WPS8314

Policy Research Working Paper 8314

# Global Data Set on Education Quality (1965–2015)

*Nadir Altinok*
*Noam Angrist*
*Harry Anthony Patrinos*

WORLD BANK GROUP
Education Global Practice Group
January 2018

Policy Research Working Paper 8314


## Abstract

This paper presents the largest globally comparable panel database of education quality. The database includes 163 countries and regions over 1965–2015. The globally comparable achievement outcomes were constructed by linking standardized, psychometrically-robust international and regional achievement tests. The paper contributes to the literature in the following ways: (1) it is the largest and most current globally comparable data set, covering more than 90 percent of the global population; (2) the data set includes 100 developing areas and the most developing countries included in such a data set to date—the countries that have the most to gain from the potential benefits of a high-quality education; (3) the data set contains credible measures of globally comparable achievement distributions as well as mean scores; (4) the data set uses multiple methods to link assessments, including mean and percentile linking methods, thus enhancing the robustness of the data set; (5) the data set includes the standard errors for the estimates, enabling explicit quantification of the degree of reliability of each estimate; and (6) the data set can be disaggregated across gender, socioeconomic status, rural/urban, language, and immigration status, thus enabling greater precision and equity analysis. A first analysis of the data set reveals a few important trends: learning outcomes in developing countries are often clustered at the bottom of the global scale; although variation in performance is high in developing countries, the top performers still often perform worse than the bottom performers in developed countries; gender gaps are relatively small, with high variation in the direction of the gap; and distributions reveal meaningfully different trends than mean scores, with less than 50 percent of students reaching the global minimum threshold of proficiency in developing countries relative to 86 percent in developed countries. The paper also finds a positive and significant association between educational achievement and economic growth. The data set can be used to benchmark global progress on education quality, as well as to uncover potential drivers of education quality, growth, and development.


This paper is a product of the Education Global Practice Group. 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://econ.worldbank.org. The authors may be contacted at hpatrinos@worldbank.org.

# Global Data Set on Education Quality (1965-2015)

Nadir Altinok*, Noam Angrist, Harry Anthony Patrinos (1)

**Key words:** Quality, Human Capital, Education, International, Achievement, Database, PISA, TIMSS, SACMEQ, PASEC, LLECE.

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

(1) * Corresponding author, Nadir Altinok, BETA, CNRS & University of Lorraine (France), Address: BETA, UFR Droit, Sciences Economiques et Gestion 13 place Carnot C.O. 70026 - 54035 Nancy cedex, France. Tel: +33 372 748 452. Email nadir.altinok@univ-lorraine.fr

Noam Angrist, Oxford University, Email noam.angrist@bsg.ox.ac.uk

Harry Patrinos, The World Bank, Email Hpatrinos@worldbank.org

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.

## Introduction

A country's education level is critical for its economics success. For many years, the economics literature focused on the positive effects of education quantity on growth (Barro, 1991; Mankiw, 1992). However, a growing body of evidence suggests it is not only the *quantity* of schooling, measured by average years of schooling or enrollment rates, but also the *quality* of schooling, proxied by student achievement tests, that contributes to growth. It is not about being in school but what is learned in school that matters. Over 15 years of literature now supports this conclusion (Hanushek and Kimko, 2000; Pritchett, 2001; Hanushek and Woessmann, 2008; Hanushek and Woessmann, 2012). The evidence shows that in cross-country regressions when student achievement conditional on years of schooling – rather than years of schooling alone - is correlated with growth, the association and explanatory power of growth models is significantly higher. The most recent World Development Report (World Bank, 2017) highlights this finding. Moreover, Hanushek and Woessmann (2012) use differences-in-differences and instrumental variables methods and find a plausibly causal link between cognitive skills and growth.

This insight comes at a time when the availability and coverage of International Student Achievement Tests (ISATs) – which are carefully constructed, psychometrically-tested, standardized assessments – is growing. ISATs first started in the 1960s and are carried out by institutions such as the OECD and the International Association for the Evaluation of Educational Achievement (IEA). One of the largest ISATs, PISA, covered 71 countries in 2015, and another large ISAT, TIMSS, covered 65 countries in 2015. The growth of these assessments enables credible global comparison of education quality levels and changes over time.

While critically useful, these international achievement tests have a series of limitations. First, while PISA and TIMSS tests are highly correlated (Rindermann and Stephen, 2009), they have meaningful differences in both their rigor and scaling. Thus, when comparing them, it is important to adjust for these differences. Second, since these assessments only started being implemented consistently and in a standardized fashion in the 1990s and 2000s, they are limited in their ability to conduct longitudinal and panel analysis. Third, these assessments often include mostly OECD countries, omitting developing countries which have the most to gain from a quality education. For example, the first PISA in 2000 included 28 OECD countries and four non-OECD countries. While PISA has grown substantially, and in 2015 included 71 countries, none of these countries where from Sub-Saharan Africa. Thus, implications of studies

analyzing PISA and TIMSS results are limited in their inclusion of and application to developing countries. Despite these drawbacks, ISATs provide a strong foundation to obtain globally comparable estimates of education quality.

We build on a literature that aims to produce comparable estimates of cognitive skills across countries and over time, leveraging the emergence and growth of ISATs, and proposing methodological innovations to deal with some of the shortcomings listed above. Our methodology builds on seminal work done by Barro and Lee (1996) and Barro (2001) and provides a global update of previous papers (Altinok and Murseli, 2007, Angrist, Patrinos and Schlotter, 2013, Altinok, Diebolt, de Meulemeester, 2014; Altinok, 2017). We also build on methodologies used by Hanushek and Kimko (2000) as well as extensions by Barro and Lee (2015), Hanushek and Woessmann (2012) and Hanushek and Woessmann (2016).

In a pioneering paper, Barro and Lee (2001) used a simple regression technique to obtain different constants between each test, thus allowing for test differences. Hanushek and Kimko (2000) then created more credible over-time comparisons by adjusting ISATs between 1964-1995 using the *National Assessment of Educational Progress* (NAEP) in the United States as an anchor, since the United States participated in both the NAEP and each ISAT. To this end, they use the United States' performance in NAEP over time to adjust for varying difficulty and scaling across ISATs and construct comparable over-time achievement data. Recent work by Hanushek and Woessmann (2016) aims to address issues of equating *variation* across ISATs in addition to equating *levels*. To do this, the authors express performance in terms of standard deviations and project the standard deviation of a relatively homogenous and stable group of OECD countries – termed the "OECD Standardization Group" (OSG) of countries – and then transform these standard deviations into scores using the standardized PISA scale.[12] However, as the authors acknowledge, this does not apply for countries far off the OSG scale since ISATs may be too difficult and irrelevant for this sub-set of countries, distorting the variance equating exercise. This bias is particularly important for analyses focused on developing countries.

Altinok and Murseli (2007) provide the first attempt to include a significant number of developing countries in internationally comparable estimates. Many developing countries do not participate in international tests such as PISA and TIMSS. However, they do participate in

[1] The criteria chosen for the "OECD Standardization Group (OSG)" includes: the countries have to be member states of the relatively homogenous and economically advanced group of OECD countries over all ISATs observations. Second, the countries should have had a substantial enrollment in secondary education in 1964.
[2]The OSG countries are: Austria, Belgium, Canada, Denmark, France, Germany, Iceland, Japan, Norway, Sweden, Switzerland, the United Kingdom, and the United States.

regional assessments, which if made comparable to international assessments, would provide further insight into achievement in developing regions. For example, Latin American countries participate in the UNESCO Laboratorio Latinoamericano para la Evaluación de la Calidad de la Educación (LLECE) and many African countries participate in the South and Eastern African Consortium for Monitoring Educational Quality (SACMEQ) or the Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC). Altinok and Murseli (2007) use a similar anchoring approach as Hanushek and Kimko (2000) – creating an index to adjust for scaling and difficulty. Instead of doing this over time, they do this across assessments, linking Regional Standardized Achievement Tests (RSATs) and ISATs. The first database of this kind produced by Altinok and Murseli (2007) included 105 countries. Extensions by Angrist, Patrinos and Schlotter (2013) included 128 countries for any test score from 1965-2010 and Altinok et al. (2014) included up to 103 countries in primary education and 111 countries in secondary education.

In this paper, we build on datasets constructed by Altinok et al. (2014), Angrist et al. (2013) and Altinok and Murseli (2007). We deploy a similar methodology linking international assessments such as PISA, TIMSS, PIRLS, and their precursors, as well as include more regional student achievement tests (RSATs), such as MLA, LLECE, SACMEQ or PASEC[3], which are only partially included in previous papers. This enables us to obtain original data on the quality of student achievement for the largest set of countries to date, and the largest number of developing countries. To our knowledge, this paper presents the largest globally comparable panel database of cognitive achievement, including 163 countries and regions, 32 of which are from Sub-Saharan Africa, over the last 50 years (1965-2015).

The size of our database has a few ramifications beyond sheer coverage. Most notably, we can include many developing countries – the countries that have the most to benefit from educational reform and educational progress. Second, because the methodology we use to link assessments hinges on the existence of enough overlap in countries which take *both* an RSAT and an ISAT, the larger the database, the more overlap, and the more robust *all* transformations. Thus, a larger database enables both the inclusion of developing countries as well as enhances the robustness of the methodology used to include them, making each update significant. Finally, since this database has rich panel data over time and across countries, it can be used to

[3] Respectively the Monitoring Learning Achievement (MLA), the Latin American Laboratory for Assessment of the Quality of Education (LLECE), the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) and the Program on the Analysis of Education Systems (PASEC)

deduce more credibly causal estimates between various drivers of education quality and growth, an elusive yet critical endeavor.

In addition to the size of the database, this is the most current database. Given a series of recent global initiatives which focus on education quality – such as *Education for All* (UNESCO, 2010) and the recent *World Development Report 2018 (*World Bank, 2017) – there is significant demand for the most current, credible and globally comparable measures of education quality. This database provides the largest, most comparable and current learning data.

We include numerous methodological improvements in this paper over prior papers linking ISATs with RSATs by Altinok et al (2014), Angrist, Patrinos and Schlotter (2013) and Altinok and Murseli (2007). First, while previous research relied on a single methodology to anchor assessments, we provide several methods for anchoring and indexing, enhancing the reliability and robustness of our dataset. Previous research papers assume that the distribution of scores on ISATs and RSATs for each country are homogenous across subpopulations and different percentiles. By using alternative anchoring methodologies to link assessments, we provide results which account for varying distributions across sub-populations within a country and test.

Another important contribution of our database is the inclusion of the proportion of students within each country who reach three different international benchmarks (minimum, intermediate and advanced benchmarks) in mathematics, science and reading. This provides an estimate of the distribution of performance, which is essential for our understanding of population-level human capital formation. This is especially critical in societies where inequality is high, since the mean score will be a particularly biased estimate relative to the population at large.

We also include measures of variance and provide confidence intervals of our estimates. This enables explicit quantification of the degree of reliability of each estimate. To do this, we restrict the linking of assessments to those which have micro data available. While this might reduce the coverage of countries we can include, it is a worthwhile trade-off, since it substantially enhances our ability to estimate performance with reliable degrees of confidence.

Finally, we disaggregate results by gender, socioeconomic status, rural/urban, language, and immigration status, thus enabling greater precision and equity analysis. This disaggregation ensures estimates for each sub-population are precise and relevant. Moreover, it enables equity analysis and more detailed understanding of a country's human capital development.

In summary, we contribute to the literature in several ways. First, we present the largest and most current globally comparable dataset. Second, we include the largest number of developing countries. Third, our dataset contains credible measures of globally comparable achievement distributions as well as mean scores. Fourth, we use multiple methods to link assessments, including mean linking and distribution-related linking methods, enhancing robustness of the dataset. Fifth, we include standard errors for our estimates, enabling explicit quantification of the degree of reliability of each estimate. Finally, this dataset has multiple types of disaggregation enabling targeted as well as equity analysis. Overall, we obtain at least one measure of education quality for approximately 163 countries/areas.

## 2. Data

### 2.1. International and Regional Standardized Achievement Tests

This section describes the achievement tests we use to construct our database of *Harmonized Learning Outcomes* (HLOs) which can be compared globally and over time. We divide the assessments into two main groups: The first consists of international assessments; the second is regional assessments. A detailed summary of these assessments is provided in Table 1.

#### *2.1a. International Standardized Achievement Tests (ISATs)*

The Early ISATs (1960 to mid-1990s): **FIMS, FISS, SIMS, SISS, SRC, RLS, MLA and IAEP.** The *International Association for the Evaluation of Educational Achievement* (IEA) was the first body to measure individual learning achievement for international comparison. Tests began in the early 1960s. These tests were precursors to their more current counterparts: *Trends in International Mathematics and Science Study* (TIMSS) and *Progress in International Reading Literacy Study* (PIRLS). The precursors to TIMSS included: pilot studies in 1960, the First International Mathematics Study (FIMS) in 1964, the First International Science Study (FISS) in 1970, the Second International Mathematics Study (SIMS) in 1980-1982, the Second International Mathematics Study (SISS) from 1982-1986, and the *International Assessment of Educational Progress* (IAEP) conducted in 1988 and 1991. Precursors to PIRLS included: Study of Reading Comprehension Study (SRC) in 1970, and the Reading Literacy Study (RLS) in 1990-1991. According to the test developers, the earlier studies served as a model for the later studies (Campbell, Kelly, Mullis, Martin and Sainsbury, 2001; Elley, 1994).

An additional early international assessment - a joint UNESCO and UNICEF project called the *Monitoring Learning Achievement* (MLA) program - covers more than 72 countries and ranges from early childhood, basic and secondary education to non-formal adult literacy (Chinapah,

2003). A series of results reports exist for MLA I across 11 African countries of interest (Botswana, Madagascar, Malawi, Mali, Morocco, Mauritius, Niger, Senegal, Tunisia, Uganda and Zambia; see UNESCO, 2000). However, much of the data has not been published. Since microdata is sparse or often unavailable for the MLA and IAEP data, we include these series only for mean scores and for the total population metrics.

