WPS8592

Policy Research Working Paper 8592

Background Paper to the 2019 World Development Report

Global Dataset on Education Quality

A Review and Update (2000–2017)

Harry Anthony Patrinos
Noam Angrist

WORLD BANK GROUP
Education Global Practice
September 2018

Policy Research Working Paper 8592

Abstract

This paper presents the most expansive and current cross-country dataset on education quality. The database includes 164 countries/territories covering over 98 percent of the global population from 2000–2017. Of these countries, 111 are low or middleincome economies. Harmonized learning outcomes are produced using a conversion factor to compare international and regional standardized achievement tests. These tests include PISA, TIMSS, PIRLS, SACMEQ, LLECE and PASEC. In addition, this paper includes the Early Grade Reading Assessment (EGRA) for the first time. This enables extension of the database substantially. This methodological update paves the way to include a series of assessments that are increasingly common in developing countries that are often excluded from large international assessments. The database includes mean scores and disaggregates the data by subject, level, and gender. This paper further presents a series of methodological improvements including measures of uncertainty, fixed conversion factors for greater comparability over time and year-by-year data.

This paper—prepared as a background paper to the World Bank's *World Development Report 2019: The Changing Nature of Work*—is a product of the Education Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at hpatrinos@worldbank.org.

Produced by the Research Support Team

Global Dataset on Education Quality: A Review and Update (2000-2017)

Harry Anthony Patrinos

Noam Angrist[1]

Key words: Quality, Human Capital, Education, International, Achievement, Database

JEL Classification: C8, I2, N3, J24, O15

[1] We are grateful to Syedah Aroob Iqbal and Husein Abdul-Hamid who were essential contributors to this work. This work builds on co-authored work with Nadir Altinok. Particular thanks to Aart Kraay for detailed comments and contributions to the methodology and paper. Valuable comments were provided by Eva L. Baker, Eduardo Cascallar, Stefan Dercon, Deon Filmer, Roberta Gatti, Rachel Glennerster, Daniel Koretz, Silvia Montoya, George Psacharopoulos, Heiner Rindermann, Halsey Rogers, Jaime Saavedra, Shwetlena Sabarwal, Justin Sandefur, and Eduardo Velez. This paper benefited from multiple workshops held at the World Bank, World Congress of Cliometrics, American Economic Association, IDB, Oxford, USAID and FHI360. A great number of individuals and organizations supplied us with data. A special thanks to Rebecca Rhodes for access to many EGRAs. In the future, the database will be developed as part of a World Bank and UNESCO Institute for Statistics (UIS) partnership. Support from the World Bank's Research Support Budget is gratefully acknowledged. The views expressed here are those of the authors and should not be attributed to the World Bank Group.

1. Introduction

A country's human capital is critical for its economic success. A recent report found that human capital comprises the largest component of a nation's wealth, at 64 percent (World Bank 2018). In this paper, we focus on one measure of human capital: learning outcomes. Fifteen years of literature demonstrates that learning outcomes, as well as years of schooling, matter substantially for growth (World Bank, 2018; Hanushek and Kimko, 2000; Pritchett, 2001; Hanushek and Woessmann, 2008, 2012). This insight comes at a time when the availability of international student achievement tests (ISATs) is growing. These tests are psychometrically designed, standardized assessments of cognitive skills and enable credible cross-country and over-time comparisons of learning. However, traditional ISATs often exclude developing countries – the countries that have the most potential to gain from human capital accumulation.

In this paper, we address this gap by including the largest number of developing countries in a globally comparable database of learning outcomes to date. We build on a literature producing comparable measures of learning across countries and over time (Barro and Lee, 1996; Hanushek and Kimko, 2000; Barro and Lee, 2001, 2015; Altinok and Murseli, 2007; Hanushek and Woessmann, 2012, 2016; Angrist, Patrinos and Schlotter, 2013; Altinok, Diebolt and de Meulemeester, 2014; UIS, 2017; Altinok, Angrist and Patrinos, 2018). Our database covers 164 countries/territories from 2000-2017.

The central intuition behind the methodology we use is the production of a conversion factor between international standardized achievement tests (ISATs) such as PISA, TIMSS, and PIRLS and their regional counterparts (RSATs) such as SACMEQ, LLECE, and PASEC. This conversion factor is derived by comparing scores for countries that participate in both an RSAT and an ISAT in a given time period, schooling level (primary and secondary), and subject. Following Altinok, Angrist and Patrinos (2018), we call these countries "doubloon countries." A detailed description of each assessment is provided in Altinok, Angrist and Patrinos (2018). This approach puts all tests on a common scale enabling credible comparison across countries. These tests have been designed and scaled to be comparable over time since the late 1990s. Thus, there is no need for an intertemporal adjustment over the 2000-2017 interval.

This paper outlines a series of improvements over past databases. First, we extend our methodology to enable the first inclusion of the Early Grade Reading Assessment (EGRA). EGRA is a basic literacy assessment that has been conducted in over 65 countries since 2006. This enables us to extend our database substantially. We include 48 countries with data in the last ten years (2007-2017). Many of these countries are large developing economies previously excluded from globally comparable databases such as Ethiopia, Bangladesh, and Nepal. This methodological update paves the way to include a series of assessments that are increasingly common in developing countries and are often excluded from large international assessments.

We further expand our database with recent rounds of PASEC (2014), SACMEQ (2013) as well as PIRLS (2016). The size of our database has a few ramifications. Since the methodology we use to link assessments hinges on overlap in countries which take both an RSAT and an ISAT, the larger the database, the more overlap, and thus all scores are more robust. In addition, we include the largest number of developing countries to date. These countries have the most potential to benefit from educational progress. Thus, a more expansive database enables the inclusion of developing countries as well as enhances the robustness of the methodology used to include them, making each update significant.

In addition to the size of the database, this is the most current database with data as recent as 2017. Given a series of recent global initiatives which focus on education quality highlighted in the 2018 World Development Report, there is significant demand for current, credible and globally comparable measures of learning. This database provides the largest, most comparable, current learning data. Moreover, in this paper we provide a foundation for systematic and continual updates on a periodic basis going forward, enabling tracking of learning outcomes progress and longitudinal analysis.

We introduce a series of methodological improvements. First, we include single years, rather than 5-year intervals. Second, we use all available data to create conversion factors between assessments. Third, we construct a fixed conversion factor for data since 2000. This enables us to deduce changes in scores over time as a function of learning outcomes progress rather than changing conversion factors. Fourth, we provide measures of uncertainty in the form of plausible bounds for our estimates. Fifth, all data are derived directly from test score data. In contrast to black-box and complex imputation, this enables methodological transparency. Moreover, it produces a clear policy lever: when countries perform better on RSATs and ISATs they can be certain the HLO will improve.

2. Data

Each assessment included in prior databases is documented in depth in Altinok, Angrist and Patrinos (2018). The ISATs include old IEA assessments, PIRLS, TIMSS and PISA. The RSATs include SACMEQ, PASEC, LLECE and MLA. In this paper, we elaborate mostly on the inclusion of new data, including EGRA, as well as recent PASEC, SACMEQ and PIRLS data. Annex Table 1.0 shows all data included in this database.

2.1. EGRA

The Early Grade Reading Assessment (EGRA) is a basic literacy assessment. The assessment is conducted most often in grades 2-4. Since 2006, EGRA has been conducted in over 65 countries.

EGRA is focused on early grade proficiency on basic literacy skills. The EGRA testing framework is based on reading assessments such as DIBELS (Dynamic Measurement Group, 2008), CTOPP (Wagner et al. 1999), and a series of other assessments that measure literacy skills. The EGRA toolkit outlines the EGRA tool in depth (RTI

International, 2009). The assessment is a short oral assessment conducted with a child one-on-one. EGRA is designed to be flexible and adapted across countries and contexts, while maintaining core modules and similarities. EGRA is a timed test, enabling uniformity in how it is conducted. The tests often represent the most common features of the local language and align with the expectations of the grade level.

EGRA includes a series of sub-tasks that align with five phases of reading skills development: pre-alphabetic, partial alphabetic, alphabetic, consolidated-alphabetic and automatic. The subtasks in EGRA include:

- Orientation to print
- Letter name identification
- Letter-sound identification*
- Initial-sound identification
- Segmentation (phoneme or syllables)
- Familiar word reading
- Non-word reading*
- Oral reading fluency*
- Reading comprehension*
- Cloze
- Listening comprehension*
- Vocabulary
- Diction

Of these there is a sub-set of 'core' subtasks encouraged to be delivered across all countries and contexts marked with an asterisk above (Dubeck and Gove, 2015).

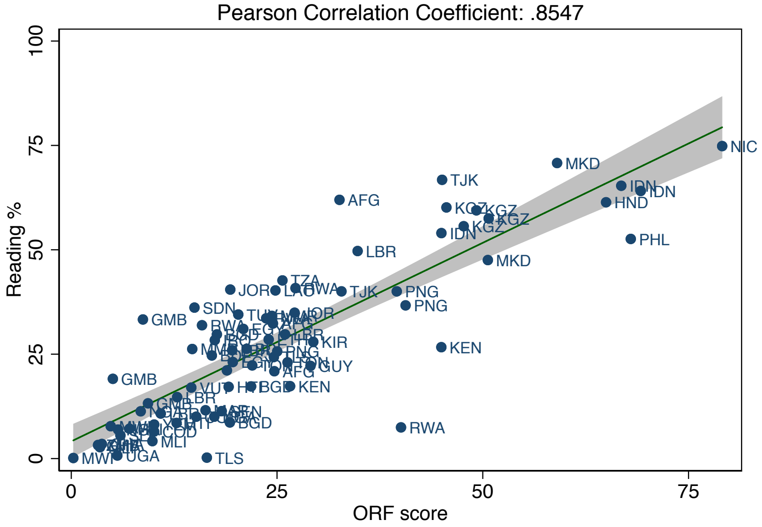
In our analysis, we analyze two specific EGRA sub-tasks: oral reading fluency (ORF) and reading comprehension. Both are core sub-tasks and therefore found across nearly all EGRA assessments. Moreover, both are conceptually linkable indicators supported by the theoretical and empirical literature.

The use of EGRA data hinges on using measures of *comparable* learning metrics, even if not perfectly *equivalent*. This approach is consistent with the approach taken by Dubeck and Gove (2015). Comparable measures must have a conceptual link to other reading assessments used in ISATs and RSATs. Moreover, indicators should be relatively invariant and robust to a myriad of changes that might occur across contexts such as language, timing, and details of implementation. For example, phonics is likely to differ across languages and is not conceptually linkable to other assessments. For this reason, we do not use the 'letter-sound identification' sub-task even though it is a core EGRA sub-task.

ORF scores have been used to make cross-country comparisons in the literature and have shown relatively high validity and reliability (Dubeck and Gove, 2015). Moreover, our two measures, comprehension and fluency, show high correlations with one another in related analyses in the range of 0.8 and 0.9 (Abadzi, 2011; Piper, 2010). Abadzi (2011) compares various measures on EGRA and refers to the tight correlation between reading fluency and comprehension, consistent with the literature (Piper, 2010; Vagh, 2009; Gove, 2009; Kim et al., 2010). Our own analysis corroborates this. Across 81 country-year observations, we find a correlation of 0.86 (see Figure 1.0).

