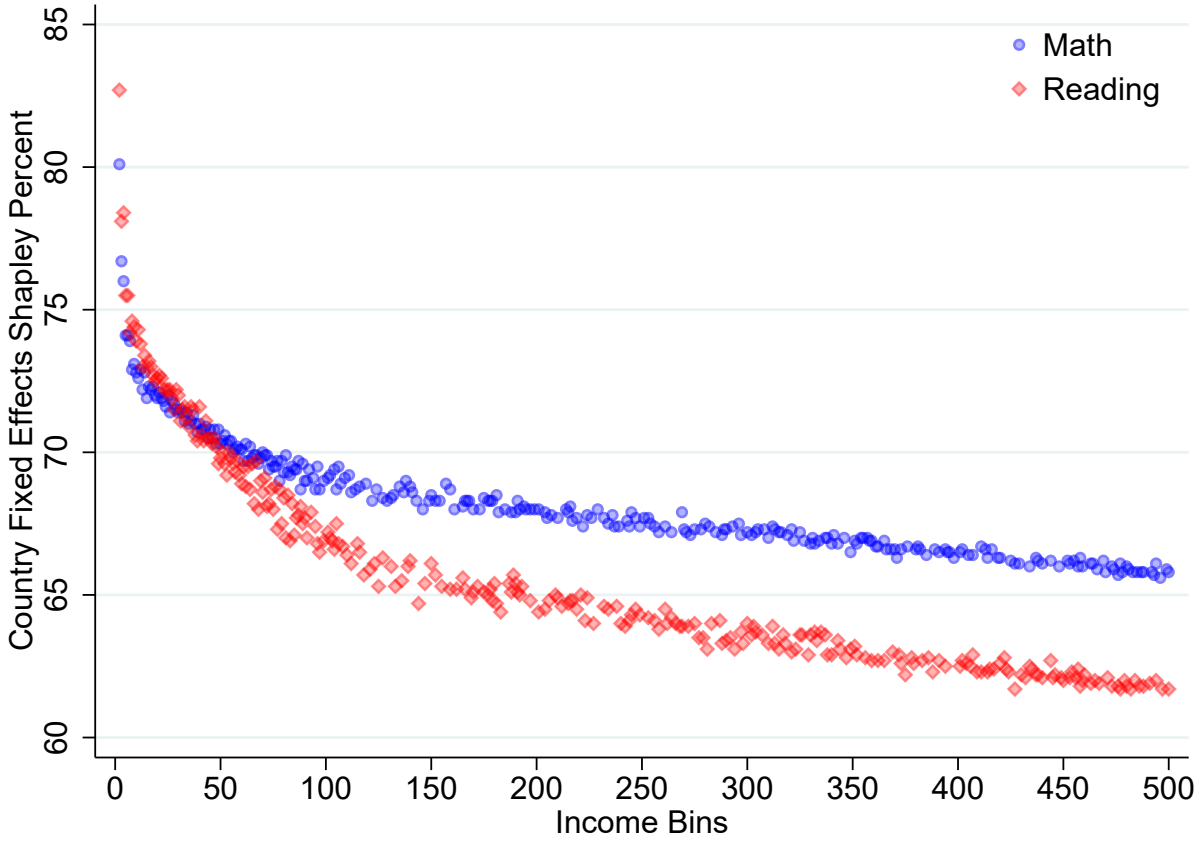


(2) Reading



Note: Figure A.11 plots the coefficients from a regression of a dummy indicating female interacted with students' rank in the global household income percentile, controlling for income percentile fixed effects. Panel (1) presents these coefficients for math, and panel (2) does the same for reading.

Figure A.12: Shorrocks-Shapley Decomposition of R^2 Across Specifications



Note: Figure A.12 blots the Shorrocks-Shapley decomposition of R^2 attributable to the country fixed effects across different specifications varying the number of world household income quintiles in the fixed effects.