The Modern ISATs (mid 1990s onward): In the mid-1990s, the emergence of standardized, psychometrically-robust and relatively consistent ISATs emerged. Below we describe the major ISATs which we use to construct our database.

**TIMSS.** The Trends in International Mathematics and Science Study (TIMSS) is one of the main survey series conducted by the IEA. Five TIMSS rounds have been held to date in Math and Science subjects covering grades 4 and 8. The first, conducted in 1995, covered 45 national educational systems and three groups of students.[4] The second round covered 38 educational systems in 1999, examining pupils from secondary education (grade 8). The third round covered 50 educational systems in 2003, focusing on both primary and secondary education (grades 4 and 8). In 2007, the fourth survey covered grades 4 and 8 and more than 66 educational systems. In 2011, the survey covered 77 educational systems across grades 4 and 8. The last round was performed in 2015 and covered 63 countries/areas. The precise content of the questionnaires varies but remains systematic across countries.

**PIRLS.** The other dominant IEA survey is the Progress in International Reading Literacy Study (PIRLS). Three rounds of PIRLS have been held to date: in 2001, 2006 and 2011. The PIRLS tests pupils from primary schools in grade 4 in reading proficiency.[5] In 2006, PIRLS included 41 countries/areas, two of which were African countries (Morocco and South Africa), 4 lower-middle-income countries (Georgia, Indonesia, Moldova, Morocco) and 8 upper-middle-income countries (Bulgaria, Islamic Republic of Iran, Lithuania, Macedonia, Federal Yugoslavian Republic, Romania, Russian Federation, South Africa). The latest round of PIRLS was carried out with TIMSS in 2011 and included 60 countries/areas.

In our database, we use all recent IEA studies across two subjects (mathematics and reading/literacy). We use results from official reports (Harmon et al., 1997; Martin et al., 2000;

[4] IEA assessments define populations relative to specific grades, while PISA assessments focus on the age of pupils. In IEA studies, three different group of pupils were generally assessed: pupils from grade 4, grade 8 and from the last grade of secondary education. In 1995, two adjacent grades were tested in both primary (3-4) and secondary schools (7-8). In order to obtain comparable trends, we restricted the sample to grades 4 and 8. Some Canadian provinces and states in the United States of America have occasionally taken part in the IEA surveys.
[5] Similar to TIMSS, pupils from Grade 4 are chosen.

Mullis et al., 2000; Mullis et al., 2003; Mullis et al., 2004; Martin et al., 2007; Mullis et al., 2008; Mullis et al., 2009; Martin et al., 2016; Mullis et al., 2016).

**PISA.** The Organization for Economic Co-operation and Development (OECD) launched the Programme for International Student Assessment (PISA) in 1997 to provide comparable data on student performance. PISA emphasizes an extended concept of "literacy" and an emphasis on lifelong learning – with the aim of measuring pupils' capacity to apply learnt knowledge to new settings. Since 2000, PISA has assessed the skills of 15-year-old pupils every three years. PISA concentrates on three subjects: mathematics, science and literacy. In 2000, PISA had a focus, in the form of extensive domain items, on literacy; in 2003, on mathematical skills; and in 2006 on scientific skills. The framework for evaluation remains the same across time to ensure comparability.[6] In 2009, 75 countries/areas participated; in 2012, 65 countries/areas participated and in 2015, 72 countries/areas participated. A main distinction between PISA and IEA surveys is that PISA assesses 15-year-old pupils, regardless of grade level, while IEA assessments assess grade 4 and 8.

### *2.1b. Regional Standardized Achievement Tests (RSATs)*

In addition to the above international assessments, three major regional assessments have been conducted in Africa and Latin America and the Caribbean.

**SACMEQ.** *The Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SACMEQ) grew out of a national investigation into the quality of primary education in Zimbabwe in 1991. It was supported by the UNESCO International Institute for Educational Planning (IIEP) (Ross and Postlethwaite, 1991). Several education ministers in Southern and Eastern African countries expressed an interest in a similar study. Planners from seven countries met in Paris in July 2004 and established SACMEQ. The current 15 SACMEQ-member education members are: Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, the Republic of South Africa, Swaziland, the United Republic of Tanzania, United Republic of Tanzania (Zanzibar), Uganda, Zambia and Zimbabwe.

The first SACMEQ round took place between 1995 and 1999. SACMEQ I covered seven different countries and assessed performance in reading at grade 6. The participating countries were Kenya, Malawi, Mauritius, Namibia, United Republic of Tanzania (Zanzibar), Zambia and Zimbabwe. The studies shared common features (research issues, instruments, target

[6]As explained in the PISA 2006 technical report, this is only the case for reading between 2000-2009, for mathematics between 2003 and 2009 and for science between 2006 and 2009. See OECD (2010) for more details.

populations, sampling and analytical procedures). A separate report was prepared for each country.

SACMEQ II surveyed grade 6 pupils from 2000-2004 in 14 countries: Botswana, Kenya, Lesotho, Mauritius, Malawi, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania (Mainland), Tanzania (Zanzibar), Uganda, and Zambia. Notably, SACMEQ II also collected information on pupil's socioeconomic status as well as educational inputs, the educational environment and issues relating to equitable allocation of human and material resources. SACMEQ II also included overlapping items with a series of other surveys for international comparison, namely the *Indicators of the Quality of Education* (Zimbabwe) study, TIMSS and the 1985-94 IEA *Reading Literacy Study*.

The third SACMEQ round (SACMEQ III) spans 2006-2011 and covers the same countries as SACMEQ II plus Zimbabwe. SACMEQ III also assess the achievement of grade 6 pupils. The latest round of SACMEQ (SACMEQ IV) began in 2013 in 15 countries, but results are not yet available.

**PASEC.** The "Programme d'Analyse des Systèmes Éducatifs" (PASEC, or "Programme of Analysis of Education Systems") was launched by the Conference of Ministers of Education of French-Speaking Countries (CONFEMEN). These surveys are conducted in French-speaking countries in Sub-Saharan Africa in primary school (grade 2 and 5) for Mathematics and French. Each round includes ten countries. PASEC I occurred from 1996 to 2003; PASEC II from 2004 to 2010 and PASEC III was conducted in 2014.

However, in contrast with other assessments, PASEC has not always been conducted simultaneously across countries and participation has varied considerably since 1994.[7] Moreover, data from the first four assessments is not available[8]. PASEC was modified significantly in 2014, rendering results hard to compare with previous PASEC items. Since scores are not fully comparable between each assessment, we anchor major items to enable international comparability.[9] Currently, we do not include PASEC III results since they require anchoring with SACMEQ IV results, which are not yet available.

[7] The following is a list of participating countries in chronological order: Djibouti (1994), Congo (1994), Mali (1995), Central African Republic (1995), Senegal (1996), Burkina Faso (1996), Cameroon (1996), Côte d'Ivoire (1996), Madagascar (1997), Guinea (2000), Togo (2001), Mali (2001), Niger (2001), Chad (2004), Mauritania (2004), Guinea (2004), Benin (2005), Cameroon (2005), Madagascar (2006), Mauritius (2006), Congo (2007), Senegal (2007), Burkina Faso (2007), Burundi (2009), Ivory Coast (2009), Comoros (2009), Lebanon (2009), Togo (2010), DRC (2010), Chad (2010). Additional countries took a slightly different test between 2010 and 2011 (Lao PDR, Mali, Cambodia and Vietnam).

[8] The first four assessments were mainly pilot studies and the purpose was not to disseminate results.

[9] We are very grateful to the PASEC team, and especially to Jean-Marc Bernard, Antoine Marivin and Vanessa Sy for their help in providing the data. More details concerning the adjustment of the PASEC database is provided in Altinok et al. (2014).

**LLECE.** The network of national education systems in Latin American and Caribbean countries, known as the Latin American Laboratory for Assessment of the Quality of Education (LLECE), was formed in 1994 and is coordinated by the UNESCO Regional Bureau for Education in Latin America and the Caribbean.

Assessments conducted by the LLECE focus on achievement in reading and mathematics. The first round was conducted in 1998 across grades 3 and 4 in 13 countries (Casassus et al., 1998, 2002). These countries include: Argentina, Bolivia, Brazil, Chile, Columbia, Costa Rica, Cuba, Dominican Republic, Honduras, Mexico, Paraguay, Peru and Venezuela (Casassus et al., 1998). The second round of the LLECE survey was initiated in 2006 in the same countries as LLECE I. In round two, called the Second Regional Comparative and Explanatory Study (SERCE), pupils were tested in grade 3 and grade 6. The Third Regional Comparative and Explanatory Study (TERCE), was done in 2013 across grades 3 and 6 and included 15 Latin American and Caribbean countries. Our analysis will include both SERCE and TERCE results, since these assessments are most similar and cover comparable grades. Table 1 summarizes availability and details of the various international and regional assessments listed above.

We link the above assessments and obtain two datasets[10]. The cross-section dataset provides measures of education quality which are aggregated at the education level. We obtain at least one measure of education quality for approximately 163 countries/areas. This covers 82.5 percent of all countries with a population greater than one million, and 90.9 percent of the global population. Out of the 163 countries/areas included, more than 100 are developing economies. Of these, 131 are unique countries. We also obtain a panel database which provides over time comparable scores for education quality between 1965 and 2015. On average, our dataset includes 3.3 observations per country at the primary level and 4.5 observations at the secondary level. Developed countries are over-represented at the secondary level. However, at the primary level, the two groups have similar coverage. Figures 12.0-12.6 present HLO data availability by 5-year interval and coverage

**Table 1. Review of Student Achievement Tests**

| *No* | *Year* | *Organization* | *Abbr.* | *Subject* | *Countries/ Areas* | *Grade/Age* | *Included* | *Survey Series* |
|---|---|---|---|---|---|---|---|---|
| 1 | 1959-1960 | IEA | Pilot Study | M,S,R | 12 | 7,8 | | - |
| 2 | 1964 | IEA | FIMS | M | 12 | 7, FS | ■ | A.1 |
| 3 | 1970-71 | IEA | SRC | R | 15 | 4,8, FS. | | A.1 |
| 4 | 1970-72 | IEA | FISS | S | 19 | 4,8, FS. | ■ | A.1 |
| 5 | 1980-82 | IEA | SIMS | M | 19 | 8, FS | ■ | A.2 |
| 6 | 1983-1984 | IEA | SISS | S | 23 | 4,8, FS | ■ | A.2 |
| 7 | 1988, 1990-91 | NCES | IAEP | M,S | 6, 19 | 4,7-8 | ■ | A.1 |
| 8 | 1990-1991 | IEA | RLS | R | 32 | 3-4,7-8 | ■ | A.1 |
| 9 | Every four years since 1995 (latest round is 2015) | IEA | TIMSS | M,S | 45, 38, 26, 48, 66, 65, 65 | 3-4,7-8, FS | ■ | A.1 (1995), A.2. (Other years - except 2011) |
| 10 | 1992-97 | UNESCO | MLA | M,S,R | 72 | 6,8 | ■ | B |
| 11 | 1997, 2006, 2013 | UNESCO | LLECE | M,S,R | 13, 16 (only 6 for science) | 3,6 | ■ | B |
| 12 | 1999, 2002, 2007 | UNESCO | SACMEQ | M,R | 7, 15, 16 | 6 | ■ | B |
| 13 | 1993-2001,2002-2012, 2014 | CONFEMEN | PASEC | M,R | 22 (before 2014), 10 | Until 2014: 2,5<br>After 2014: 3, 6 | ■ | B |
| 14 | Every five years since 2001 (latest round is 2011) | IEA | PIRLS | R | 35, 41, 55 | 4 | ■ | A.1 (2001), A.2. (Other years - except 2011) |
| 15 | Every three years since 2000 (latest round is 2015) | OECD | PISA | M,S,R | 43, 41, 57, 74, 65, 71 | Age 15 | ■ | A.1 (2000 for reading, 2003 for math, 2006 for science),<br>A.2. (remaining rounds) |

Note: For the meaning of abbreviations, please consult page 21. Only assessments for which there is an information in "Survey Series" column are included in our dataset.
Subjects: M=math; S=science; R=reading.

## 3. Methodology

We propose a methodology which enables comparison among various existing international and regional assessments. We obtain a *Harmonized Learning Outcomes* (HLO) database, which is comparable over a set of 163 countries / areas form 1965-2015. The foundation for our approach is to index across a given pair of achievement tests with results from countries that participate in both. To link results over time, we perform a similar procedure using the United States as an anchor since it has participated in all IEA assessments since 1965 as well as a consistently administered national assessment, the *National Assessment of Educational Progress (NAEP)*. Similarly, we use results from the national assessment conducted in Burkina Faso to anchor PASEC results over time. First, we present methodologies which can be used for anchoring assessments. Then, we show how we obtained the final anchored dataset.

### *3.1 Linking Methodologies*

Various methodologies can be used for linking or equating assessments. Equating is a statistical process that is used to adjust scores on tests so that scores can be used interchangeably (Kolen and Brennan, 2014). The purpose of equating is to adjust for difficulty among assessments that are built to be similar. In our case, assessments are not directly comparable since difficulty and content may differ. Instead, we use a similar approach to equating, known as *scaling to achieve comparability* according to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999). This is also known as *linking* in the terminology of Holland and Dorans (2006), Linn (1993) and Mislevy (1992). As explained in Kolen and Brennan (2014), similar statistical procedures are used in linking and equating, although their purposes are different. In this paper, we use the terminology of *linking* instead of *equating* since the tests we link used are purposefully built to be different. Notably, 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 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 a large enough instance of 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 results 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.

While IRT might not be a reliable approach when there is limited item-by-item overlap, we propose a few robustness tests in Section 5 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). We conduct a series of additional robustness tests. Namely, we compare scores and ranks of our estimates relative to ranks and raw scores for the original tests used for linking. 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 Hoskens, 2003). Specifically, we use pseudo-linear linking and equipercentile linking. Below, we describe each, starting from a foundation of mean linking.