Figure 1.0:

Correlation between Reading Comprehension and Oral Reading Fluency



Despite the viability of both measures, the oral reading fluency sub-task does not have as sound a conceptual link to RSATs and ISATs as comprehension. Moreover, fluency and words per minute used in ORF scores are likely to be more susceptible to differences in implementation and language. Reading comprehension, on the other hand, is not as tightly timed and is less sensitive to language and implementation details. Moreover, it has the most plausibly direct and substantive proficiency link to RSATs and ISATs. For example, reading comprehension on EGRA maps directly onto one of the levels in Levels 1-8 on SACMEQ, the RSAT for East and Southern Africa. Analysis and work done by ACER for the UIS Reporting Scale – a global effort to compare underlying proficiencies across assessments – focuses on reading comprehension partially for this reason (Turner et al., 2018). Abadzi (2011) also cites reading comprehension as a likely proficiency link between EGRA and ISATs and RSATs. This is born out in the data. For example, when we use reading comprehension scores, Tanzania outperforms Kenya, consistent with SACMEQ. When using ORF scores, Kenya does not outperform Tanzania, which is no longer consistent with SACMEQ. Moreover, reading comprehension was the most available sub-task across all EGRA datasets we collected.

To this end, we primarily analyze reading comprehension*,* which is available in nearly all EGRA datasets, is less sensitive to differences in context and implementation, and has a strong conceptual link to RSATs and ISATs. To ensure robustness to language effects, we only include data when students took the test in their language of instruction.

For each indicator, there is a threshold and continuous variable option. In theory, the threshold option ‘percent zero’ on either reading comprehension or oral reading fluency is most plausibly comparable. This measures the most basic foundational skills attained which should be relatively invariant to details which would affect performance higher up the learning margin. It is also easily interpretable and meaningful across contexts. However, since this variable is a threshold rather than a continuous measure of learning it is difficult to convert this to a score we can place on an international scale.

Instead we produce mean scores using a continuous variable that is plausibly comparable, the ‘percent of answers correct’ on reading comprehension. We use the percent correct since the number of test items varies slightly for each EGRA. If students did not attempt any question this is recorded as a zero. We further confirm that the threshold and continues variables are tightly correlated, with an inverse correlation of -0.9 (see Figure 2.0). We compute a standardized *Z*-score from the continuous reading comprehension metric and scale it with a mean of 500 and standard deviation of 100, akin to RSAT and ISAT scales.

EGRA data is designed for grades 2-4, although certain countries will participate out of this range. We restrict data used for our database to grades 2-4 to be consistent with the design of EGRA. This range further optimizes the coverage-bias tradeoff. If we include additional grade levels to expand coverage, the data starts to skew due to availability of data for certain countries coming from a particularly low or high grade (grades 1 or 6). If we restrict our interval to reduce grade bias, we start to lose countries. In the grade 2-4 range, we optimize this trade-off.

Figure 2.0: Correlation between Percent Zero Reading & Percent Reading Comprehension

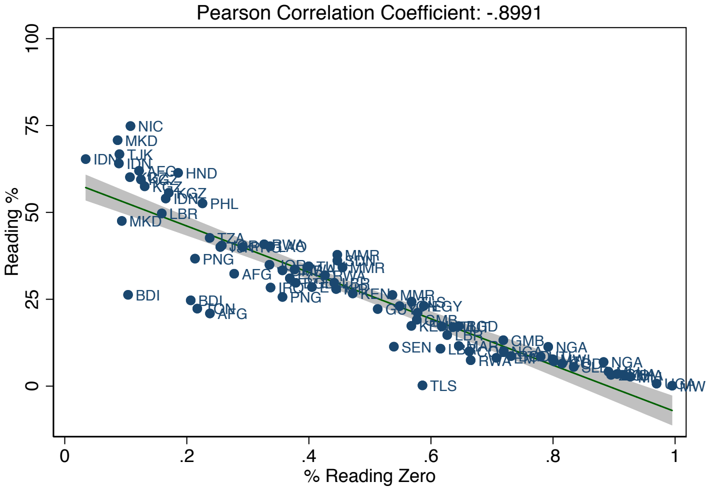
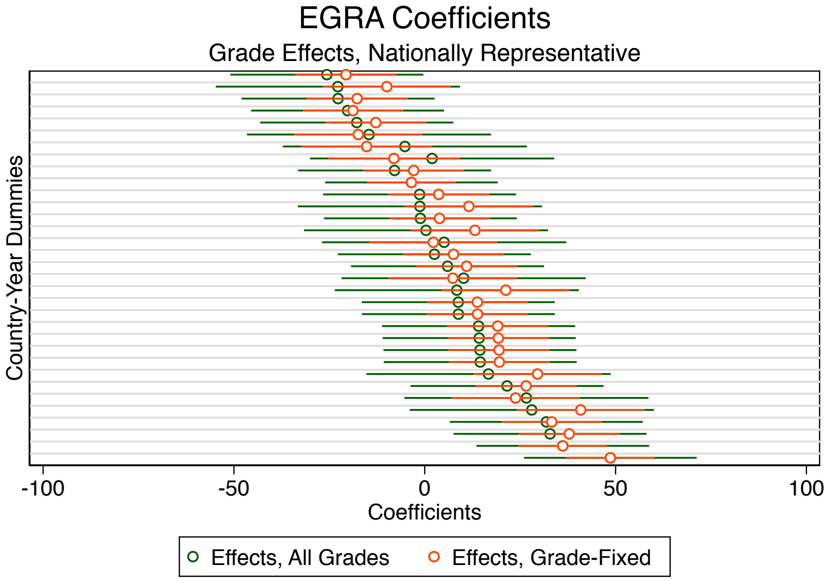


Figure 3.0 demonstrates sensitivity to grade within this interval. We run a regression with and without grade-fixed effects comparing mean scores relative to a given country which participates across all three grade levels 2-4. This gives us an indication of sensitivity to

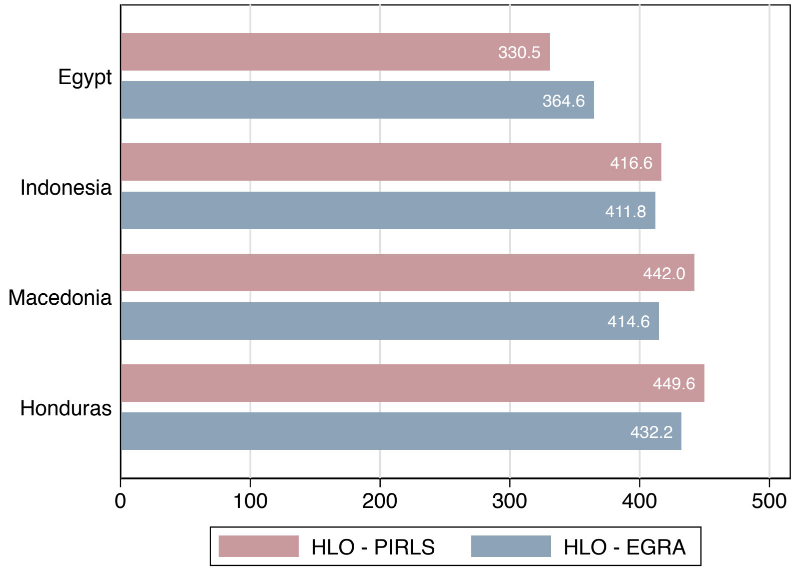
grade. We find small differences. Occasionally, we see slightly larger differences, however, in all instances the effect sizes fall within the confidence interval. This sensitivity analysis increases our confidence that the EGRA data is robust to data availability biases by grade.

Figure 3.0:



Overall, our inclusion of EGRA enables us to include over 48 countries with at least one data point in the last ten years. Moreover, it paves the way for inclusion of non-traditional assessments through careful choice of testing regime, indicators, language considerations, grade-level participation, and score scaling. Below we show sensitivity to this exercise comparing HLO scores when derived using EGRA versus PIRLS, an ISAT with scores produced using IRT. Since both assessments are conducted in reading at primary level, we can compare scores and expect similar results. We do this for countries that participate in both EGRA and PIRLS from 2000 onwards and take the average. This will be biased by data availability by year, for example, if Egypt participated in EGRA in 2013 and PIRLS in 2016. However, it gives us an indication of whether HLOs are reasonable when derived from EGRA. Figure 4.0 reveals that scores are roughly similar, within 5-28 points of one another. Moreover, we see that ranks are preserved: Honduras outperforms Macedonia, followed by Indonesia, then by Egypt.

Figure 4.0: Average HLO derived from EGRA and PIRLS, 2000-2017



2.2. PASEC (2014)

The most recent PASEC in 2014 uses Item Response Theory (IRT) and is a high-quality RSAT. Ten countries participated, including Benin, Burkina Faso, Burundi, Cameroon, Chad, Congo, Cote d'Ivoire, Niger, Senegal and Togo. We include these countries using available microdata. Madagascar also participated in 2015 and was scaled to the PASEC 2014 round. We include Madagascar in our database using estimates from reports. To this end, inclusion of the recent PASEC data adds countries as well as enhances the quality of data from sub-Saharan Africa. This marks a significant improvement over past datasets. To provide a link to past PASEC rounds, which use classical test theory and have substantially different test items, we create an inter-temporal comparison using Togo, which participated in all rounds of PASEC. However, given that PASEC did not conduct an intertemporal scaling calibration as is typical for IRT ISATs and RSATs, priority should be given to analyzing the most recent PASEC 2014 data. Any intertemporal comparisons should be conducted with caution.

2.3. SACMEQ (2013)

SACMEQ collected its fourth round of data in fifteen countries in East and Southern Africa from 2012-2014. These include: Botswana, Kenya, Lesotho, Mauritius, Malawi, Mozambique, Namibia, Seychelles, South Africa, Tanzania, Uganda, Zambia, Zanzibar and Zimbabwe. SACMEQ was designed and scaled to be comparable to past rounds. We include estimates from reports for SACMEQ since the microdata is pending.

2.4. PIRLS (2016)

The newest round of PIRLS includes 50 countries. It was designed to be comparable to prior rounds. Inclusion of PIRLS is straightforward and we use the same parameters as used in prior databases using the microdata. Overall, through the inclusion of recent rounds of PASEC, SACMEQ, and PIRLS as well as the new inclusion of EGRA, we are able to improve the quality of, and include recent data for, over 75 countries since Altinok, Angrist and Patrinos (2018).

3. Methodology

3.1. Linking Methodology

We build on a literature linking learning assessments starting with Barro and Lee in 2001. Our methodological approach in this paper extends and improves on the one detailed in Altinok, Angrist and Patrinos (2018) to build the most current database covering the largest number of countries to date. Our Harmonized Learning Outcomes (HLO) database includes 164 countries/territories from 2000-2017.

Various methodologies can be used to make assessments comparable. We do not attempt to *equate* scores. Rather, we *link* scores to achieve comparability (Kolen and Brennan, 2014; AERA, APA, NCME, 1999; Holland and Dorans, 2006; Linn, 1993; Mislevy, 1992). This distinction is important, since we do not claim tests or test scores are equivalent across tests. Rather, we aim to compare scores on a similar scale.