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 $X$ on the scale of test $Y$, which is a *Reference Test*, while test $X$ is the *Anchored Test*. For instance, we would like to link a test like *PISA 2003* on another assessment like *TIMSS 2003*. Therefore, *PISA 2003* will be the *Anchored Test X* while *TIMSS 2003* will be the *Reference Test Y*.

**Mean linking.** In mean linking, *Anchored Test X* is considered to differ in difficulty from *Reference Test Y* by a constant amount along the score scale. Define *Anchored Test X* as the new test, let *X* represent the random variable score on score *X*, and let *x* represent a particular score on *Anchored Test X*. Define Test Y as the reference test, let *Y* represent the random variable score on *Reference Test Y*, and let *y* represent a particular score on Test *Y*. Define $\mu(X)$ as the mean on Test X and $\mu(Y)$ as the mean on *Reference Test Y* for a population of pupils. In mean linking, scores on the two tests that are an equal distance away from their respective means are set equal:

$$X - \mu(X) = Y - \mu(Y) \tag{1}$$

We then solve for *y* and obtain:

$$linking_Y^m(x) = y = x - \mu(X) + \mu(Y) \tag{2}$$

In this equation, $linking_Y^m(x)$ refers to a score *x* on *Anchored Test X* transformed to the scale of *Reference Test Y* using mean equating. In other words, mean equating involves the addition of a constant $(-\mu(X) + \mu(Y))$ to all raw scores on *Anchored Test X* to find anchored scores on *Reference Test Y*. This linking methodology assumes that assessments have the same distribution, which is often unlikely.

**Linear linking.** Linear linking allows for the differences in difficulty between the two tests to vary along the score scale. In this case, scores that are an equal distance from their means in standard deviation units are set equal. Define $\sigma(X)$ and $\sigma(Y)$ as the standard deviations of *Anchored Test X* and *Reference Test Y*, respectively. The linear conversion sets standardized deviation scores (z-scores) on the two tests to be equal such that:

$$\frac{x-\mu(X)}{\sigma(X)} = \frac{y-\mu(Y)}{\sigma(Y)} \tag{3}$$

Solving for *y* in Eq. (3),

$$linking_Y^l(X) = y = \sigma(Y)\left[\frac{x-\mu(X)}{\sigma(X)}\right] + \mu(Y) \tag{4}$$

where $linking_Y^l(X)$ is the linear conversion equation for converting observed scores on *Anchored Test X* to the scale of *Reference Test Y*. By rearranging terms, an alternate expression for $linking_Y^l(X)$ is:

$$linking_Y^l(X) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X)\right] \tag{5}$$

This expression is a linear equation of the form *slope (x) + intercept* with:

$$slope = \frac{\sigma(Y)}{\sigma(X)}, \text{ and } intercept = \mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X) \quad (6)$$

In linear linking, scores on *Anchored Test X* are adjusted allowing for the tests to be differentially difficult along the score scale. Note that if the standard deviations for the two tests were equal, which assumes the distribution is the same, and Eq. (3) can be simplified to Eq. (2). In this case, we are left with an adjustment by a constant amount that is equal to the difference between the *Reference Test Y* and the *Anchored Test X* means, as in mean linking.

In summary, in mean linking we transform original to anchored scores by setting the deviation scores on the two tests equal, whereas in linear linking we set the standardized deviation scores (*z-scores*) on the two tests equal.

In our case, the difficulty between tests is different, especially between regional and international assessments. Thus, linear linking is best suited to our purposes. However, linear linking does not enable linking assessments over time, since assessments vary, rendering standard deviation comparisons misleading.

**Pseudo-linear linking.** Altinok *et al.* (2014) and Angrist, Patrinos and Schlotter (2013) use a fusion of mean and linear linking to obtain anchored scores. This estimation method uses the difference in means in the *Anchored Test X* and *Reference Test Y* as a coefficient adjustment:

$$linking_Y^{pl}(X) = y = \frac{\mu(Y)}{\mu(X)}x \quad (7)$$

where $linking_Y^{pl}(X)$ is the pseudo-linear conversion equation for converting observed scores on *Anchored Test X* to the scale of *Reference Test Y*. We prefer to use this hybrid approach instead of linear linking to preserve over-time comparability of anchored tests. If we use the linear-linking approach, this limits comparability if standard deviations are not stable over time, as is often the case.

Hanushek and Woessmann (2012) adopt an approach that adjusts the coefficient with both means and standard deviations where $linking_Y^{pl2}(X)$ is the pseudo-linear conversion equation for converting observed scores on *Anchored Test X* to the scale of *Reference Test Y*. The main drawback of this methodology is the potential variation of standard deviations for a given country over time. This assumption is particularly tenuous for developing countries, limiting the ability to make credible comparisons of education quality over time.

**Equipercentile linking.** Equipercentile linking was developed by Braun and Holland (1982). Equipercentile linking is best used when X and Y differ nonlinearly in difficulty. For instance,

*Anchored Test X* could be more difficult than *Reference Test Y* for high score but less difficult for low scores. The equipercentile linking function is developed by identifying scores on *Anchored Test X* that have the same percentile ranks as scores on *Reference Test Y*. Consider the following definitions of terms, where X and Y are continuous random variables.

*F(x)* is the cumulative distribution function of *X* in the population. This is defined as the proportion of examinees in each population who score at or below *x* on test X for a given population *T*. Formally: $F(x) = P\{X \leq x \mid T\}$ where $P\{\,. \mid T\}$ is the probability or population proportion in each population *T*.

*G(y)* is the cumulative distribution function of *Y* in the population. This is defined as the proportion of examinees in each population who score at or below *y* on test Y for a given population *T*. Formally: $G(y) = P\{Y \leq y \mid T\}$ where $P\{. \mid T\}$ is the probability or population proportion in each population *T*.

In equipercentile linking, we set the cumulative distributions of X and Y equal:

$$F(x) = G(y) \tag{9}$$

When the cumulative distribution functions are continuous and strictly increasing, we can always solve for y:

$$linking_Y^e(X) = G^{-1}[F(x)] \tag{10}$$

where $G^{-1}$ is the inverse of the cumulative distribution function *G(y)*.

In summary, equipercentile linking is broken into three main steps: we first find the percentile rank of *x* in the *Anchored Test X* distribution. Then, we find the score that has the same percentile rank in the *Reference Test Y* distribution. Then we find the equivalent score of *Reference Test Y* for *Reference Test* X based on their common percentile.

A limitation of simple equipercentile linking is that when score scales are discrete, which is the case for ISATs and RSATs, we are not able to find corresponding scores for test scores or percentiles not observed in the sample. For example, if in my observed sample, the closest percentile matches are a score with a 47.2 percentile on *Reference Test X* and a score on *Reference Test Y* with a 47.6 percentile, I have rough equivalence, but do not have an exact percentile match.

One approach to dealing with this limitation is to use percentile ranks. However, this might not yield adequate precision. Moreover, this approach does not enable future linking above the highest or lowest observed scores used for equating. Increasing sample sizes can alleviate these

concerns to an extent, but is often insufficient. To this end, smoothing methods have been developed to deal with sampling error and produce estimates of the empirical distributions and equipercentile relationship best characterizing the underlying population. This enables interpolation at each point on the curve, enhancing precision of the equating exercise.

Two general types of smoothing can be conducted. In *presmoothing*, the score distributions are smoothed using polynomial loglinear presmoothing (Holland and Thayer, 2000); in *postsmoothing*, the equipercentile equivalents are smoothed using cubic-spline postsmoothing (Kolen, 1984). We use the *presmoothing* loglinear method, which is the same method used by the ETS, and is based on Von Davier et al. (2004) and Holland and Thayer (1987, 2000)[11].

Three assumptions must hold for the linking methods above to be valid. First, they must test the same underlying population. Given we are using sample-based ISATs and RSATs and equate using overlapping countries, this assumption is satisfied if the population tested is similar and participation rates reach a certain threshold or non-participation is random. Second, tests should measure similar proficiencies. We link across precise dimensions such as subject and schooling level (primary vs. secondary) to increase the likelihood of proficiency overlap. Finally, the distribution of proficiency should be similar across tests. We address this assumption by equating using an average across countries that participate in both tests. The reliability of the equating exercise is enhanced with an increase in the number of countries that take both tests being equated. We include robustness checks to demonstrate the sensitivity of our results to this effect. We also include confidence intervals for our estimates to quantify the degree of uncertainty.

We compute two types of education quality metrics: (a) the proportion of students achieving international benchmarks of performance; and (b) mean scores. For the first set of metrics, threshold levels of achievement, we use the presmoothed equipercentile method to capture the distribution of scores. For the second metric, mean scores, we use pseudo-linear linking. This methodology enables credible over-time comparisons, a central feature of our panel dataset, and is consistent with a growing literature in economics on globally comparable education quality data.

### *3.2 Features of the Methodology*

#### *3.2.1. Comparability across Countries and over Time*

[11] We used R Statistics software for the equipercentile linking. In particular, we use the "equate" package. See Albano (2016) for more information.

In the linking theories above, the anchoring process is done by adjusting results from the same population between two tests, or with the same items used in different tests. In our case, we use the former approach. We examine the same population between two tests to determine the relationship between *Reference Test X* and *Anchored Test Y*. To this end, we compare the same countries at the same point in time which took an ISAT and an RSAT. Since ISATs and RSATs are psychometrically-robust, sample-based test designed to be nationally representative, they represent the same underlying population at the country-level. Thus, by comparing *doubloon* countries which participate in both tests being linked, we can index difficulty and scales across tests. Table 2 provides the list of countries that overlap in assessments[12]. This enables inclusion of Regional Standardized Achievement Tests (RSATs) from Latin America and Sub-Saharan Africa and thus international comparison. This is a significant addition, since many developing countries have participated in RSATs (LLECE, SERCE, PASEC, and SACMEQ) but rarely or never in ISATs (PISA, TIMMS, PIRLS). Transformation of regional scores into an internationally comparable value is most accurate the more *doubloon* countries are available. If our index relies on just one *doubloon* country (if it is the only country participating in both surveys), it is ambitious to convert all other regional scores using this quotient. We provide robustness tests on the sensitivity to the number of doubloon countries.

In addition to indexing learning outcomes across assessments using *doubloon* countries, we anchor assessments across time using the *National Assessment of Educational Progress* (NAEP) in the United States. This is possible since the U.S. participated in NAEP and various international achievement test at every interval. For example, if the performance of the United States changed in NAEP in a given year but did not in the same subject and year in which in the IEA assessment the U.S. took part in, it would mean that the IEA study is upward or downward biased. To correct for this under- or over-estimation, we adjust old IEA studies by trends on NAEP results.[13] We only include the NAEP adjustment for scores before the 1990s since standardized ISATs began to be conducted consistently from the 1990s onwards and are therefore comparable over time.

[12] For example, when linking PISA 2003 and TIMSS 2013 the 15 overlapping countries are: countries are Australia, Hong-Kong, Hungary, Indonesia, Italy Japan, Republic of Korea, Latvia, Netherlands, New Zealand, Norway, Russian Federation, Slovakia, Sweden, Tunisia and the United States).

[13] A similar methodology was used for linking PASEC assessments over time. We used results for a national assessment in Burkina Faso, which provides over-time comparable scores, and also took part at PASEC in 2006 and 2014. After linking PASEC assessments onto a single scale, we used the participation of Mauritius in both PASEC and SACMEQ for linking PASEC to our internationally anchored scale. However, PASEC is an assessment for Francophone countries, while SACMEQ focuses on Anglophone countries. This might bias the anchoring process for adjusted reading scores. Since Mauritius has been tested in both languages in PASEC (English and French) we can use this to correct for language differences.

Using this approach, which builds on Altinok *et al.* (2014), Angrist, Patrinos & Schlotter (2013), and Altinok and Mureseli (2007), regional test scores for countries participating in an RSAT but not in an ISAT can be transformed into an internationally comparable score over time. These test scores allow the inclusion of developing countries which participate only in regional assessments to be included in our international achievement data set. We conduct this analysis for all countries, including those that participated in an ISAT to apply this transformation equally and limit bias. As a result of this relatively simple methodological innovation, we can build a globally comparable database of *Harmonized Learning Outcomes (HLOs)* for 163 countries/areas from 1965-2010.[14]

*3.2.2. Grade, Subject and Year Grouping*

While it would be ideal to have a test score for every year and grade, test frequency is too low and sporadic. To this end, we group test scores into five-year and grade range intervals. Specifically, we construct a score for each subject (Math, Reading and Science) and grade range (Primary or Secondary) for every 5-year interval. This increases data and country coverage substantially and is aligned to the approach taken by Barro and Lee (2001) for educational attainment. We conduct this exercise by grouping test scores that are comparable by subject and grade if they are administered a few grade levels or years apart. If countries participated in several comparable tests in or around a specific year, we build the average over the tests.

This approach has obvious limitations. Namely, an extra grade or year of learning would mechanically improve learning. For example, grouping grade 4 and grade 6 into 'primary' schooling or 2000 and 2001 tests in to a 2000-year interval ignores the fact that students perform better as the moves across grades and over time. However, these differences within group (primary vs secondary, or a few years over or above a five-year step) are often small, limiting bias. Moreover, since this methodology is applied across all countries and intervals it unlikely that one country is transformed differently from another, further limiting the scope for bias. The main instance in which bias might arise is if data availability is correlated with a countries' education quality our progress. For example, if countries that perform worse only have available

[14]We standardize our final dataset with a standard deviation of 30 and mean of 500. This is analogous to many of the ISAT and RSATs means and standard deviation at the country-level is approximately the same as the observed value in our dataset. We do this for a group of stable countries, the OECD, in line with the methodology proposed by Hanushek and Woessmann (2016). This enables us to know where Finland, for example, lies relative to the average OECD country. We conduct this ex post standardization for all countries. However, for countries off the OECD scale, this standardization is biased and less relevant. For these countries, their relative position and rank is robust, although their absolute score difference should be caveated appropriately.

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 than they are. While this bias might exist, it creates a conservative metric. Moreover, we put up with this limitation to enhance data availability and coverage.

Similar limitations exist when using the index to transform across tests as when grouping intervals across grades and time within tests. A regional test might measure a different grade or be administered in a different year than an international test. For example, the regional SERCE test is specific to grade 6, while the international TIMSS test might be specific to grade 8. Furthermore, the SERCE test was conducted in 2006 while the TIMSS test was conducted in 2007. Therefore, even if the mean score for all countries that took a regional test such as SERCE in 2006 is unbiased, when we divide the SERCE 2006 mean by the TIMSS 2007 mean, we might be concerned about the integrity of the index.

This potential bias, however, does not seriously affect the outcome of our methodology for two important reasons. First, we use the index to translate all original scores. Since the same index is used for all original scores, each score is transformed equally. Second, it is unlikely that tests changed from year to year in a way that differentially affected certain countries. For example, even if TIMSS 2007 was made more challenging because of 2006 SERCE test scores, which is relatively unlikely to begin with, this change should not impact Colombia more than Bolivia. Thus, the index we produce can be a powerful and relatively unbiased tool to link international achievement tests with regional tests.

Additional assumptions and limitations revolve around data availability and coverage. 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 the meaning of each score by eliminating the ability to differentiate math and reading scores. To this end, we construct disaggregated measures as well as aggregated ones. This enables us to conduct analyses with each, considering the trade-offs.

#### *3.2.3. Specific choices made in the anchoring process*

We use all available information on performance for each country to obtain the most precise and accurate panel dataset between 1965 and 2015.

For linking early time intervals, where some countries took part in the FISS and SISS without participating in FIMS and SIMS, we estimate a countries' performance by regressing their

scores in FISS on FIMS. The constant in the regression captures the potential difference between the two subjects. In addition, we estimate the score of countries which took part in FISS and SISS based on the variation of their performance, instead of only using the level, capturing trends in schooling performance over time. The data for trends between FIMS and SIMS comes from Robitaille and Garden (1989) while the data concerning FISS and SISS can be found in Keeves (1992).

For countries that took part in a PIRLS assessment, without participating in a TIMSS test we estimate scores for countries which took part in both PIRLS 2001 and TIMSS 2003, and then compute performance by using the growth rate between PIRLS 2001 and PIRLS 2006 (instead of estimating the PIRLS 2006 scores based on the TIMSS 2007 dataset). A similar process was used for data between PIRLS 2006 and PIRLS 2011.

Anchoring for PISA 2000 was made with a similar approach. Scores from PISA 2000 were estimated in mathematics using the TIMSS 1999 assessment. For countries which took part in both TIMSS 2003 and TIMSS 2007, we use the growth rate of scores between PISA assessments to estimate performance. When a country took part in both PISA and TIMSS assessments, we used only results from TIMSS. When possible, PISA trends are directly used. PISA assessments permit over-time comparability for reading between 2000 and 2015, for mathematics between 2003 and 2015 and for science between 2006 and 2015[15].

It should be noted that to conduct all linking methods, we need to access to the micro data of all assessments. Unfortunately, this is not always possible. Therefore, when the micro data is not available, we restrict our anchoring methodology to the pseudo-linear method used by Altinok *et al.* (2014) and Angrist, Patrinos and Schlotter (2013). This is the case for MLA assessments for which raw data is not available.

### *3.2.4 Inclusion of Standard Errors*

We include standard errors for each test based on definitions provided in each test's technical report. This enables us to quantify to an extent the uncertainty around our estimates. For doubloon countries, we use the standard error of the original assessment. In future iterations of this dataset, we aim to include additional metrics of uncertainty to provide a series of plausible bounds on our estimates.

### *3.2.5 Construction of Proficiency Thresholds*

[15] Given the fact that some countries took part in the PISA 2009 study in 2010, their results have been adjusted for 2009 by predicting their performance level in 2010.

We construct results for the proportion of students achieving minimum, intermediate and advanced benchmarks using the presmoothed equipercentile linking method. Recent research on education quality includes only mean scores, without information on within country distributions of cognitive skills. Distributional information on education quality is important for understanding the dynamics of education quality and growth, especially in often unequal developing economies.

In the growth literature, there are two main views regarding the channel through which education enhances growth. The first view argues for investing in the top performers who would boost innovation (Nelson and Phelps, 1966; Aghion and Howitt, 1998; Aghion, Meghir & Vandenbussche, 2006; Galor, 2011) while the alternative view argues for a more egalitarian school system to ensure well-educated masses (Mankiw, Romer & Weil, 1992). Aghion and Cohen (2004) distinguish economies of imitation from economies of innovation. This motivates investment in primary and secondary schooling and attainment of basic skills in developing economies to support imitation. In contrast, high income countries might be best off investing in higher education, supporting innovation on the technological frontier. These alternative views are reflected in different policy goals. For example, "Education 2030" focuses attention on providing most pupils with a minimum level of proficiency in mathematics and reading in developing countries (UNESCO, 2015).

To this end, we provide new measures at three different benchmarks (minimum, intermediate and advanced) to enable analysis of educational performance with a distributional lens. While the proportion of students reaching the minimum benchmark would better fit with an egalitarian economy, the share of students at the advanced level may be more suited for economies which aim at innovating. Moreover, if developing countries focus on high performers, this can bias mean scores up relative to society at large. If there is interest in performance for the median citizen, distributional information on performance can help triangulate analysis along this dimension, addressing biases inherent in mean scores which are susceptible to outliers.

Table 3.1-3.3 summarizes the benchmarks used across primary and secondary education for each subject and provides a description of expected competencies at each. For primary education, we use benchmarks defined by PIRLS and TIMSS. These are 400, 475, and 625 for low, intermediate and advanced benchmarks, respectively, across all three subjects. For secondary education, we use the PISA benchmarks. The low threshold is approximately 400, the intermediate, roughly 475, and the advanced somewhat above 600, although for secondary education each benchmark varies slightly by subject. Notably, the use of international

thresholds might not be relevant for developing countries, where small percentages of countries pupils might attain the upper benchmarks. In the future, we hope to use alternative thresholds.

## 4. Results

We construct two complementary *Harmonized Learning Outcomes (HLO)* datasets:

(a) *Panel Data Set.* Our panel database provides over time comparable scores for education quality from 1965 to 2015. On average, our dataset includes 3.3 observations per country at the primary level and 4.5 observations at the secondary level.

(b) *Cross-Sectional Data Set*. Our cross-sectional dataset provides measures of education quality averaged across time and subject. We obtain at least one measure of education quality for approximately 163 countries/areas, 100 of which are developing economies and 30 in Sub-Saharan Africa.

We provide estimates for mean scores, standard errors, and low, intermediate and advanced proficiency benchmarks. We also include disaggregated estimates across: subject, school level (primary, secondary); gender, socioeconomic status, language (if the test language is the same language spoken at home), geographic location (urban, rural), and immigration status.

### *4.1. International Comparison of Education Quality using the Cross-Sectional Dataset*

Figures 1 and 2 present educational achievement among regions. Asian countries seem to outperform countries from other regions in the primary and secondary level, followed by North America and Europe. Latin America and the Caribbean and Northern Africa are the next best performers, followed by Sub-Saharan Africa. The regions that perform worst, Sub-Saharan Africa and Southern Asia, have larger gaps in primary education performance than secondary education performance. Among middle income countries, those in Eastern Europe and Central Asia perform the best. Developing countries perform worse in both primary and secondary education than developed countries, and have much larger variance, especially at the primary level. While variation is high, and the top-performing country in Sub-Saharan African still performs lower than the lowest performing country in developed economies.

An important note when interpreting these results is that they might capture a series of selection effects. One such selection effect is driven by immigration. If a country tends to attract the most able students from around the world, it will have high mean scores. For instance, PISA 2009 results showed that approximately 15 percent of pupils from Singapore have an immigrant background, while this proportion is more than 40 percent in Qatar, which might drive mean scores up due to immigration rather than school quality alone. We can check this empirically

for each country and overall by analyzing whether results are higher or lower conditional on immigration status. Another selection effect is driven by enrollment. For example, if some countries do not have support and infrastructure to enable retention from primary to secondary schooling, the remaining students in secondary might be the highest performers. To this end, increases in secondary schooling performance could be driven by a selection effect rather than value-added learning.

Figures 3-5 show results by gender. Regions with an average above the zero-line are the ones where the female learning premium is positive and girls outperform boys. In the Middle East and South-eastern Asia females tend to most outperform males. Developing regions have higher gender-based variance in performance than developed countries. Overall, our results show small gender gaps conditional on girls being enrolled in school: most regions and countries have a gender gap of less than 5 points. In terms of direction of the gender gap, there is not a consistent pattern, with the female premium toggling between positive and negative depending on the region.

Figure 6a and 6b show results for a sub-set of developing and developed countries by percentage of students reaching one of three proficiency primary and secondary schooling benchmarks: minimum, intermediate and advanced. This provides crucial information on the distribution of performance on a global scale, a key feature of our dataset, in addition to the level of performance. We compare figure 6b for primary scores to Figure 7 which includes mean scores for the same countries. A few notable trends emerge.

First, on a global scale, developing countries in Sub-Saharan Africa and to a lesser extent Latin America consistently place last, both according to mean scores as well as minimum proficiency thresholds. Second, of Sub-Saharan African countries shown, less than 50 percent of students meet the minimum proficiency threshold on a global scale; developed countries are consistently above 80 percent. Moreover, often a higher percentage of students in developed countries achieve the intermediate benchmark than the percentage of students who achieve the minimum benchmark in developing countries.

Third, there is significant information that mean scores alone miss. For example, Finland has a higher percentage of students meeting the minimum and intermediate threshold than Japan, but since Japan has more students achieving advanced performance, higher a higher mean score. Similarly, the United States has fewer students meeting minimum and intermediate proficiency benchmarks than Germany, but has more students at the advanced level, and thus a higher mean score. Since the mean score is highly responsive to outliers, it is possible a small group of top

performers in the United States can skew the mean; if we care about a wide pool of the population acquiring basic skills, however, the mean is a distortionary metric. This effect is similar when comparing South Africa to Tanzania, where South Africa has fewer students meeting the minimum threshold, but more students reaching intermediate and advance thresholds, so has a higher mean performance. A notable comparison of distributions is Zimbabwe and Swaziland. Zimbabwe has a lower percentage of students meeting the minimum threshold relative to Swaziland, but the percentage of students meeting the intermediate benchmark is much higher, clustered close to the percentage meeting the minimum, whereas in Swaziland the percentage of students meeting the intermediate benchmark is clustered closer to those meeting the advanced threshold. Thus, Zimbabwe, has a higher mean score.

When analyzing the entire dataset, we see that less than 50 percent of students reach the minimum global threshold of proficiency in developing countries relative to 86 percent in developed countries. For intermediate benchmarks, 25 percent of students reach the threshold relative to 66 percent in developed countries; and for advanced benchmarks only 2 percent of students reach the global threshold relative to 10 percent in developed countries.

Overall, distributional information reveals critical information on the absolute level of education quality attained by a larger pool of society, the egalitarian nature of education quality, and can shed light on the debate over whether economic performance is driven by a few innovative members of society at the top, or an education society at large.

### *4.2. Long-term Performance Trends (1965-2015) using the Panel Data Set*

Our database provides the largest, most current globally comparable panel database on education quality. Table 4 provides summary statistics for each year and level of schooling in our *Harmonized Learning Outcomes* (HLO) database. We observe that at the primary level, mean scores have fluctuated over time, but overall increased. Moreover, variation has also fluctuated but has stabilized more recently. At the secondary level, performance overall has decreased. This is likely driven by enrollment selection effects, where more poor performers stay in the system.

In Table 5, we rank countries by the variation in their secondary schooling quality from 1980 and 2015. We observe gains for most countries with data availability over this period, with annual growth rates in achievement ranging from 0.10 to 0.62 percent. We see the largest gains for Hong Kong, followed by Iran and Finland. Notably, a few countries have experienced a decline in performance, including France, Hungary, Thailand and Chile. Figure 8 present results

of those countries with 50 years of data extending all the way back to 1965. These provide important examples of the potential effects of successful versus failed policy reforms, with Thailand experience fluctuations in performance and an overall drop, Israel experiencing fluctuations and an overall increase, and Finland experiencing a relatively steady increase in performance.

Figures 9.0-9.6 present results for a few select countries with error bars capturing standard errors of the original tests. This enables us to capture the measure of uncertainty around our estimates. We observe relatively tight confidence intervals for Finland relative to Germany for example, revealing that Finland's education progress is robust, while Germany's recent gains in the last fifteen to twenty years might in fact be closer to flat progress, although there has been significant progress since the mid-1990s.

In Figures 11.0-11.4 we map out HLO coverage and learning trends for each country from 1965-2015. Notably, only 26 countries have test score data extending continuously back 20 years, and only 8 countries have test scores extending continuously back 50 years. Despite this, this dataset presents the richest panel dataset on globally comparable education quality to date.

Figure 10 demonstrates a statistically significant and positive relationship between educational achievement and economic growth. This association is consistent with results from Hanushek and Woessmann (2012).