We do not link using Item Response Theory (IRT) – the technique used to generate scores for each respective international and regional assessment. IRT models the probability a given pupil answers a given test item correctly as a function of pupil and item-specific characteristics. While this methodology is used within each of the international and regional tests we use, to use it across ISATs and RSATs would require overlap in test items. This is not true for a significant enough set of tests and time intervals to create a globally comparable panel dataset. Moreover, even when there is overlap, for IRT to be reliable there must be large item-specific overlap. When this overlap is small, standard maximum likelihood estimates will reflect both true variance and measurement error, overstating the variance in the test score distribution. Das and Zajonc (2010) elaborate on the various challenges of estimating IRT parameters with limited item-specific overlap.

It is possible to empirically test the conditions under which IRT produces reliable estimates by examining differential item functioning (DIF). Sandefur (2016) equates SACMEQ and TIMSS with IRT methods. Sandefur (2016) measures the DIF as the distance between the item-characteristic curve (ICC) for the reference population and actual responses for the focal group, an approach first proposed by Raju (1988). The resulting DIF is high, casting doubt on the IRT approach in a context with limited item overlap, which is the case for most RSATs we aim to include in our database.

While IRT might not be a reliable approach when there is limited item-by-item overlap, we conduct a few robustness tests where overlap is larger. We compare our results to the *Linking International Comparative Student Assessment* (LINCS) project which uses IRT methods and has significant overlap in items for a subset of international studies focused on reading at primary school from 1970 onwards (Strietholt, 2014; Strietholt and Rosén, 2016). If our expanded HLO database can produce similar results to original scores and IRT methods where there is overlap, we gain confidence in our results as well as an expanded dataset.

We note that while mean scores might vary by linking methods, and should be caveated appropriately, ranks and relative performance are relatively robust. While Sandefur (2016) finds large variation on mean scores depending on the equating method chosen, the Spearman rank correlations of the country averages are .97 or higher.

In building globally comparable education quality estimates, we rely on classical test theory (Holland and Hosken, 2003). Suppose that a population of pupils, sampled from the target population *T*, takes two different assessments *X* and *Y*. Here, we suppose that any differences in the score distributions on *X* and *Y* can be attributed entirely to the assessments themselves, since group ability is assumed to be constant.

The goal of linking is to summarize the difference in difficulty between two tests *X* and *Y*. We would like to link test *X* , the *Anchored Test*, to the scale of test *Y*, the *Reference Test*. For example, if we want to link *PISA 2003* to *TIMSS 2003*, *PISA 2003* is the *Anchored Test X* while *TIMSS 2003* is the *Reference Test Y*.

Our main estimation method is ratio linking. To conduct ratio linking we index the difference in means in the *Anchored Test X* and *Reference Test Y* as follows:

$$(1)\ linking_Y^{pl}(X) = y = \frac{\mu(Y)}{\mu(X)}x$$

where $linking_Y^{pl}(X)$ is the equation for converting observed scores on *Anchored Test X* to the scale of *Reference Test Y*. A limitation of this methodology is that it is sensitive to the scale of each test being similar. We address this potential limitation by using tests with a mean of 500 and standard deviation of 100.[2] This ensures that ratios capture differences in test difficulty rather than differences in scaling.

Below we specify each step in the process. The conversion rate is calculated as follows. Mean scores for tests X and Y, $\mu(X)$ and $\mu(Y)$, are first calculated for a given round *r*, subject *s* and level *l*. We consider tests to be in the same round if they are five years apart and optimize to have the rounds as tight as possible. Most often this translates to within one to two years. In some cases, this extends to three to five years apart. In a few

[2] In rare cases where tests do not have initial scores on this scale, we rescale the test using the microdata.

exceptions, we average adjacent years across one another as outlined in Annex Table 2. This minimizes the likelihood that test differences are a function of time, proficiency, schooling level, or data availability and are an accurate reflection of difficulty across tests. *r* is the number of testing rounds over the time period of the fixed conversion rate. We then calculate the fixed conversion rate d_{sl} from test X to test Y in a given subject and level such that:

$$(2)\quad d_{sl} = \frac{1}{r}\sum_{r=1}^{r}\frac{\mu(X_{rsl})}{\mu(Y_{rsl})}$$

where

$$\mu(X_{rsl}) = \frac{1}{n_{rsl}}\sum_{i\ \in Y_{rsl}\cap X_{rsl}} X_{irsl}$$

and

$$\mu(Y_{rsl}) = \frac{1}{n_{rsl}}\sum_{i\ \in Y_{rsl}\cap X_{rsl}} Y_{irsl}$$

where *i* is a country in the set of n_{rsl} doubloon countries that participate in both tests X and Y in a given round, subject, and schooling level ($Y_{rsl} \cap X_{rsl}$).

We then apply this conversion rate to a given country *j* that participates in test X but not test Y to produce a harmonized score:

$$(3)\ linking_{Y}^{pl}(X_{jysl}) = Y_{jysl} = \frac{X_{jysl}}{d_{sl}}$$

where *y* is the official year of anchored test X.

When producing aggregate scores across levels and subjects, we average across subjects and then across levels, weighting each equally.

A notable advantage of ratio linking is that it is easy to interpret and policy-friendly. This enables construction as well as consumption of comparable learning outcomes.

Linking methods include mean, linear linking and equipercentile linking, explained in detail in Altinok, Angrist and Patrinos (2018). A common linking method is linear linking – a shift in scores using the difference in means (rather than the ratio of means). This is an additive adjustment that accounts for standard deviations in a given testing round. Hanushek and Woessmann (2012) adopt this approach for OECD countries assuming standard deviations are constant across testing rounds. Hanushek and Woessmann acknowledge this assumption and construct an 'OECD Standardization Group' of countries with stable enrollments, ISAT participation, and standard deviations over time. However, this assumption is tenuous and unlikely to hold for developing countries, where standard deviations change significantly across testing rounds given growing enrollments and changing participation of countries across rounds. To this end, while ratio linking does not account for standard deviations, itself a limitation, it is also not biased by them.

Altinok, Angrist and Patrinos (2018) calculate scores using linear linking as a robustness check. Below is a comparison of the HLO scores produced using linear versus ratio linking. Figure 5.0 shows that the difference is centered around 0 with a small percentage of scores falling within a 10-25-point difference. When we compare ranks in Figure 5.1. we see an even smaller difference, showing similar empirical results.

Figure 5.0:
Difference between HLO scores
(2000-2015)
(linear vs. ratio method)

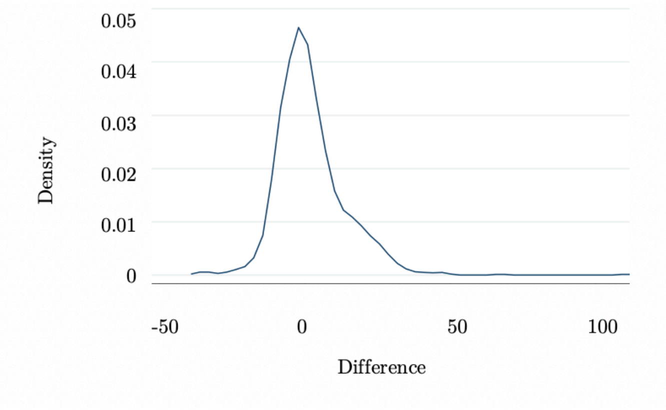
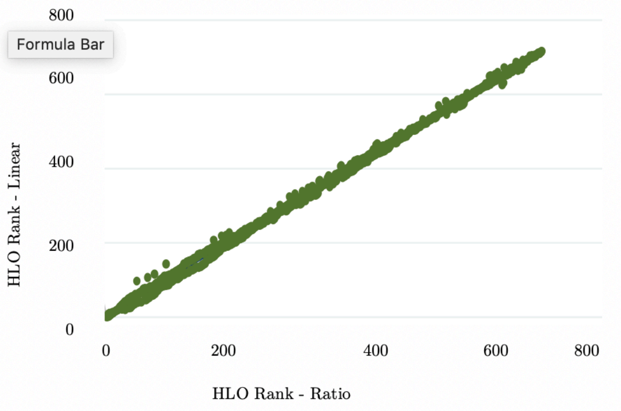


Figure 5.1:
Difference between HLO ranks
(2000-2015)



3.2. Assumptions

Three overall assumptions must hold for the linking methods above to be valid. First, linked tests must test the same underlying population. We satisfy this assumption by using sample-based ISATs and RSATs which are nationally representative for the population of students in school. We link tests using overlapping countries, which we call *doubloon countries*. This ensures that the underlying population tested is the same, and we capture differences between tests. An alternative approach to linking would involve the same items across different tests. Given there is limited item overlap, we adopt the approach using population-overlap across tests. A related assumption here is that participation rates reach a certain threshold or that non-participation is random. In cases where countries do not have nationally representative data, we ensure the linking function is based on data that is nationally representative and caveat the data appropriately. For EGRA, only a sub-set of countries have nationally representation data. Of 72 country-year observations, 42 are nationally representative. We produce the linking function based on countries that took part in nationally representative tests to guarantee population overlap. We apply the linking function to all countries with nationally representative EGRA results producing internationally comparable results. We also apply this linking function to non-national EGRA assessments. While we have confidence in the former estimates, we caveat the latter estimates. We exclude TIMSS results for Yemen which TIMSS highlighted as having a flooring effect and has an unrealistically low score. Our HLO Yemen score is thus derived from EGRA data. For Sri Lanka, we use a score derived from the National Assessment which was linked to TIMSS (Dundar et al., 2017). In the case of China and India we validate or extrapolate nationally representative data, described in Annex 3 and 4.

Second, tests should measure similar proficiencies. We link across precise subjects (math, reading and science) to ensure proficiency overlap. For EGRA, since it often has multiple languages of instruction, we only include scores for tests where the language of instruction is the same as the language of the test. This ensures we capture underlying cognitive skills rather than language effects.

Third, test differences should capture difficulty differences rather than country-fixed effects. We address this assumption in a few ways. As noted above, we link tests across *doubloon countries* that participate in tests that are nationally representative. The reliability of this linking exercise increases the larger the set Y ∩ X, the number of countries that take both tests being linked. Each update increases this overlap, making the inclusion of new countries in this update significant.

A novel contribution of this update which increases the reliability of this last assumption is the fixing of our conversion rate across a 15-20-year interval for recent years from 2000-2017. This enables us to increase the sample size in the set Y ∩ X for each doubloon index. For example, if Botswana participates in SACMEQ I and II as well as TIMSS 2003 and TIMSS 2007, we now have four data points instead of two, and two rounds of conversion factors instead of one between SACMEQ and TIMSS. In fixing the conversion factor, we assume that the relationship between tests stays constant across rounds. This assumption is reasonable and is true by design for RSATs and ISATs since the mid 1990s. This is not the case prior to the 1990s, limiting our ability to apply a fixed conversion factor to earlier data. The fixing of conversion factors increases the likelihood that we capture test-specific rather than country-specific differences. It further guarantees that any changes in test scores over this interval are due to realized progress rather than changing conversion factors.

3.3. Additional Methodological Parameters

Over-time comparability. ISATs and RSATs have been designed to be comparable since the late 1990s and early 2000s. Thus, the use of these modern assessments enables comparability over time.

Time Intervals. While this is one of the largest and most comprehensive comparable learning outcomes database produced to date, it is still sparse given limited test frequency. In past databases we smoothed time intervals over 5-year periods. This produced continuously spaced intervals, was designed to reduce noise by averaging results within these intervals and was comparable to the Barro-Lee approach for years of schooling. In this update, we have moved away from this approach. We now provide the year of the Reference Test X as documented in official reports. Since the doubloon index is a fixed conversion rate over a given interval, we produce a score for Anchor Test Y in the same year as Reference Test X. This enables greater granularity and precision of the data and enables the users of the database to make trade-offs at their discretion.

Schooling Levels. We construct a score for each grade and pool across grades within each schooling level to produce primary and secondary school scores. We include the default grade for TIMSS as the anchor grades across levels: grade 4 for primary and grade 8 for secondary. If we have data for additional grades, we include the closest grade with a preference for older grades since learning trajectories are steepest in the early years of learning introducing more scope for grade bias at earlier grades. If we have multiple test scores, we include the closest grade, and exclude the rest, rather than averaging. For example, we include PASEC grade 5 scores and exclude grade 2 scores. We include SACMEQ grade 6 scores along with PIRLS grade 4 scores at primary. If the test is designed for an age group (for example, PISA) we code it at the schooling level (for example, secondary for PISA).

Conceptually, the broader categories of primary and secondary scores enable us to categorize learning at critical schooling levels across assessments which span multiple grades and age groups. At the same time, we specify an approach to including specific grade levels to ensure we have a tight grade interval to minimize scope for grade-specific differences. While the interval is relatively small, it still leaves room for grade-fixed effects rather than test-fixed effects when linking tests. For example, linking PIRLS 2001 grade 4 with SACMEQ 2000 grade 6 might capture a grade difference in PIRLS (which is likely to increase performance) in addition to difficulty. However, to enable greater coverage, we put up with the need to expand beyond single grade level intervals. Moreover, these differences are often small and since the linking functions are applied across the entire set of countries being linked, any bias will be applied equally such that cross-country bias is eliminated for each pair-wise link and regional ranks will be preserved.

We distinguish primary from secondary schooling since enrollment rates drop off quickly between levels in many developing countries. This introduces a potential selection term in secondary school scores, with the highest performing students progressing in the system, biasing scores up due to selection rather than actual learning. We give this methodological parameter increased importance in this update. In the previous version of the database we linked SACMEQ scores in primary school to TIMSS scores in secondary school to increase the number of doubloon countries (Botswana and South Africa). In this version, we reduce the number of doubloon countries (just Botswana) and link at the primary school level only given the scope for selection between levels.

Subjects. We link tests within three specific subjects: mathematics, reading and science. While this is not granular at the test item level, this ensures that there is significant proficiency overlap in linked tests.

Data Availability. The availability of data is sparse. This introduces bias if data availability is correlated with education quality or progress. For example, if countries that perform worse only have available data in later years (since they were later to introduce assessments), this would mean the data that exists is more recent, likely biasing the average up due to testing later rather than stronger performance. Thus, countries that have historically performed poorly might appear to do better. Since we now provide year-by-year test scores, this can be accounted for rather than buried in 5-year intervals.

Relatedly, when aggregating data across subjects, levels and time, there is a trade-off between coverage and disaggregation. For example, while constructing averages across subjects would increase data coverage, it also makes it harder to interpret and introduces potential sources of bias. To this end, we construct disaggregated measures as well as aggregated ones. This enables analyses at each level, considering the trade-offs.

Subsamples. We construct a doubloon index over the full sample. When calculating Harmonized Learning Outcomes by gender we apply the overall doubloon index to raw scores for each sub-sample, rather than constructing sub-sample specific doubloon indices. While performance is likely to vary across sub-samples in a given test, the relationship between test X and Y is unlikely to vary across sub-samples or relative to the full sample.

Doubloon. In constructing the doubloon index, we use as much data as possible. This is an update over our previous methodology, where we choose one specific anchor year in which to construct the index. This enables us to verify if the doubloon index is stable over time, reduces noise, and increases the sample size of the doubloon index.

Steps in the Linking Function. When converting RSAT and EGRA results to an international scale, we have two options: (1) conversion first to regional assessments or (2) directly through an international assessment.

Approach (1) will best preserve underlying regional rankings. However, it will distort rankings at the margin of regional groupings. Approach (2) is direct, simple and uniform, however might distort underlying regional rankings. Each approach will produce more reliable estimates depending on the number of doubloon countries that is possible for each. For example, for EGRA the RSAT transformation for sub-Saharan Africa has three doubloons, while the direct PIRLS transformation also has three doubloons. In this case the doubloon advantage is ambiguous.

We go with approach (2) where possible since it is the simplest and most direct approach and leaves the least room for error, since each step in the linking function introduces uncertainty. Exceptions include PASEC and MLA which only have overlap with SACMEQ. To this end, scores are first made comparable to SACMEQ scores, before they are converted to our final HLO score.

Hierarchy of Tests. We construct a hierarchy of tests to facilitate year-level-subject matching between anchored and reference tests. For math and science, scores are converted to TIMSS-comparable scores at primary and secondary level. For reading, scores are converted to PIRLS-comparable scores at primary level and PISA-comparable scores at secondary level.

This hierarchy of tests can be extended to adjudicate which HLO score to use when multiple scores are available across sources tests. (e.g. a country that took part in EGRA and SACMEQ). The first HLO choice is: PIRLS, PISA or TIMSS depending on the subject and schooling level. If a country took part in PISA and not in TIMSS for math and reading, we next convert PISA scores into subject-specific secondary HLO scores. If a country took part in LLECE, we convert these scores into a subject-specific primary HLO score. Next, we take SACMEQ HLO scores, followed by PASEC HLO scores, then EGRA HLO scores, and finally by MLA HLO scores.

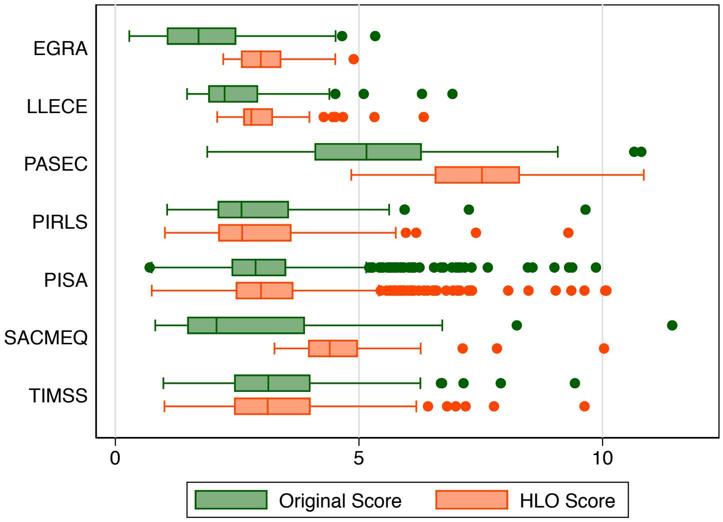
Measures of Uncertainty. We include measures of uncertainty to quantify the degree of confidence around our estimates. We capture two sources of uncertainty: scores on the original test and uncertainty in the calculation of the conversion factor across tests. Simplifying equations (2) and (3), the HLO for a given country on test X is:

$$HLO = \frac{X}{d_{sl}}$$

We calculate the variance of the HLO by bootstrapping. We assume country-level average scores are asymptotically normally distributed . We take 1000 draws from the distribution of subject-level-grade average test scores for each testing regime. We create doubloon indices and HLOs from each bootstrapped sample. We take the 2.5th and 97.5th percentiles of the distribution and use this to construct lower and upper bounds of uncertainty for our HLO scores.

We find small uncertainty intervals overall, with an average of 3.5 points and ranging from 1 to 11 points (see Figure 6.0). This is consistent with original standard errors from each respective testing regime. We find larger uncertainty for HLO scores relative to original scores when testing regimes have fewer doubloon countries, such as SACMEQ and PASEC. By quantifying this uncertainty, we can reliably bound our estimates.

Figure 6.0: Standard Errors for Original and HLO Scores by testing regime



Scale. Our methodology enables us to link countries participating in regional standardized achievement tests (RSATs) on an international scale. On this scale, 625 represents advanced attainment and 300 represents minimum attainment. This interpretation is derived by taking an ISAT benchmark for high performance on the upper end of the distribution and an RSAT benchmark (expressed in HLO units) on the lower end of the distribution. This approach enables us to capture performance across the entire distribution and accounts for floor and ceiling effects that would be introduced by just taking the ISAT or RSAT benchmarks alone on both ends of the spectrum.

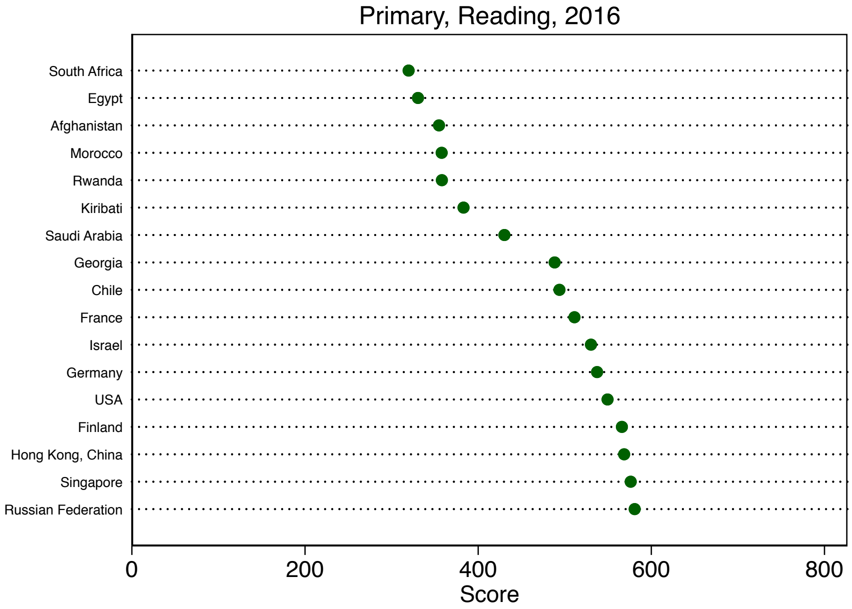
4. Illustrative Analysis

We include a few key figures from our database as motivation for the utility of the database. We analyze disaggregated data by subject and schooling levels to explore the full texture of the data.

Figure 7.0 shows select countries on a global scale in reading in 2016. We see a few notable trends. Russia has taken the lead – a new development. Chile nearly outperforms France and has started to outperform Eastern European countries such as Georgia. Saudi Arabia places near the bottom outperforming only African, Pacific or war-affected countries. Rwanda outperforms the regional behemoth, South Africa. The performance gap between Egypt and Russia is nearly double.