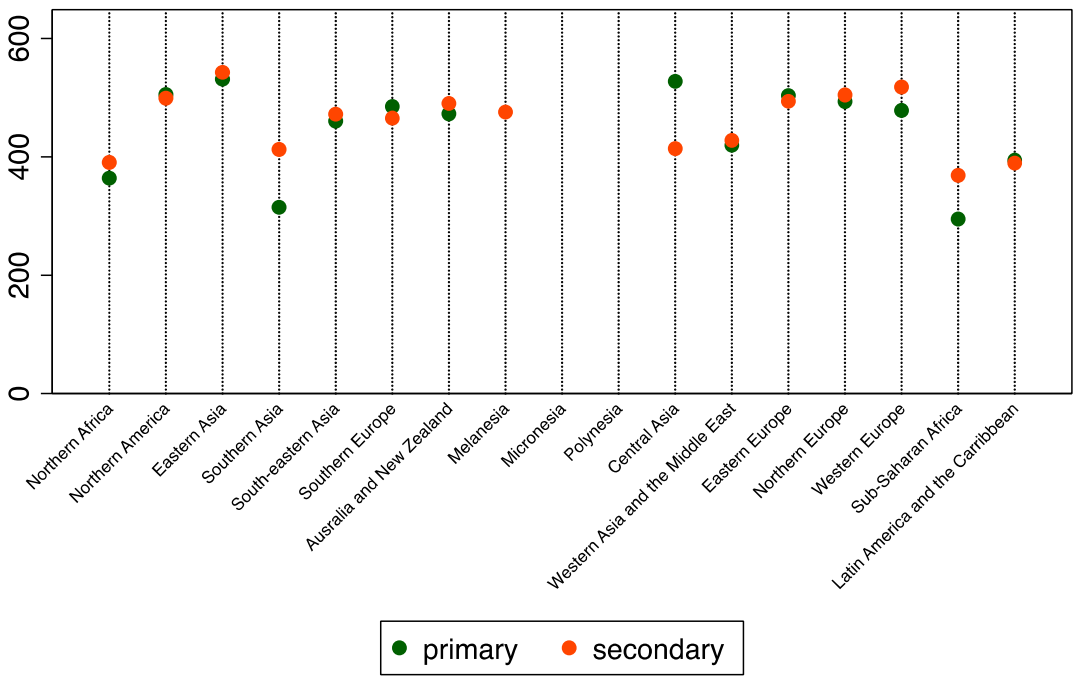



**Figure 2:**

Performance on Average across 1965-2015 for Primary, Secondary and All Scores

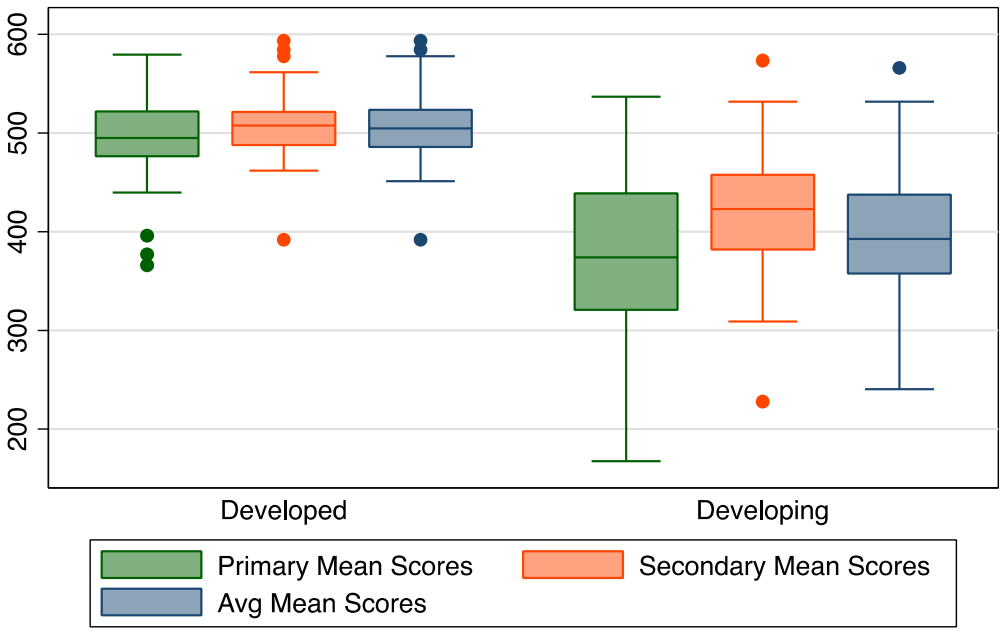


**Figure 3:**

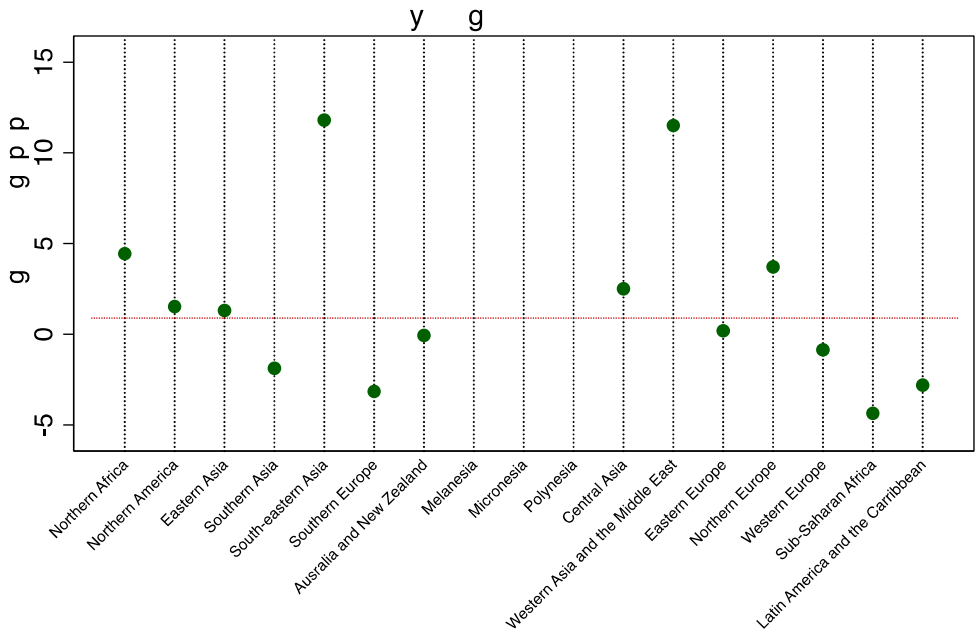



**Figure 4:**

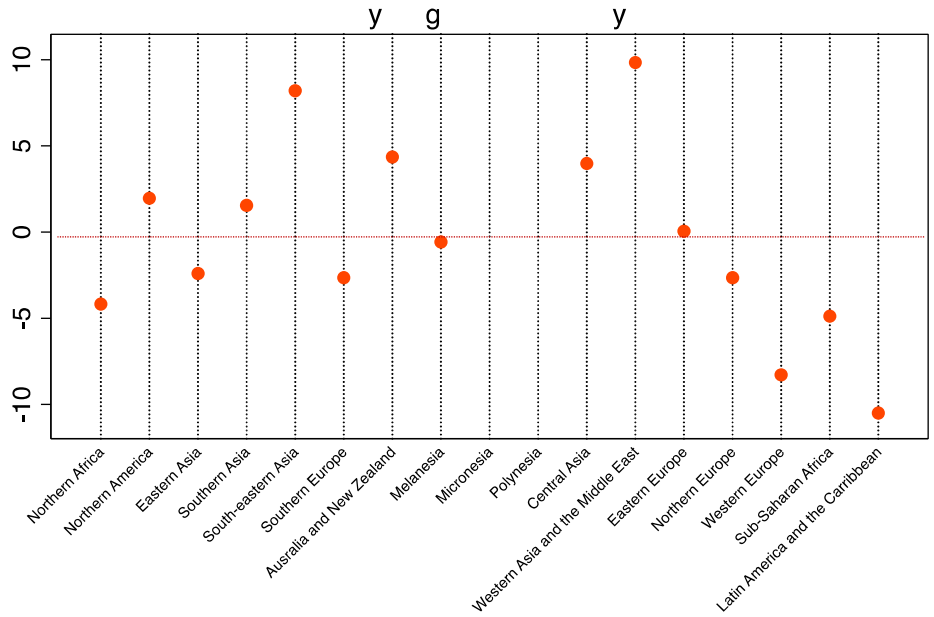


**Figure 5:**

Average Gender Gap across 1965-2015 for Primary, Secondary and All Scores

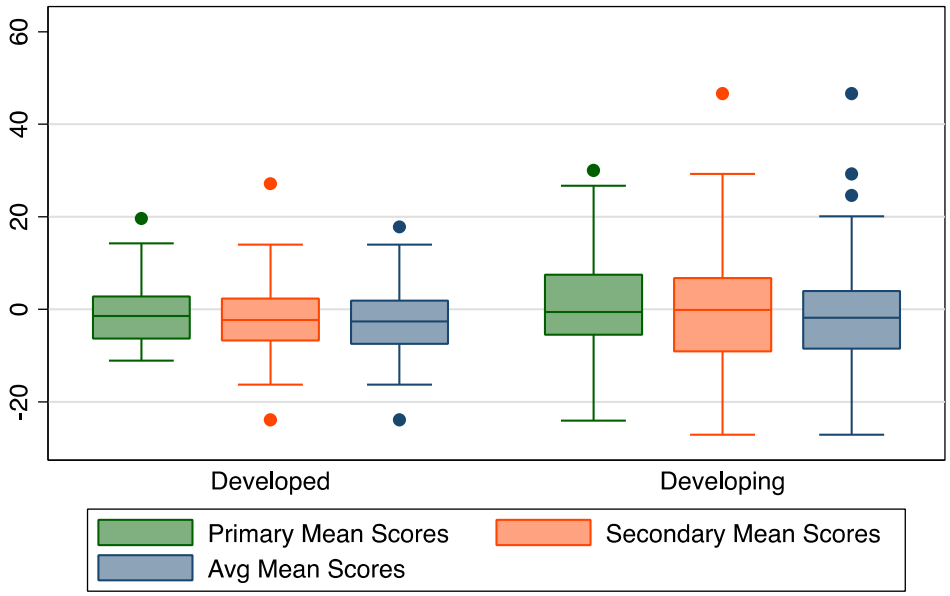


**Figure 6a:** Percent of Students Achieving Low, Intermediate and Advanced Average

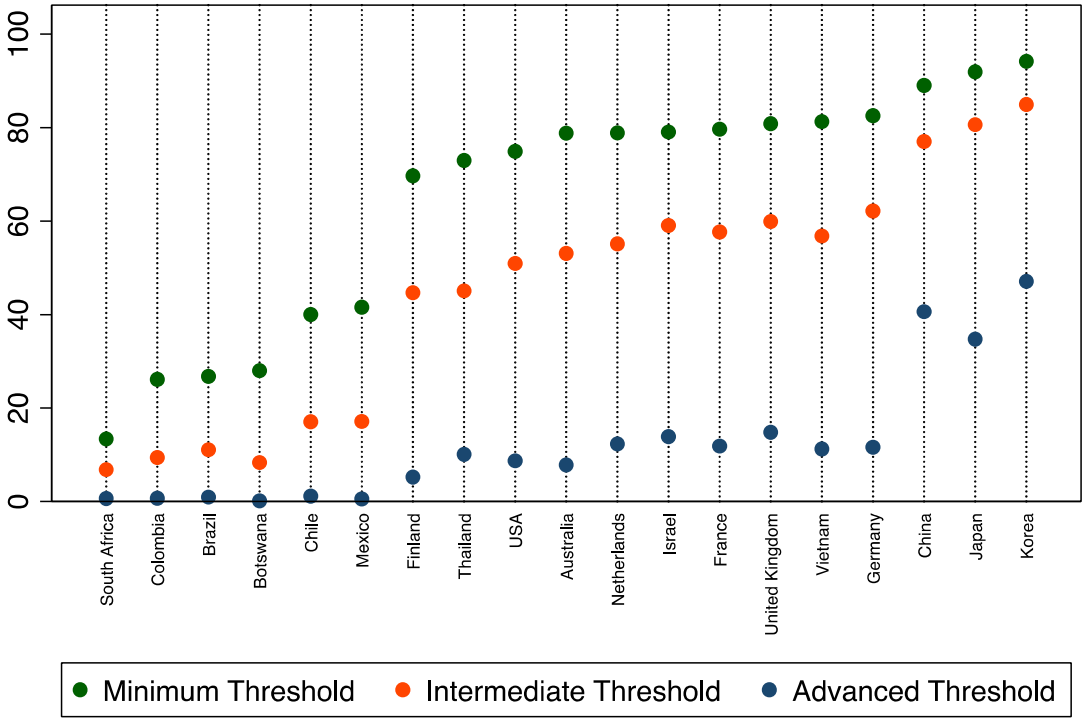



**Figure 6b:** Percent of Students Achieving Low, Intermediate and Advanced Average Primary

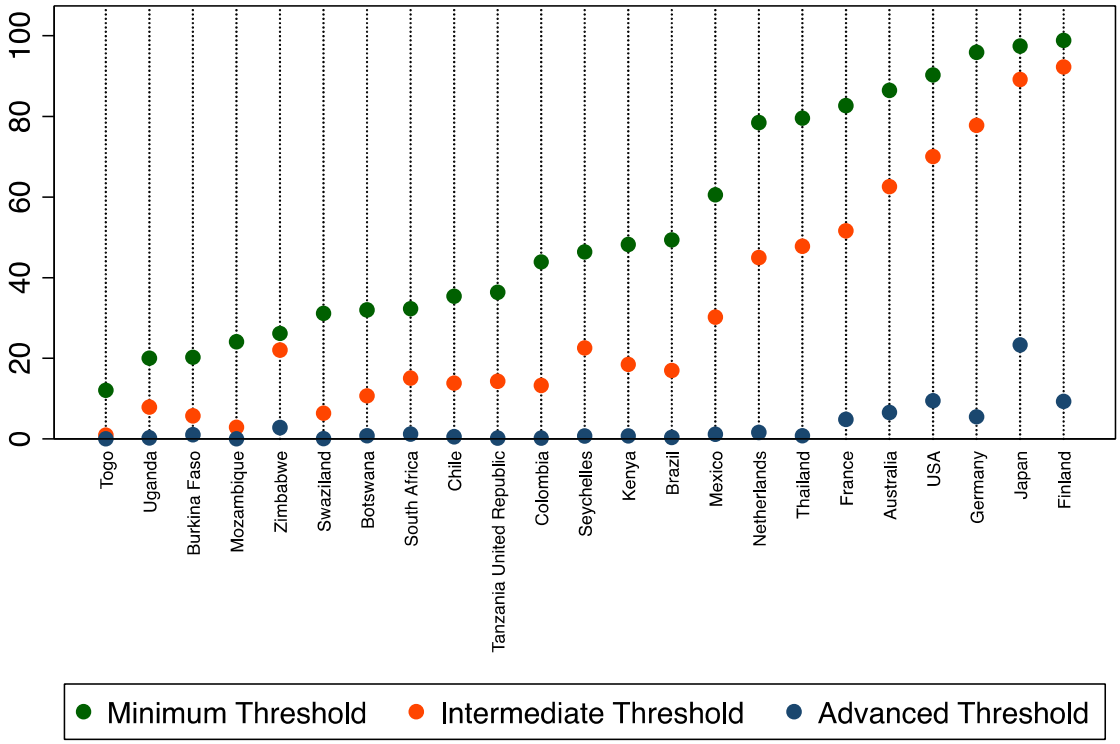


**Figure 7:** Mean Average Secondary Score in sub-set of Developing and Developed

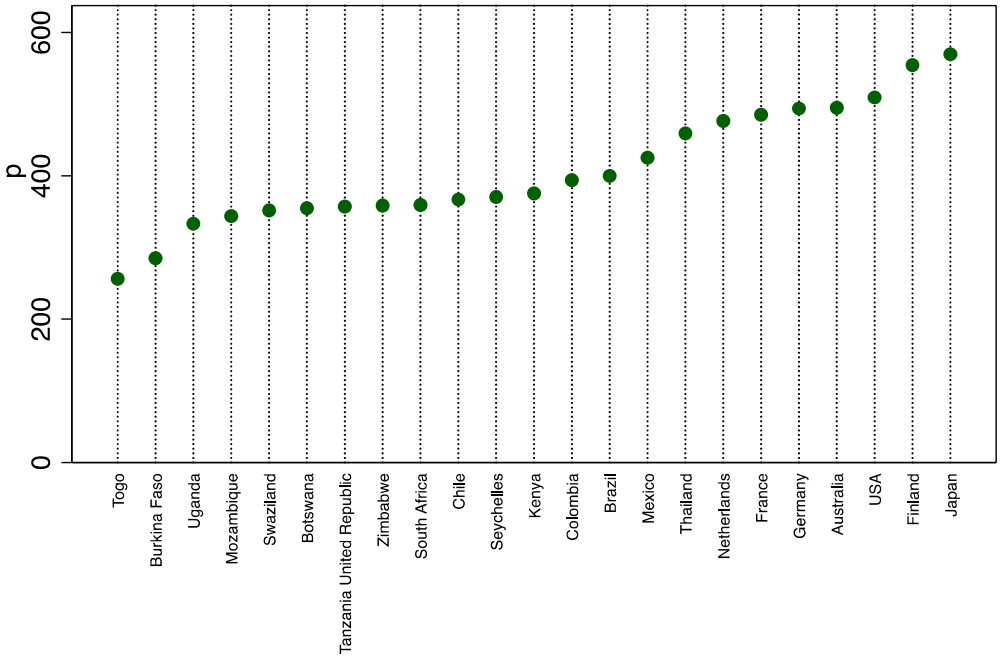


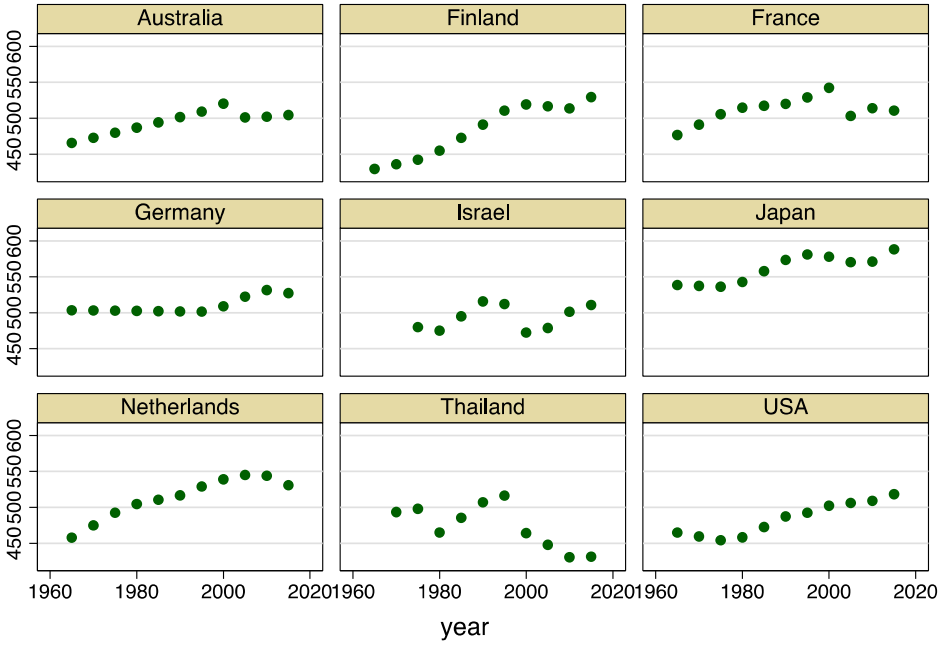
**Table 4: Summary Statistics for Panel Dataset**

| Year | Countries/Areas | Mean | Standard Deviation | Minimum | Maximum |
|---|---|---|---|---|---|
| Primary Level | | | | | |
| 1970 | 16 | 454.06 | 60.41 | 313.31 | 558.11 |
| 1975 | 10 | 401.08 | 155.46 | 3.60 | 559.92 |
| 1980 | 10 | 403.76 | 153.24 | 12.40 | 561.74 |
| 1985 | 22 | 470.94 | 56.88 | 331.92 | 572.11 |
| 1990 | 39 | 444.48 | 88.06 | 28.42 | 569.66 |
| 1995 | 62 | 421.30 | 94.06 | 197.56 | 600.04 |
| 2000 | 74 | 425.85 | 106.75 | 167.49 | 626.45 |
| 2005 | 110 | 430.47 | 101.44 | 190.08 | 628.58 |
| 2010 | 87 | 457.52 | 89.14 | 182.73 | 604.31 |
| 2015 | 69 | 484.80 | 68.08 | 352.14 | 617.37 |
| Secondary Level | | | | | |
| 1965 | 10 | 487.51 | 36.87 | 429.55 | 538.55 |
| 1970 | 22 | 476.14 | 38.10 | 345.70 | 537.43 |
| 1975 | 19 | 480.01 | 37.46 | 359.16 | 536.31 |
| 1980 | 28 | 473.74 | 33.45 | 373.15 | 542.88 |
| 1985 | 35 | 483.26 | 29.48 | 387.67 | 557.97 |
| 1990 | 53 | 475.87 | 84.09 | 227.81 | 645.59 |
| 1995 | 46 | 492.46 | 66.30 | 277.71 | 608.59 |
| 2000 | 65 | 482.33 | 74.13 | 271.83 | 604.66 |
| 2005 | 97 | 466.68 | 68.55 | 283.64 | 598.88 |
| 2010 | 96 | 464.90 | 66.66 | 325.61 | 609.21 |