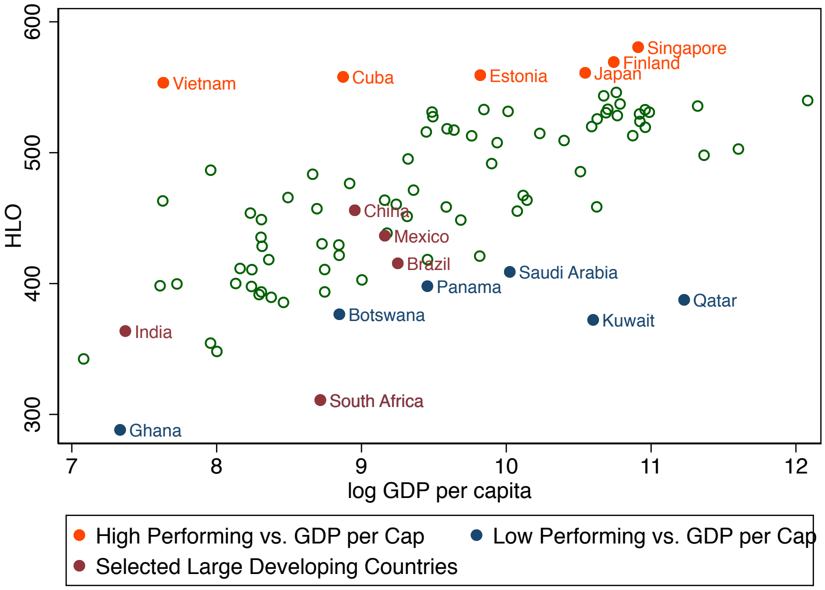
Figure 8.0 shows the relationship between average harmonized learning outcomes in science from 2000-2017 and 2015 GDP per capita.

Figure 7.0: Select Countries, Primary, Reading, 2016



In figure 8.0, we highlight countries in red that outperform relative to their incomes status, such as Singapore, Japan, Finland, Estonia, Cuba, and Vietnam. Countries that underperform on learning outcomes relative to their income status include Qatar, Saudi Arabia, Kuwait, South Africa, Panama, Botswana, and Ghana. We also highlight large developing countries on the same graph. These include India, China, Mexico, Brazil and South Africa. China outperforms its counterparts, while India and South Africa trail behind. These graphs demonstrate a tractable 'distance to the frontier' – large developing economies who might benefit their economies by focusing on human capital development.

Figure 8.0: Average HLO Science Score (2000-2017) vs. 2015 GDP per Capita



5. Robustness Checks

We compare Harmonized Learning Outcomes (HLOs) for primary reading scores with scores generated using IRT linking by the *Linking International Comparative Student Assessment* (LINCS) project. The LINCS project leverages overlap in items for a subset of ISATs focused on reading at primary school from 1970 onwards (Strietholt, 2014; Strietholt and Rosén, 2016).

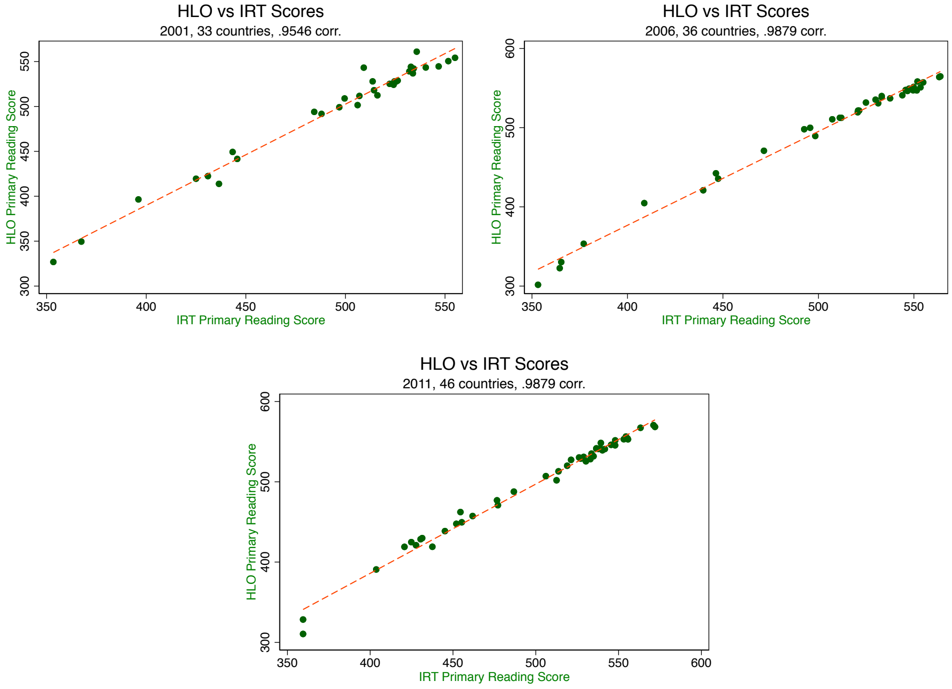
Figures 9.0-9.6 below show our results. The HLO and IRT scores have large overlap with a correlation coefficient consistently above 0.95 for scores as well as ranks. This indicates we are able to produce similar results to an IRT methodology where there is overlap. This increases our confidence in the robustness of our estimates.

We further compare ranks derived from HLO scores versus raw scores and verify the original regional rankings are preserved. We are able to fully preserve underlying regional ranks within a year subject and schooling level. This is true by design due to a fixed exchange rate applied at this level of disaggregation. We do not show these results since they are true mechanically.

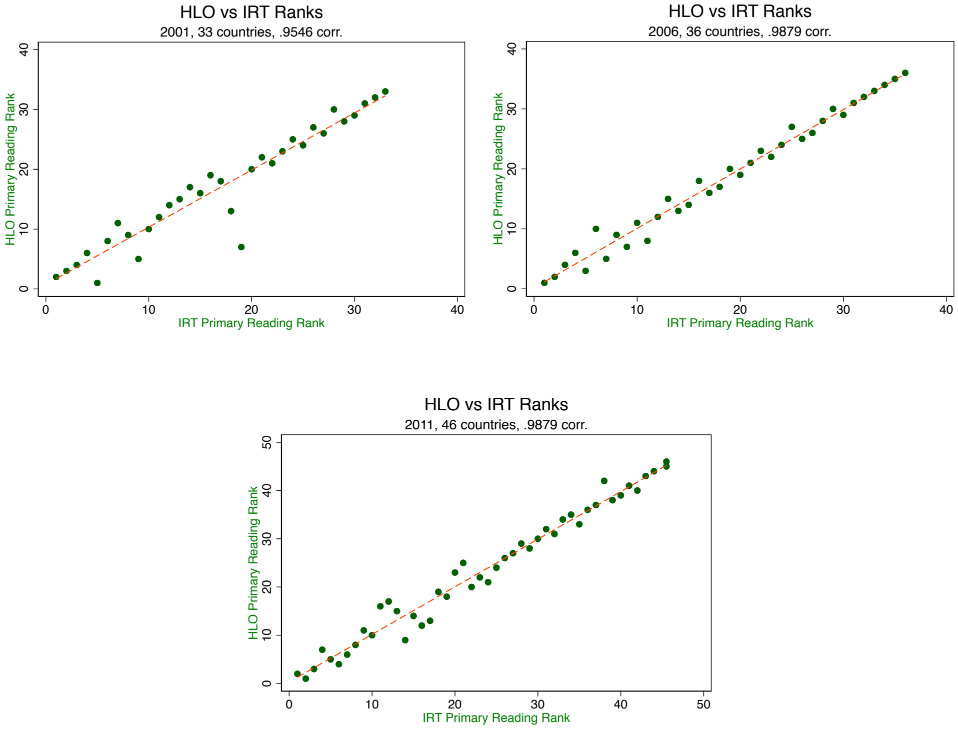
We conduct two additional robustness tests, shown in Annex 5. The first explores HLOs derived from multiple possible source tests and how this affects international and regional rankings. The second is a comparison of test scores by data availability of each test. These robustness tests confirm that even when HLO scores come from different tests they are broadly consistent.

A future robustness check could be the inclusion of psychometric adjustments. Jerrim et al. (2017) highlight a few features of ISATs that are important to consider when conducting aggregate test score comparisons. Jerrim et al. (2017) conclude that results of a paper they re-analyze by Lavy (2015) are robust to inclusion of these elements. This enhances our confidence that such adjustments are not strictly necessary, but we hope to include them as a robustness test in the future. For example, if we want a regional average we can weight countries to construct a reliable 'regional average' that is not distorted by artificially weighting each country equally since the true region consists of larger countries.

Figures 9.0-9.3: Robustness - HLO vs. IRT Primary Reading Scores by Year



Figures 9.4-9.6: Robustness - HLO vs. IRT Primary Reading Ranks by Year



Conclusion

There is policy and academic demand for a measure of education quality that is comparable across countries and over time. The growth of international standardized achievement tests is a huge step in this direction. However, the countries that participate in these tests are often high and middle-income countries. This limits the ability to track, compare, or understand education patterns in developing countries – the countries that often have the most to gain from quality education and human capital formation.

One option is to wait to make comparisons for low-income countries to participate in international assessments. Although this is a worthy aspiration and we hope it occurs, it will take time and render a rich array of retrospective data null, limiting longitudinal data analysis. Alternatively, we can use a rigorous approach – albeit with caveats – to harmonize available learning data across different types of international and regional assessments. This is the approach we take in this paper, creating a Harmonized Learning Outcomes (HLO) database. Until there is a global proficiency assessment for numeracy and literacy, and to enable rich longitudinal panel data analysis, the harmonization of existing learning assessments provides the next best alternative to compare education quality on a global scale. Moreover, as more countries join international and regional assessments, and do so for longer, the accuracy and robustness of the harmonization exercise will improve.

We build a globally comparable database of 164 countries/territories from 2000-2017, approximately two-thirds of which are in developing economies, representing more than 98 percent of the global population. While our methodology has limitations, our robustness tests indicate that this dataset produces similar results to each underlying assessment used, as well as Item Response Theory (IRT)-linking methodologies where there is overlap. We include double the number of countries of any individual assessment or IRT-linking methodology while demonstrating consistent estimates where there is overlap.

We contribute to the literature in several ways. This is the most current and expansive cross-country dataset on education quality, including the most developing countries. We include, for the first time, the Early Grade Reading Assessment (EGRA) in addition to TIMSS, PISA, PIRLS, SACMEQ, LLECE, PASEC and MLA. This enables us to extend our database substantially, including 48 countries with recent data in the last ten years, all of which are developing economies. This methodological update paves the way to include a series of assessments that are increasingly common in developing countries that are often excluded from large international assessments. We introduce a series of methodological updates, including measures of uncertainty, fixed conversion factors for greater comparability over time, and year-by-year data. All data are derived directly from test score data. In contrast to black-box and complex imputation, this enables methodological transparency. Moreover, it produces a clear policy lever: when countries perform better on RSATs and ISATs they can be certain the HLO will improve.

Our goal in this paper is not to provide a perfect measure of education quality. Rather, we provide a practical yet rigorous and globally comparable set of estimates with large and inclusive country coverage over time. We hope this dataset can be used to reveal important descriptive trends in human capital formation across developed and developing countries. We also hope to enable analysis of factors correlated with and that have plausible causal links to the formation of human capital and economic growth. Finally, we hope this dataset can be useful for monitoring and evaluation of important policy goals.

Future iterations of this dataset will continue to expand coverage across countries and time as countries join existing assessments and by including additional assessments such as early grade reading and mathematics assessments. Moreover, we aim to build a dataset that enables over-time isolation of value-added learning by including variables which can account for various selection effects. This includes linking quality of education data to quantity of education data, as well as including measures of enrollment and retention across schooling levels. We also aim to enable further identification of the link between education quality and economic growth by including variables such as comparable estimates on the returns to education. We hope this dataset, and future iterations, will enable a deeper understanding of mechanisms driving human capital formation, the link to development, and useful policy applications.

References

Abadzi, Helen. (2011). Reading fluency measurements in EFA FTI partner countries: Outcomes and improvement prospects.

Abdul-Husein, Noam Angrist, Syedah Aroob Iqbal, Aart Kraay and Harry Patrinos. (2018). Note on Estimating Nationally-Representative Test Scores for China. In process, World Bank.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. American Educational Research Association.

Altinok, Nadir, and Hatidje Murseli. (2007). International database on human capital quality. *Economics Letters* 96.2 237-244.

Altinok, Nadir, Claude Diebolt, and Jean-Luc Demeulemeester. (2014). A new international database on education quality: 1965–2010. *Applied Economics* 46.11: 1212-1247.

Altinok, Nadir, Noam Angrist, and Harry Patrinos. (2018). Global Dataset on Education Quality 1965-2015. *World Bank Policy Research Working Paper No. 8314*.

Altinok, Nadir. (2017). Mind the Gap: Proposal for a Standardised Measure for SDG 4-Education 2030 Agenda. *UIS Information Paper* 46.

Angrist, Noam, Harry Anthony Patrinos, and Martin Schlotter. (2013). An expansion of a global data set on educational quality: a focus on achievement in developing countries. The World Bank.

Barro, Robert J., and Jong Wha Lee. (1996). International measures of schooling years and schooling quality. *The American Economic Review* 86.2: 218-223.

Barro, Robert J., and Jong-Wha Lee. (2015). Education matters. Global schooling gains from the 19th to the 21st century. *Oxford University Press.*

Barro, Robert J., and Jong-Wha Lee. (2001). International data on educational attainment: updates and implications. *Oxford Economic Papers* 53.3: 541-563.

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

Dubeck, Margaret M., and Amber Gove. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development* 40: 315-322.

Dundar, Halil, et al. (2017). Sri Lanka Education Sector Assessment: Achievements, Challenges, and Policy Options. The World Bank.

Gove, A. (2009). Early Grade Reading Assessment Toolkit. RTI International, USAID and the World Bank.

Hanushek, Eric A., and Dennis D. Kimko. (2000). Schooling, labor-force quality, and the growth of nations. *American economic review* 90.5:1184-1208.

Hanushek, Eric A., and Ludger Woessmann. (2012) Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of economic growth* 17.4: 267-321.

Hanushek, Eric A., and Ludger Woessmann. (2008). The role of cognitive skills in economic development. *Journal of economic literature* 46.3: 607-68.

Hanushek, Eric A., Stephen J. Machin, and Ludger Woessmann. (2016). *Handbook of the Economics of Education*. Elsevier.

Hintze, John M., Amanda L. Ryan, and Gary Stoner. (2003). Concurrent validity and diagnostic accuracy of the dynamic indicators of basic early literacy skills and the comprehensive test of phonological processing. *School Psychology Review* 32.4: 541-556.

Holland, Paul W., and Machteld Hoskens. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika* 68.1: 123-149.

Holland, Paul W., and Neil J. Dorans. (2006). Linking and equating. *Educational measurement* 4: 187-220.

Jerrim, John, et al. (2017) What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review* 61: 51-58.

Kim, Young-Suk, et al. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology* 102.3: 652.

Kolen, Michael J., and Robert L. Brennan (2014). Nonequivalent groups: Linear methods. *Test equating, scaling, and linking*. Springer, New York, NY. 103-142.

Lange, Glenn-Marie, Quentin Wodon, and Kevin Carey. (2018). The changing wealth of nations 2018: Building a sustainable future. The World Bank.

Lavy, Victor. (2015). Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries. *The Economic Journal* 125.588: F397-F424.

Linn, Robert L. (1993). Educational assessment: Expanded expectations and challenges. *Educational evaluation and policy analysis* 15.1: 1-16.

Mislevy, Robert J., et al. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* 29.2: 133-161.

Piper, Benjamin. (2010). Ethiopia early grade reading assessment data analytic report: Language and early learning. *Prepared for USAID/Ethiopia under the Education Data for Decision Making (EdData II) project, Task Order Nos. EHC-E-07-04-00004-00 and AID-663-BC-10-00001 (RTI Tasks 7 and 9). Research Triangle Park. NC: RTI International. Retrieved January* 15: 2015.

Pritchett, Lant. (2001). Where has all the education gone? *The world bank economic review* 15.3: 367-391.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53(4):495–502.

Sandefur, Justin. (2018). Internationally comparable mathematics scores for fourteen African countries. *Economics of Education Review* 62. 267-286.

Steinmann, I., Rolf Striethold, and Wilfried Bos. (2014). Linking International Comparative Student Assessment. LINCS Technical Report.

Strietholt, Rolf, and Monica Rosén. (2016). Linking large-scale reading assessments: Measuring international trends over 40 years. *Measurement: Interdisciplinary Research and Perspectives* 14.1: 1-26.

Turner, Ross, et al. (2018). Development of Reporting Scales for Reading and Mathematics: A report describing the process for building the UIS Reporting Scales.

Vagh, Shaher Banu. (2010). Validating the ASER testing tools: Comparisons with reading fluency measures and the Read India measures. *Unpublished report. Available at http://img. asercentre. org/docs/Aser% 20survey/Tools% 20validating_the_aser_testing_tools_oct_2012_3. pdf.*

Wagner, R., et al. (1999). CTOPP: Comprehensive Test of Phonological Processing–Second Edition.

World Bank. (2018). World Development Report 2018: Learning to Realize Education's Promise. Washington, DC.

Annex Table 1: Review of Student Achievement Tests

Subjects: M=math; S=science; R=reading.

No	*Year*	*Organization*	*Abbr.*	*Subject*	*Countries/ Areas*	*Grade/Age*	*Included*
1	Every four years since 2003 (latest round is 2015)	IEA	TIMSS	M,S	45, 38, 26, 48, 66, 65, 65	3-4,7-8, FS	■
2	2000	UNESCO	MLA	M,S,R	72	6,8	■
3	2006, 2013	UNESCO	LLECE	M,S,R	13, 16 (only 6 for science)	3,6	■
4	2000, 2003, 2007, 2013	UNESCO	SACMEQ	M,R	7, 15, 16	6	■
5	2006, 2014	CONFEMEN	PASEC	M,R	22 (before 2014), 10	Until 2014: 2,5 After 2014: 3, 6	■
6	Every five years since 2001 (latest round is 2016)	IEA	PIRLS	R	35, 41, 55	4	■
7	Every three years since 2000 (latest round is 2015)	OECD	PISA	M,S,R	43, 41, 57, 74, 65, 71	Age 15	■
8	2007-2017	RTI/USAID	EGRA	R	65	Grades 2-4	■

Annex Table 2: Test Linking Architecture and Corresponding Doubloon Countries

PISA-TIMSS Linking (Math and Science)

Reference Test X	Anchor Test Y	Subject	Doubloon Countries
PISA 2000	TIMSS 1999	Math / Science	Australia, Bulgaria, Canada, Chile, Czech Republic, Finland, Hong Kong – China, Hungary, Indonesia, Israel, Italy, Japan, Korea - Republic of, Latvia, Macedonia F.Y.R., Netherlands, New Zealand, Romania, Russian Federation, Thailand, USA.
PISA 2003	TIMSS 2003	Math / Science	Australia, Hong Kong – China, Hungary, Indonesia, Italy, Japan, Korea Republic of, Latvia, Netherlands, New Zealand, Norway, Russian Federation, Slovakia, Sweden, Tunisia, USA
PISA 2006	TIMSS 2007	Math / Science	Australia, Bulgaria, Chinese Taipei, Colombia, Czech Republic, Hong Kong - China, Hungary, Indonesia, Israel, Italy, Japan, Jordan, Korea - Republic of, Lithuania, Norway, Qatar, Romania, Russian Federation, Serbia, Slovenia, Sweden, Thailand, Tunisia, Turkey, USA
PISA 2009/2012 (average)	TIMSS 2011	Math / Science	Australia, Chile, Chinese Taipei, Finland, Georgia, Hong Kong – China, Hungary, Indonesia, Israel, Italy, Japan, Jordan, Kazakhstan, Korea – Republic of, Lithuania, Malaysia, New Zealand, Norway, Qatar, Romania, Russian Federation, Singapore, Slovenia, Sweden, Thailand, Tunisia, Turkey, USA, United Arab Emirates
PISA 2015	TIMSS 2015	Math / Science	Argentina – Buenos Aires, Australia, Canada, Chile, Chinese Taipei, Georgia, Hong Kong – China, Hungary, Ireland, Israel, Italy, Japan, Jordan, Kazakhstan, Korea – Republic of , Lebanon, Lithuania, Malaysia, Malta, New Zealand, Norway, Qatar, Russian Federation, Singapore, Slovenia, Sweden, Thailand, Turkey, USA, United Arab Emirates

SACMEQ Linking

Reference Test X	Anchor Test Y	Subject	Doubloon Countries
SACMEQ 2007/2013	PIRLS 2011	Reading	Botswana
SACMEQ 2007/2013	TIMSS 2011 - grade 4	Math	Botswana

LLECE Linking

Reference Test X	Anchor Test Y	Subject	Doubloon Countries
LLECE 2013	PIRLS 2011/2016 (avg)	Reading	Colombia, Chile, Honduras
LLECE 2006	TIMSS 2007	Math/Science	Colombia, El Salvador
LLECE 2013	TIMSS 2011/2015 (avg)	Math/Science	Chile, Honduras

PASEC Linking

Reference Test X	Anchor Test Y	Subject	Doubloon Countries
PASEC 2006	SACMEQ 2007	Reading/Math	Mauritius
PASEC 2014	PASEC 2006	Reading/Math	Togo

EGRA Linking

Reference Test X	Anchor Test Y	Subject	Doubloon Countries
EGRA 2007-2017	PIRLS 2011/2016 (avg)	Reading	Egypt, Honduras, Indonesia

MLA Linking

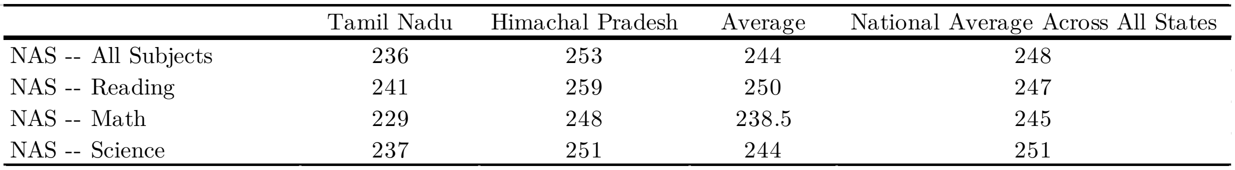
Reference Test X	Anchor Test Y	Subject	Doubloon Countries
MLA 2000	SACMEQ I (2000)	Math/Reading	Botswana, Malawi, Mauritius, South Africa, Uganda, Zambia

Annex 3: Validating Nationally Representative India HLO score[3]

India participated in PISA 2009 represented by two states, Himachal Pradesh and Tamil Nadu. Given we aim to include nationally representative data in our HLO database, we confirm whether scores from these states reflect the national average.

We examine the NAS Report for Grade 8 (Cycle 3) – one the largest surveys conducted in India and the world. It is officially described as a “transparent and credible exercise done under third party verification” (Prakash Javadekar). It was conducted for Classes 3, 5 and 8 in government and government aided schools. The survey tools used multiple test booklets with 45 questions in Classes III and V related to language, mathematics and 60 questions in Class VIII in Mathematics, Language, Sciences and Social Sciences. Along with the test items, questionnaires pertaining to students, teachers and schools were also used. The scale is from 0 to 500. The average score for the entire population is initially set at 250. The standard deviation of the scale is initially set at 50. This means that most students (about 70%) will have scores in the range 200 to 300.