| 2015 | 89 | 476.18 | 62.37 | 339.34 | 620.96 |

**Table 5:** Long-term secondary trends on schooling quality for 22 economies, 1980-2015

| Country | 1980 Score | 2015 Score | Variance (points) | Variance (%) | Annual Growth Grate |
|---|---|---|---|---|---|
| Hong Kong, China | 489.34 | 594.99 | 105.65 | 21.59 | 0.62 |
| Iran Islamic Republic of | 373.15 | 436.35 | 63.20 | 16.94 | 0.48 |
| Finland | 454.83 | 529.49 | 74.66 | 16.41 | 0.47 |
| USA | 458.21 | 518.21 | 60.00 | 13.09 | 0.37 |
| Luxembourg | 450.07 | 503.28 | 53.21 | 11.82 | 0.34 |
| Sweden | 448.09 | 500.72 | 52.63 | 11.75 | 0.34 |
| England | 466.61 | 517.86 | 51.25 | 10.98 | 0.31 |
| Japan | 542.88 | 588.36 | 45.48 | 8.38 | 0.24 |
| Canada | 490.26 | 527.28 | 37.02 | 7.55 | 0.22 |
| Israel | 475.10 | 510.89 | 35.79 | 7.53 | 0.22 |
| Canada, Ontario | 485.73 | 522.30 | 36.57 | 7.53 | 0.22 |
| United Kingdom | 481.92 | 510.22 | 28.30 | 5.87 | 0.17 |
| Belgium | 497.79 | 525.25 | 27.46 | 5.52 | 0.16 |
| Netherlands | 504.57 | 530.71 | 26.14 | 5.18 | 0.15 |
| New Zealand | 468.75 | 492.72 | 23.97 | 5.11 | 0.15 |
| Germany | 502.61 | 527.29 | 24.68 | 4.91 | 0.14 |
| Italy | 473.48 | 494.39 | 20.91 | 4.42 | 0.13 |
| Australia | 487.01 | 504.53 | 17.52 | 3.60 | 0.10 |
| France | 514.70 | 510.68 | -4.02 | -0.78 | -0.02 |
| Hungary | 526.16 | 514.48 | -11.68 | -2.22 | -0.06 |
| Thailand | 464.91 | 431.42 | -33.49 | -7.20 | -0.21 |
| Chile | 475.80 | 427.57 | -48.23 | -10.14 | -0.29 |

**Figure 8.** Long-term trends for selected countries, 1965-2015



## Figures 9.0-9.6: Long-Term Trends of Selected Countries with Confidence Intervals

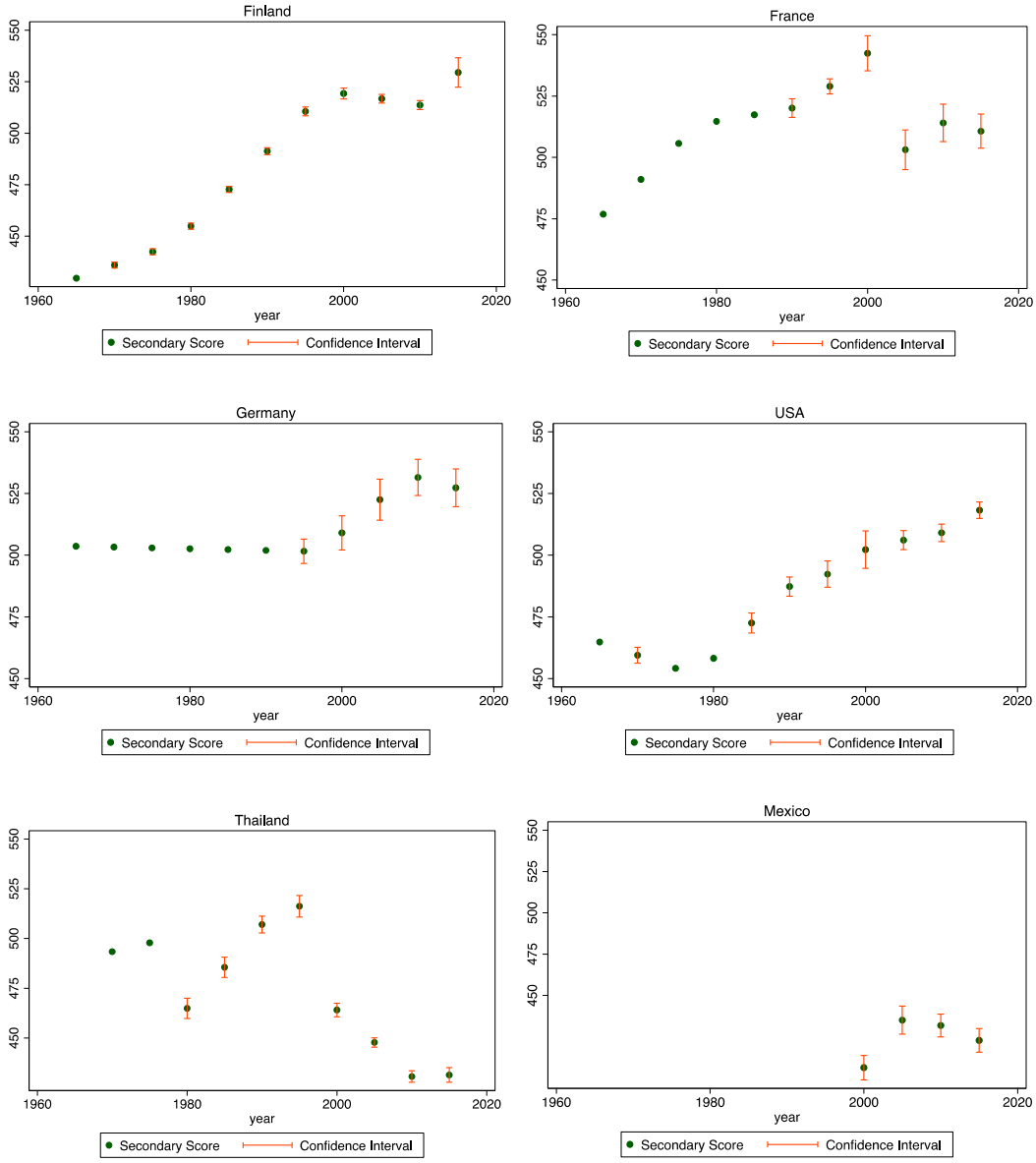
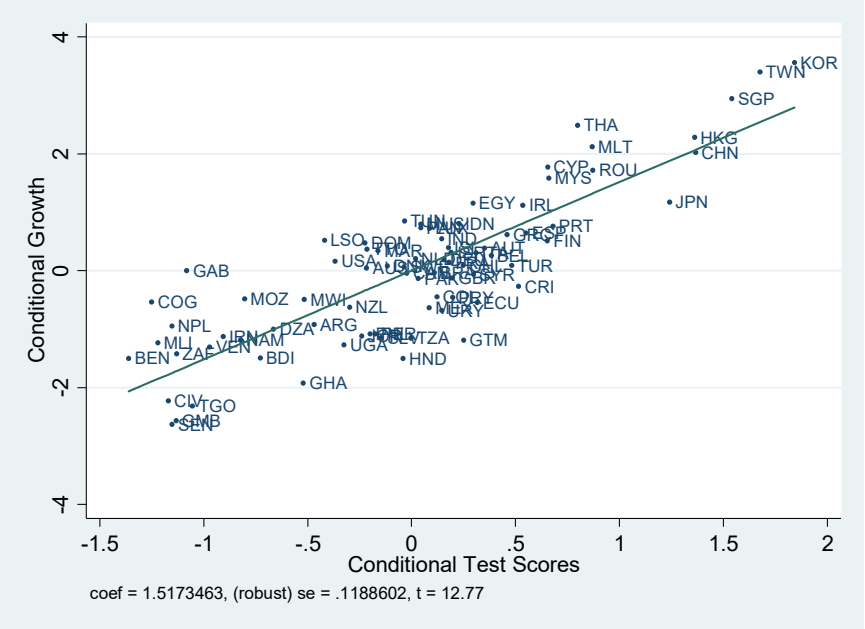


**Figure 10.0: Educational Achievement and Economic Growth (1965-2015)**




Note: Added-variable plot of a regression of the average annual rate of growth (in percent) of real GDP per capita in 1965-2015 on the initial level of real GDP per capita in 1960 and average scores on international and regional student achievement tests.