On the Grade 8 NAS (Cycle 3), we find that Tamil Nadu and Himachal Pradesh’s average is nearly a one-to-one match to the national average, with a difference of 3-7 points across subjects and 3.5 points on average. This indicates that the states participating in PISA are approximately representative of the national average based on the distribution of NAS scores across states.[4]

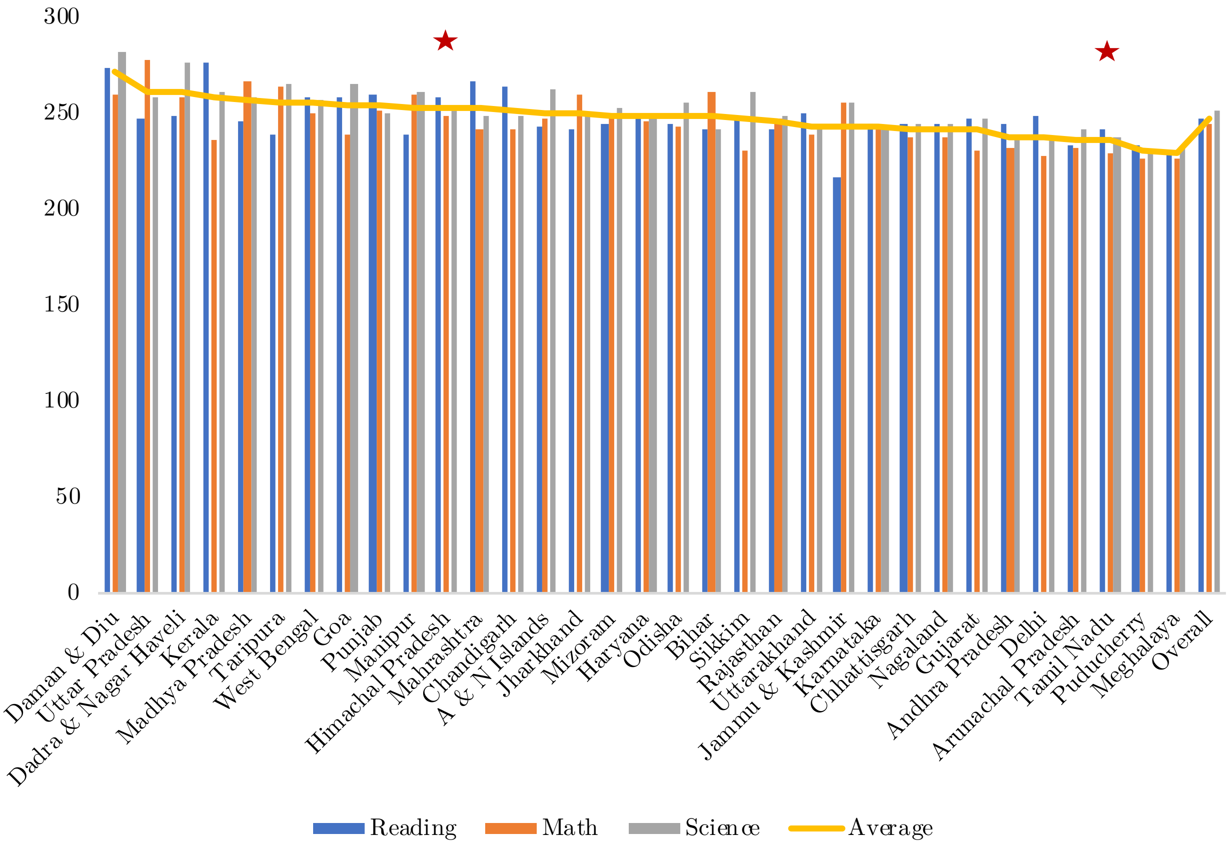


	Tamil Nadu	Himachal Pradesh	Average	National Average Across All States
NAS -- All Subjects	236	253	244	248
NAS -- Reading	241	259	250	247
NAS -- Math	229	248	238.5	245
NAS -- Science	237	251	244	251

The two PISA states are neither top nor bottom performers in NAS. Himachal Pradesh ranks 10th overall, while Tamil Nadu is 31st out of 33. Below is a graph of state by state performance across subjects on NAS.

[3] This annex was written with Syedah Aroob Iqbal.

[4] This calculation does not weight for population.



We compare PISA 2009 data to NAS data and conduct a ratio linking exercise to extrapolate an HLO score. The conversion factor between NAS and PISA is 1.376. Applying this, gives us a score of 341 based on NAS across all states. This is remarkably close to the PISA HLO which is 336, shown below by state and on average.

	India PISA 2009 original scores			
State	Reading	Math	Science	Average
Himachal Pradesh	317	338	325	327
Tamil Nadu	337	351	348	345
Average	327	345	337	336

This exercise gives us confidence that the HLO scores derived from PISA are approximately nationally representative for India.

Annex 4: Nationally Representative China Data[5]

Background

With nearly 1.4 billion people, the People's Republic of China is the world's most populous country and has the world's largest education system. While special administrative regions of China, namely, Hong Kong and Macao, have been regular participants in international standardized assessments (ISATs) like TIMSS/PIRLS and PISA, China, as a country, has yet to participate.

However, a few provinces have participated in PISA over time, which has provided some insight into the quality of education in China. In 2009 and 2012, the province of Shanghai participated in PISA and was the top-performing economy in both years. Shanghai is predominantly urban with 88 percent of the residents in urban areas. The city's income per capita is also more than twice as high as the national average.[6] In 2015, four provinces, Beijing, Shanghai, Jiangsu and Guangdong (B-S-J-G) participated in PISA. With the inclusion of relatively less urban and less affluent provinces of Jiangsu and Guangdong as compared to Shanghai, PISA scores for China dropped from first to tenth place. These four provinces of China cover 17 percent of China's population and give us an insight into education quality in a less affluent region than Shanghai alone: per capita income in B-S-J-G is 1.49 times the national average. However, this income gap is still large and so PISA scores for B-S-J-G are unlikely to be representative of China as a whole.

Apart from the participation of four provinces in PISA, some independent efforts have been conducted to assess education quality in other provinces in China. One such effort is led by the Rural Education Action Program at Stanford. In 2015, researchers at REAP led a reading assessment to assess education quality in Shaanxi province and rural areas of Guizhou and Jiangxi provinces.[7] The reading tests were constructed by trained psychometricians and used test items from the Progress in International Reading Literacy Study (PIRLS), thereby allowing international comparison. Results show that Chinese provinces of Jiangxi and Guizhou stood last in comparison to other countries participating in PIRLS 2011. The provinces represent 5.8 percent of the population in China and give insight into the education quality in rural, less affluent regions of China: average per capita income in these three regions is just 60 percent of the national average.

Extrapolating to China National Average Test Scores

[5] This annex was written with Husein Abdul-Hamid, Syedah Aroob Iqbal and Aart Kraay.

[6] Throughout this annex we measure per capita income as household per capita disposable income, as reported by NBS based on China's household survey. Available from the NBS website at http://data.stats.gov.cn/english/easyquery.htm?cn=E0103.

[7] Gao, Qiufeng, Yaojiang Shi, Hongmei Yi, Cody Abbey, and Scott Rozelle (2017). "Reading Achievement in China's Rural Primary Schools: A Study of Three Provinces". Stanford University, Freeman Spogli Institute Working Paper, available at https://fsi.stanford.edu/publication/reading-achievement-chinas-rural-primary-schools-study-three-province.

We combine information from PISA scores as provided by OECD and PIRLS scores as provided by REAP to estimate the average education quality in China. For both programs, the fundamental problem is that the test scores are obtained in provinces that are unlikely to be nationally-representative of educational quality given the income gaps noted above. Even taking a population-weighted average of PISA and PIRLS scores for the seven provinces is unlikely to result in nationally-representative scores. This is because the average per capita income of these areas is still much higher than the national average. Average household income in B-J-S-G is 1.49 times the national average, while average household income in all seven provinces covered by PISA and PIRLS is 1.26 times the national average.

Therefore, in addition to these estimates for seven provinces, we approximate national-level test scores by extrapolating the observed PISA and PIRLS scores based on per capita income. Specifically, we extrapolate PISA scores for Shanghai (2012) and B-S-J-G (2015) to the national average using log per capita disposable income, separately by subject. Below we include tables which showcase this calculation.

Extrapolating PISA Scores by Subject:

	Log of HH Income	Actual and Imputed Test Scores				
		Math	Reading	Science	Overall	HLO
Shanghai (PISA 2012)	10.90	613	570	580	588	608
B-S-J-G (PISA 2015)	10.47	531	494	518	514	532
China National (Extrapolated	10.08	455	424	461	446	462
Doubloon Index Conversion Factor		1.05	1	1.05		

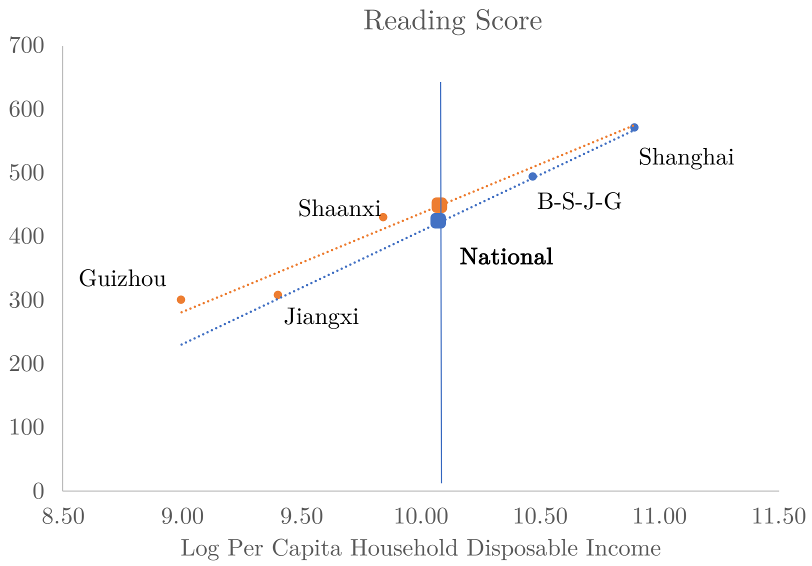
Extrapolating PIRLS Scores:

	Log HH Income	Actual and Imputed Test Scores	
		Reading	HLO
Shaanxi	9.85	430	
Jiangxi (Rural Only)	9.40	308	
Guizhou (Rural Only)	9.00	300	
China National (Extrapolated	10.08	449	449
Doubloon Index Conversion Factor		1	

This gives an extrapolated national PISA score of 446, as compared with PISA scores of 588 and 514 for Shanghai and B-S-J-G, respectively. This corresponds to an extrapolated national-level Harmonized Learning Outcome (HLO) of 462. We do the same using the PIRLS scores for reading only. This gives an extrapolated national-level reading score of 449. Since the units of PIRLS and HLO are the same, this is also the HLO.