## 5. Limitations and Robustness Checks

### *5.1. Limitations*

*5.1.1. Differences in distributions.* When we link assessments, we assume that our reference and anchor test have a similar distribution of scores. In practice, this assumption is tenuous. For example, if test items on SACMEQ are easier than on TIMSS, this would skew the distributions. In Table 6 we include various robustness tests to this effect. These tests show that indeed the distributions differ, especially when comparing the PASEC and SACMEQ to ISATs. We address this limitation by using the equipercentile method to link assessments according to percentile ranks, adjusting for differences in the distribution. Moreover, we generate an internationally standardized threshold of performance which is linked to matching underlying competencies. Notably, the use of international thresholds might not be relevant for developing countries, where small percentages of countries pupils might attain the upper benchmarks. In the future, we hope to use alternative thresholds. In this context, means scores are limited, and equipercentile thresholds are likely more meaningful, since they account for changes in the

distribution. As an additional measure, we provide standard errors to capture the measure of uncertainty around linking assessments.

*5.1.2. Differences in underlying populations.* Since PISA measures 15-year olds, and TIMSS measures grades (4 and 8), it is possible that the underlying populations do not perfectly match. However, despite occasional diverging results, Rindermann and Stephen (2009) have found the correlation between PISA and TIMSS to be high. While overall the results are relatively consistent, we account for some of the differences that do exist by using TIMSS as our default anchor reference assessment, rather than averaging across both TIMSS and PISA.

An additional concern in ensuring similarity of the underlying populations is non-participation. Since ISATs and RSATs are sample-based and designed to be representative at the country-level, in principle, doubloon countries should ensure the underlying populations are similar. However, if randomly sampled schools and students do not participate for non-random reasons, this would jeopardize the representativeness of the sample of a given test at the country level. As an assurance against this, ISATs and RSATs have strict rules on threshold participation rates. For example, PISA requires an 85 percent participation rate at the school level, and an 80 percent participation rate within school at the student level. If these benchmarks are not met, the results are excluded or caveated appropriately. This is a safeguard to ensure ISATs and RSATs accurately represent the nation's underlying population.

**Table 6. Robustness check: Comparison of main statistics between assessments for the restricted doubloon countries samples**

| Number | Assessment 1 | # of countries | Mean | SD | Skewness | Kurtosis | Assessment 2 | Mean | SD | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | LLECE I math | 1 | 488.5 | 71.3 | 0.1 | 5.5 | TIMSS 1995, grade 8 | 343.9 | 80.3 | 0.2 | 3.3 |
| 2 | LLECE I Reading | 2 | 506.4 | 88.4 | 0.1 | 3.5 | PIRLS 2001 reading | 429.5 | 86.5 | -0.2 | 2.9 |
| 3 | LLECE III math | 2 | 549.7 | 99.4 | 0.4 | 3.1 | TIMSS 2011 math | 442.3 | 86.5 | 0.0 | 2.8 |
| 4 | LLECE III reading | 2 | 512.3 | 81.5 | 0.2 | 3.0 | PIRLS 2011 reading | 460.9 | 72.3 | -0.1 | 2.9 |
| 5 | SACMEQ II math | 2 | 497.0 | 96.8 | 0.8 | 5.1 | TIMSS 2003, grade 8 | 304.2 | 102.7 | 0.2 | 2.9 |
| 6 | SACMEQ III reading | 1 | 497.9 | 115.0 | 0.6 | 2.9 | PIRLS 2006 reading | 295.3 | 123.5 | 0.5 | 3.2 |
| 7 | PISA 2000, 15 years old students, math | 12 | 492.1 | 113.5 | -0.3 | 2.7 | TIMSS 1999, grade 8, math | 521.5 | 99.9 | -0.4 | 3.3 |
| 8 | PISA 2003, 15 years old, math | 12 | 490.8 | 108.5 | -0.1 | 2.7 | TIMSS 2003, grade 8, math | 517.2 | 91.9 | -0.2 | 3.0 |
| 9 | PASEC II, math | 1 | 482.2 | 235.7 | -0.2 | 2.2 | SACMEQ III, math | 619.2 | 135.9 | 0.2 | 2.6 |
| 10 | PASEC II, reading | 1 | 500.6 | 249.7 | -0.5 | 2.0 | SACMEQ III, reading | 570.9 | 120.2 | 0.1 | 2.3 |

*5.1.3. Content differences across tests*. Some tests measure core competencies while others measure content knowledge, and exactly which domains are tested varies. To this end, we aim to provide separate results for three core subjects: reading, math and science. While imperfect, since, for example, some math tests might cover double digit recognition, while others might focus on higher-order multiplication problems, subject-specific scores provide a reasonable level of meaningful differentiation for essential cognitive skills and education quality estimates correlated with growth outcomes.

*5.1.4. Doubloon country robustness*. When the number of doubloon countries is small, there is scope for significant bias to arise in our linking function from country-specific or time-specific factors, rather than test-specific differences. As the number of doubloon countries is increased, this bias is reduced. We perform a few empirical tests to measure the sensitivity of our results to this bias. First, we compare the linking function between PISA and TIMSS across various simultaneous instances: 2003 and 2015. Second, instead of using all doubloon countries, we split the sample into two parts to measure the stability of the linking function. If the linking function is stable, using half of the sample should produce similar results to the full sample. Table 7 presents the results of this exercise for the United States. Using the presmoothing equipercentile method, we see no difference in scores in 2003, and a difference of 6-7 points in 2015. For the pseudo-linear linking method, there is a difference range of about 20 points in 2013, which is reduced to around 7 points in 2015. This exercise indicates that although there are some notable differences depending on the method and time period, the differences are often small. This suggests the doubloon methodology is relatively robust. When linking regional to international assessments it is likely the methodology is less robust since the overlap in doubloon countries is small. Since this dataset, albeit imperfect, significantly expands the overlap in doubloon countries relative to prior similar datasets, we present estimates that are the most robust to doubloon country sensitivity to date.

*5.1.5. Disaggregation*. We provide estimates for subsamples where microdata is available across a few dimensions: gender, geographic location, socioeconomic status, immigration status, and language. It is important to note that each individual ISAT and RSAT is not necessarily designed to be nationally representative for each subpopulation. This means that when linking assessments, the use of doubloon countries does not guarantee the underlying population represented is the same. This is only the case when the scores being examined are representative at the country-level. To this end, while disaggregation provides more precision in theory, in practice the linking function is less robust. However, given the importance of equity analysis, we put up with the limitation and provide disaggregated estimates with this caveat.

*5.1.6. Threshold definitions.* The primary TIMSS benchmark for low performance, when linked by underlying competency, is approximately equivalent to Level 6 in SACMEQ, Level III in LLECE and Advanced Level in PASEC before 2014.[16] However, the minimum benchmark from SACMEQ provides a more realistic benchmark for low-income countries; indeed, it aligns much better with other RSAT thresholds and is roughly equivalent to Level 1 in LLECE and the intermediate benchmark in PASEC before 2014. To this end, while the low benchmark from TIMSS is technically easier to standardize globally, since, for example, Levels 1-5 in SACMEQ have no equivalent on TIMSS, it might be less relevant for developing countries. In the future, we hope to use alternative thresholds.

*5.1.7. Differences across linking methodologies.* As mentioned earlier, our results vary across linking methodologies. To this end, we include estimates from the methodologies most 'fit for purpose.' For mean scores, this is the pseudo-linear linking method, and for thresholds it is the equipercentile linking method. In the future, we provide a measure of variance across estimates produced by linking methodologies to capture this additional element of uncertainty.

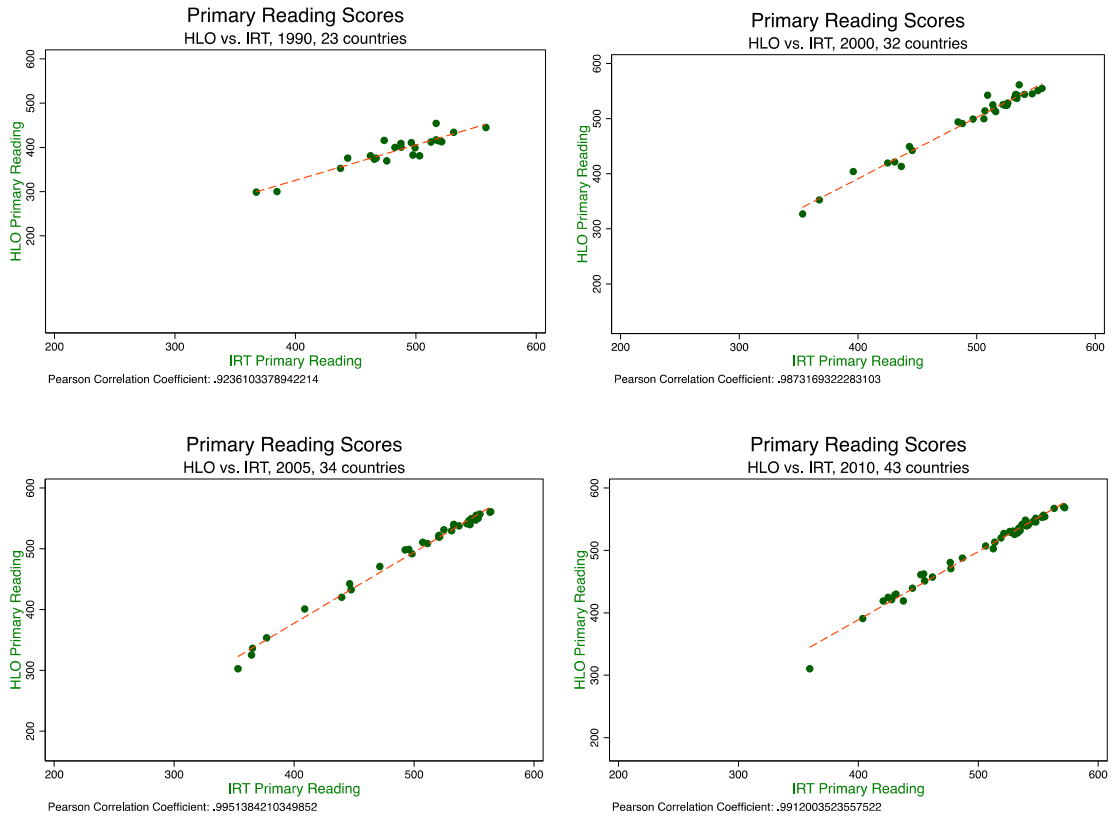
*5.2. Robustness Checks*

*5.2.1. Comparison to LINCS.* We compare our Harmonized Learning Outcomes (HLO) for primary reading scores with scores generated using an IRT linking methodology by the LINCS project which leverages overlap in items for a subset of ISATs focused on reading at primary school from 1970 onwards (Strietholt, 2014; Strietholt and Rosén, 2016). Where there is overlap in country coverage, these results consistently indicate a high Pearson correlation coefficient of

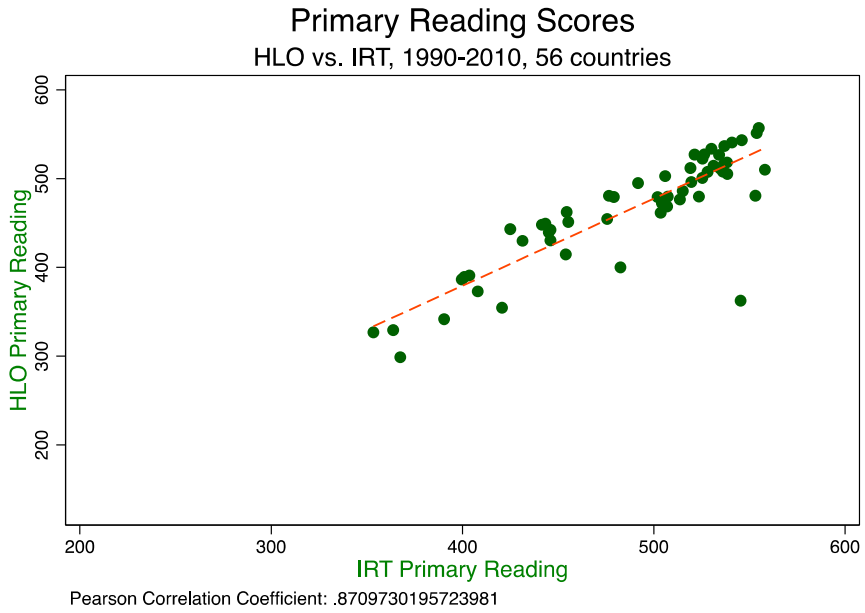
[16] We considered that the PASEC study (before 2014) has three different thresholds: minimum (20 points), intermediate (40 points) and advanced (60 points). For practical reasons, we multiplied the PASEC scores by 10 to obtain scores in a scale comparable to other assessments.