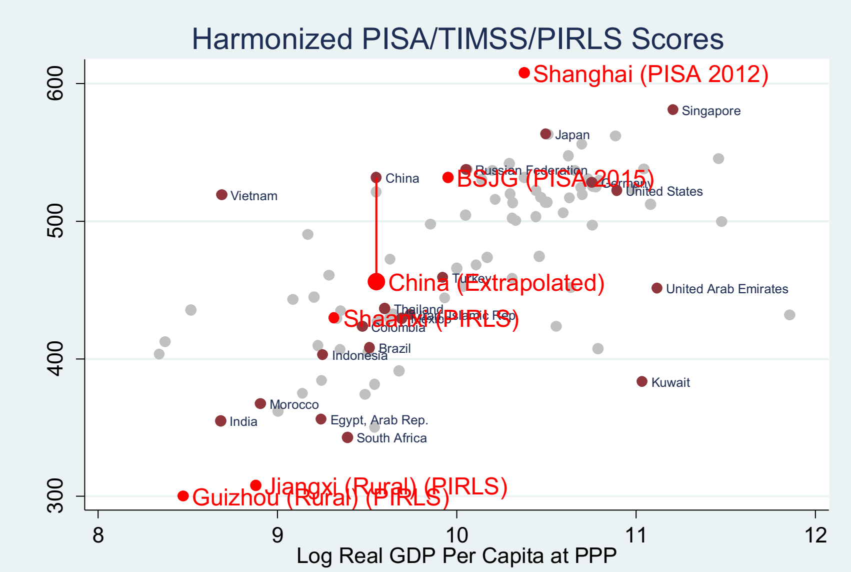
The PISA and PIRLS-based extrapolations give remarkably similar results. This is most apparent from comparing the PISA reading score extrapolation with the PIRLS extrapolation. The graph below plots reading scores from the two programs against log per capita household disposable income in 2016. The small blue (orange) dots represent the data points for PISA (PIRLS). The dashed lines represent the corresponding extrapolations. Finally, the two large data points show the extrapolated values for China national-level scores, which are 462 and 449 for PISA and PIRLS respectively.

Comparing PISA and PIRLS Extrapolated Reading Scores:



The figure below places China's extrapolated test score in international perspective. The gray dots in the graph report the most recently-available PISA, TIMSS and PIRLS scores for countries participating in these programs. The test scores are on the vertical axis and are calculated as the average of whichever of TIMSS 2015, PISA 2015, and PIRLS 2016 are available for the country, after harmonizing to HLO units. The horizontal axis is log real GDP per capita. Over top of this we superimpose (a) the extrapolated China national test score (large red dot in center of graph; the vertical line connects it to what would be a China score based on PISA 2015 scores for B-S-J-G alone) and (b) the actual PISA and PIRLS scores for Chinese provinces described above (for these provinces we assume that the ratio of GDP per capita to the national average is the same as the ratio of household income per capita to the national average).

China's Extrapolated National HLO In International Perspective:



Note: Data are preliminary and subject to revision (as of August 31, 2018)

Summary and Caveats

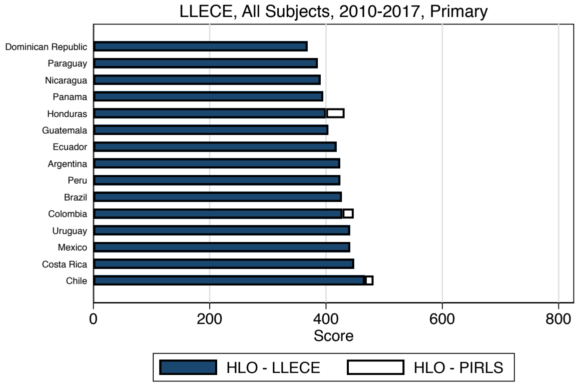
In summary, extrapolating PISA scores based on income differences suggests a national-level HLO of 462, while doing the same using PIRLS scores suggests a national-level HLO of 449. We take the average of these two figures, 456, as the best estimate of the national-level HLO for China and use the values of 449 and 462 as lower and upper bounds.

Naturally these extrapolations are tentative, as they are based on only two data points for PISA, and three data points for PIRLS. However, they are necessitated by the absence of published nationally-representative test score data for China. It is worth noting that a high-quality national level assessment exists with provincial results. That assessment – known as NAEQ – is linked with PISA 2012. However, the results have yet to be published. The publication of such data could help us validate the extrapolated scores. Their publication could also make the extrapolated scores obsolete.

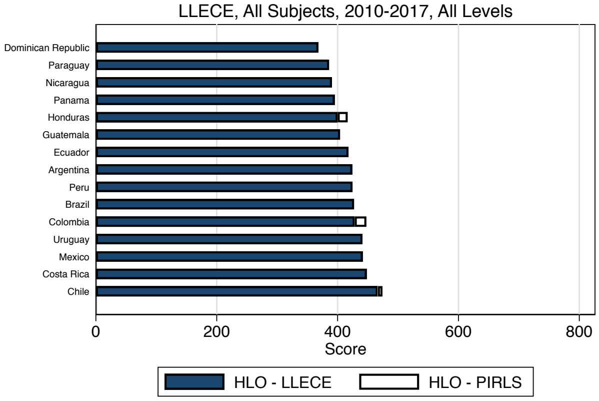
Annex 5: Robustness Test - Sensitivity to Multiple Possible HLOs

We conduct a series of sensitivity tests comparing multiple possible HLOs. For example, when we convert LLECE scores to an ISAT-comparable score, we can either take the PIRLS score or the LLECE-transformed score for primary reading scores. Our hierarchy of tests dictates that we take the PIRLS score. While this will preserve the doubloon countries relationship to other countries within their ISAT ranking, it could distort underlying regional rankings. Below we show how this choice affects regional rankings.

We see (top panel) that if you use the PIRLS HLO score, Chile > Colombia > Honduras. If you use the LLECE HLO scores change only slightly, however the rank order among the three countries with HLO scores from both tests is preserved. Within-region ranks shift by one rank for Colombia and 2-3 ranks for Honduras, with Honduras underperforming Guatemala and Ecuador using the LLECE HLO. Thus, we see that where there is overlap ranks are preserved but that underlying regional rankings shift by taking the PIRLS HLO for countries that have them.



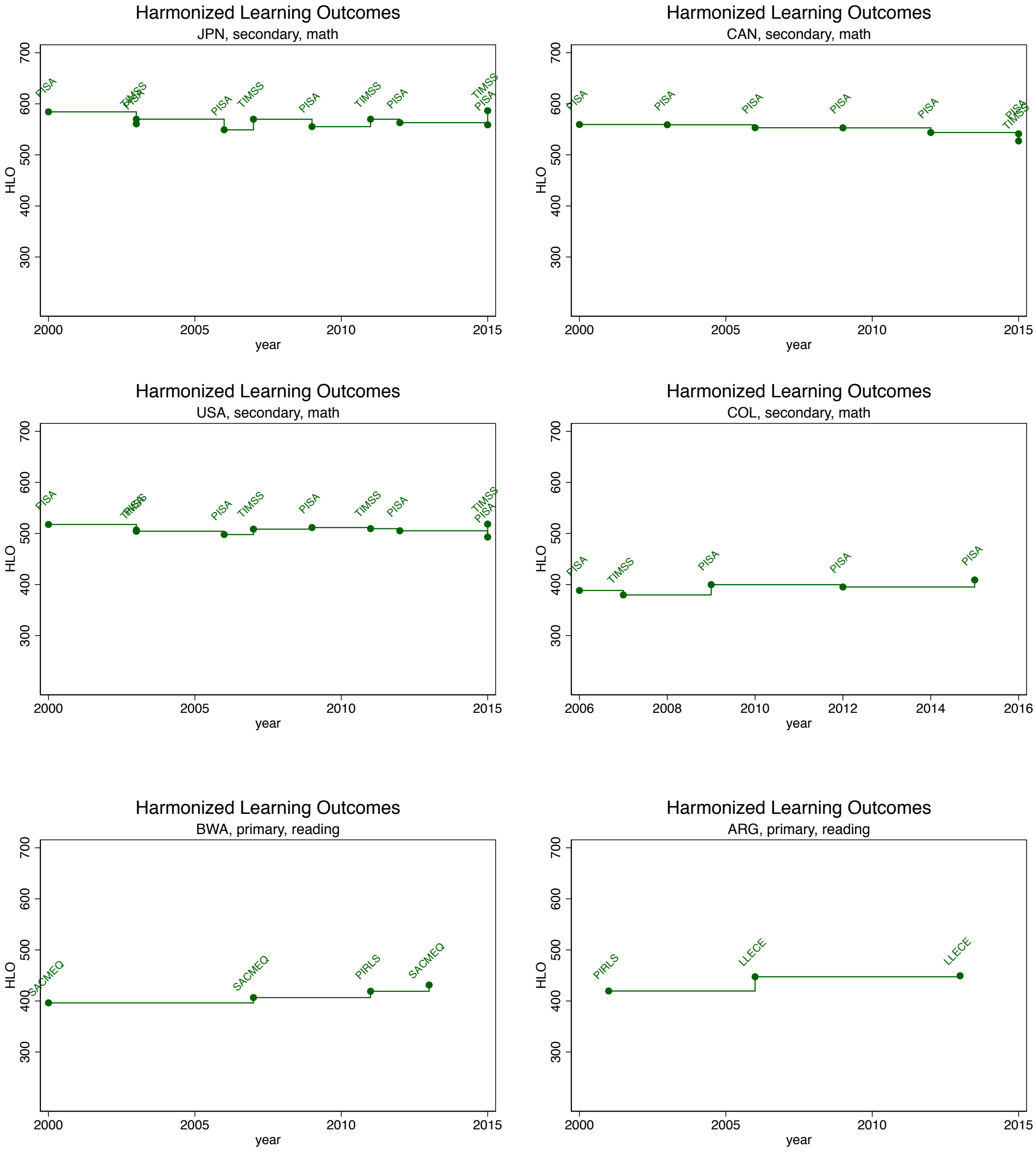
When we average across subjects and levels (bottom panel), we see that scores and ranks are preserved more robustly. Honduras and Colombia now only shift one rank within region. This exercise demonstrates broad robustness to the source test for the HLO as well as the trade-offs in the final score choice and sensitivities to our designated hierarchy of tests.



Annex 6: Robustness Test - Sensitivity to Test Availability

We next demonstrate the sensitivity of the hierarchy of test choices to the underlying source test with a few examples. We show the sensitivity for a group of countries for secondary math scores and primary reading scores. These figures demonstrate that HLOs stay on a relatively stable path regardless of the source test from which the HLO was derived.

HLO scores by Source Test, 2010-2015:


Harmonized Learning Outcomes
JPN, secondary, math
PISA
TIMSS
PISA
PISA
TIMSS
PISA
TIMSS
PISA
TIMSS
PISA
Harmonized Learning Outcomes
CAN, secondary, math
PISA
PISA
PISA
PISA
PISA
PISA
TIMSS
Harmonized Learning Outcomes
USA, secondary, math
PISA
TIMSS
PISA
PISA
TIMSS
PISA
TIMSS
PISA
TIMSS
PISA
Harmonized Learning Outcomes
COL, secondary, math
PISA
TIMSS
PISA
PISA
PISA
Harmonized Learning Outcomes
BWA, primary, reading
SACMEQ
SACMEQ
PIRLS
SACMEQ
Harmonized Learning Outcomes
ARG, primary, reading
PIRLS
LLECE
LLECE
HLO
year