.92 or above. We exclude 1970 since we do not have Harmonized Learning Outcomes for primary reading in this year. In 1990 our Pearson correlation coefficient is .92, in 2000 it is .98 and in 2005 and 2010 it is roughly .99. This indicates that our methodology performs similarly for this subset of scores. Figures 11.0-11.3 below depict the results in graph form. When we average results over 1990-2010, the Pearson correlation coefficient is .87. This indicates we are able to produce similar results to the IRT methodology where there is overlap, in addition to doubling country coverage, with over 100 HLO countries relative to 59 in LINCS.

**Figures 11.0-11.3: Robustness - HLO vs. IRT Primary Reading Scores by Year**



**Figures 11.4: Robustness - HLO vs IRT Primary Reading Scores from 1990-2010**




We conduct a similar exercise using ranks within each respective year. Figures 13.0-13.3 in the Appendix summarize the results. We find consistently high correlations above .95 for all years, except 1990 where the Spearman rank correlation is lower than raw scores at .84. This indicates that the relationship between our HLO database and the LINCS project database is high even when using rank correlations which might be sensitive to small changes in point estimates.

*Comparison to ISAT and RSAT raw scores.* As an additional robustness check, we compare our Harmonized Learning Outcome (HLO) database to raw scores from PISA, TIMSS, SACMEQ and LLECE. If our methodology is able to preserve the integrity of the original scores and rankings, in addition to expanding coverage, this would enhance our confidence in the robustness of our approach. Figures 14.0-14.9 demonstrate the results. Of note, we conduct comparisons from 1995-2015 for all assessments, except LLECE where we compare results form 2005-2015 since we did not include the first LLECE in our HLO database.[17] For PISA and TIMSS we see almost a direct mapping to our HLO database. With SACMEQ and LLECE, we see similar trends, different by a relatively constant factor. This is the expected result: similar trends over time scaled down when placed on an international scale.

[17] As noted earlier, the first LLECE is not analogous to the SERCE and TERCE due to varying grade-level coverage.

We would expect TIMSS and PISA results to vary little after transformation since they are already on an international scale. Table 8.0 quantifies this change. We see that TIMSS scores change less than a single point on average. While PISA math and science scores are underestimated by our HLO by around 20 points, this is a relatively small difference, and for reading scores there is essentially no difference between PISA scores and our HLO outcomes. This indicates a relatively high degree of robustness in comparison to raw ISAT scores.

We expect scores for developing countries to change in our HLO database since indeed this is the purpose of the linking function. However, ranks should be preserved relative to their original rank within each group of countries that participated in the original RSAT used in the anchoring process. We verify whether this assumption holds.

Although we see RSAT scores change significantly, especially for SACMEQ, ranks remain stable. The average rank change in math for SACMEQ countries is 0, with a standard deviation of .2 for math scores; LLECE countries' math scores are slightly more sensitive to rank changes, with an average rank change of 0 and a standard deviation of 1. However, in both instances, original RSAT ranks are consistent with our HLO ranks, indicating a high degree of robustness. Qualitatively, we see that the first SACMEQ has no change in ranks, and SACMEQ II has one change in ranks with a one-rank swap between Tanzania and Seychelles. For SERCE there are a few rank shifts ranging from shifts of one to five. In TERCE there are only two shifts of 1-rank swaps between Nicaragua and Panama and Colombia and Ecuador. We conduct a similar exercise across TIMSS primary and secondary math scores from 1995-2015, as well as PISA secondary reading scores from 2000-2015. We find an overall average rank changes of 0 for all assessments, with standard deviations ranging from 1.4-2.5. If you include RSATs, the average rank change remains 0, and the average standard deviation shift is 1.4. Table 8.1 and Figures 15.0-15.16 show these results in depth. Additional results are available on request.

Overall, we find that our HLO database while more expansive, produces similar results to both raw scores, ranks and IRT-equated scores where there is overlap. This increases our confidence in the robustness of our estimates. We also provide standard errors to provide an added measure of reliability by quantifying the uncertainty around our methodology and estimates.

A future robustness check would be the inclusion of psychometric adjustments. Jerrim et al. (2017) highlight a few features of ISATs that economists often ignore. 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 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.

## 7. Conclusion

To meaningfully compare learning across countries, we need a measure of learning that is comparable. The growth of international standardized achievement tests, which are carefully constructed, psychometrically tested, standardized assessments implemented globally is a huge step in this direction. However, the countries that participate in these tests are often high and middle-income countries. This limits our ability to track, compare, or understand education patterns in developing countries – the countries that often have the most to gain from education.

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 happens, it will take a long time. Moreover, this approach would render a rich array of retrospective data null, limiting longitudinal and panel 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* database which builds on previous work.

Recently, UNESCO endorsed the harmonization of learning outcomes as a useful approach as part of the Global Alliance to Monitor Learning (GAML) (UIS, 2017). This database is part of a World Bank and UNESCO Institute for Statistics (UIS) partnership to advance this effort. In the long term, the ambition and ideal is to deploy a worldwide proficiency assessment for numeracy and literacy. Until then, and to enable rich longitudinal panel data analysis, 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.

The crux of our harmonization methodology hinges on construction of an index that enables us to include developing countries, and more countries overall. We use *doubloon* countries that participate in both regional and international assessments as an anchor. This enables inclusion of regional assessments from Latin America and Sub-Saharan Africa. Because of this relatively

simple methodological innovation, we build a globally comparable database of 163 countries/areas from 1965-2015, approximately two-thirds of which are in developing economies, and 30 in Sub-Saharan Africa, representing more than 90 percent of the global population. We build both a cross-sectional and panel database.

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 relatively consistent estimates where possible.

We contribute to the literature in several ways. This is the largest and most current globally comparable dataset, including the most developing countries. In addition to mean scores, this dataset also contains measures of globally comparable achievement distributions, namely, threshold attainment of low, intermediate and advanced proficiency thresholds. Moreover, this dataset uses multiple methods to link assessments, including pseudo-linear linking and presmoothed equipercentile linking methods. This enhances the robustness of each estimate. To enhance the robustness and reliability of the dataset further, we include standard errors of our estimates, enabling explicit quantification of the degree of certainty around each estimate. We also include estimates that are disaggregated across multiple parameters: gender, socioeconomic status, rural/urban, language, and immigration status, thus enabling greater precision and equity analysis.

A first analysis of this dataset reveals a few important trends:

1. Learning outcomes in developing countries often cluster at the bottom of a global scale
2. Although variation in performance is high in developing countries, the top performers still often perform worse than the bottom performers in developed countries
3. Gender gaps are relatively small, with high variation in the direction of the gap by region
4. Distributions reveal meaningfully different trends than mean scores, with less than 50 percent of students reaching the minimum global threshold of proficiency in developing countries relative to 86 percent in developed countries.

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 both 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


Aghion P. and Cohen E., (2004). Éducation et Croissance. La Documentation française, Paris, 2004.

Aghion P. and Howitt P., (1998). *Endogeneous Growth Theory*. MIT Press, Cambridge.

Altinok, N. (2017). Mind the Gap: Proposal for a Standardised Measure for SDG 4-Education 2030 Agenda. Montreal: UNESCO Institute of Statistics (UIS) Information Paper 46.

Altinok, N., Diebolt, C., & de Meulemeester, J.-L. (2014). A New International Database on Education Quality: 1960-2010. Applied Economics 46 (11), 1212-1247.

Altinok, N., Murseli, H. (2007). "International database on Human Capital Quality". *Economics Letters*, 96(2), pp. 237-244.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Angrist, N., Patrinos, H.A., & Schlotter, M. (2013). An Expansion of a Global Data Set on Educational Quality (No. WPS6536). The World Bank Policy Research Working Paper.

Barro, R.J. (2001). "Education and Economic Growth". in Helliwell, J.F. (Ed.), *The Contribution of Human and Social Capital to used Economic Growth and Well-Being* (pp. 14-41). Paris: OECD Press.

Barro, R.J., Lee, J.W. (1993). "International Comparisons of Educational Attainment". *Journal of Monetary Economics*, 32, pp.363-394.

Barro, R.J., Lee, J.W. (2001). "International Data on Educational Attainment: Updates and Implications". Center for International Development Working Paper no. 45, Harvard University.

Barro, R.J., Lee, J.W. (2010). " New Data Set of Educational Attainment in the World: 1950-2010", *NBER Working Paper*, No. 15902.

Barro, R.J., Lee, J.W. (2012). "A New Data Set of Educational Attainment in the World: 1950-2010".

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic.

Brown, G., Micklewright, J., Schnepf, S.V., Waldmann, R. (2005). "Cross-National Surveys of Learning Achievement: How Robust are the Findings?". Southampton Statistical Sciences Research Institute, Applications & Policy Working paper, A05/05.

Campbell, J. R., Kelly, D. L., Mullis, I. V. S., Martin, M. O., & Sainsbury, M. (2001). Progress International Reading Literacy Study (PIRLS). *International Association for the Evaluation of Educational Achievement (IEA), Second Edition. Chestnut Hill, MA, USA: PIRLS International Study Center*.

Card, D., Krueger, A.B (1992). "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States," *Journal of Political Economy*, 100(1), pp.1-40, February.

Chinapah, V. (2003). "Monitoring Learning Achievement (MLA) Project in Africa", *ADEA Biennial Meeting 2003.* Grand Baie, Mauritius, 3-6 December.

Ciccone A., Papaioannou E. (2005). "Human Capital, the Structure of Production, and Growth". Barcelona, Universitat Pompeu Fabra.

Cohen, D., Soto, M. (2007). "Growth and Human Capital: Good Data, Good Results". *Journal of Economic Growth*, 12(1), pp.51-76.

Coulombe, S., Tremblay, J.-F. (2006). "Literacy and Growth", *Topics in Macroeconomics*, 6(2). Berkeley Electronic Press.

Coulombe, S. and J.F. Tremblay (2007), "Skills, Education and Canadian Provincial Disparity", *Journal of Regional Science*, vol. 47, nr. 5, pp. 965-991, December.

Das, J., & 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.

De la Fuente, A., Domenech, R. (2002). "Educational Attainment in the OECD, 1960-1995". *CEPR DP 3390*.

De la Fuente, A., Domenech, R. (2006). "Human Capital in growth regression: How much difference does quality data make?". *Journal of the European Economic Association*, 4(1), pp. 1-36.

Demeulemeester and Diebolt (2011), "Education and Growth: What Links for Which Policy?", *Historical Social Research*, 36, pp. 323-346.

Demeulemeester, J.-L. and Rochat (1997), "Convergence versus divergence between European countries: the case of higher education systems", *Brussels Economic Review*, ULB - Université Libre de Bruxelles, 153, pp. 3-19.

"The EFA movement". United Nations Educational, Scientific and Cultural Organization. Retrieved 11 Sep 2010.

Galor, O. (2011). Unified Growth Theory and Comparative Development. Rivista di Politica Economica, SIPI Spa, issue 2, 9-21, April-Jun.

Gurgand, M. (2000). "Capital humain et croissance: la littérature empirique à un tournant?", *Économie Publique*, 6, pp. 71-93.

Hanushek, E.A., Rivkin, S.G., Taylor, L.L. (1996). "Aggregation and the Estimated Effects of School Resources", *NBER Working Papers*, 5548, National Bureau of Economic Research.

Hanushek, E.A., Kimko, D.D. (2000). "Schooling, Labor-Force Quality, and the Growth of Nations". *American Economic Review*, 90(5), pp. 1184-1208.

Hanushek, E.A., Woessmann, L. (2007). "The Role of Education Quality in Economic Growth". World Bank Policy Research Working Paper, 4122. Washington, D.C.

Hanushek, E.A., Woessmann, L. (2012). "Do Better School Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation". *Journal of Economic Growth*, 17(4), pp. 267-321.

Hanushek, E.A., Woessmann, L. (2015). *The Knowledge Capital of Nations: Education and the Economics of Growth*, MIT Press Books, The MIT Press, edition 1, volume 1, number 0262029170, July.

Heath, A., Kilpi-Jakonen, E. (2012). "Immigrant Children's Age at Arrival and Assessment Results", *OECD Education Working Papers*, n°75, OECD Publishing.

Heckman, J., Layne-Farrar, A., Todd, P., (1996). Human Capital Pricing Equations with an Application to Estimating the Effect of Schooling Quality on Earnings, *Review of Economics and Statistics*, 78(4), pp. 562-610.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187- 220). Westport, CT: American Council on Education and Praeger.

Holland, P. W., & Hoskens, M. (2002). Classical test theory as a first-order item response theory: application to true-score prediction from a possibly nonparallel test. *ETS Research Report Series*, *2002*(2).

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25,* 133-183.

Holland, P. W., & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions. *ETS Research Report Series*, *1987*(2).

Jerrim, J., Lopez-Agudo, L., Marcenaro-Gutierrez, O. D., & Shure, D. (2017).” What Happens When Econometrics and Psychometrics Collide? An Example Using the PISA Data.”

Keeves, J.P. (1992). *The IEA Science Study III: Changes in Science Education and Achievement: 1970 to 1984*. Oxford: Pergamon Press.

Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9,* 25-44.

Kolen, M.J., Brennan, R.L. (2014). *Test Equating, Scaling, and Linking. Methods and Practices*. Series Title: Statistics for Social and Behavioral Sciences, Springer-Verlag New York.

Kumar, S., Barakat, B., Goujon, A., Skirbekk, V., Sanderson, W., Lutz, W. (2010). "Projection of Populations by Level of Educational Attainment, Age and Sex for 120 Countries for 2005-2050". International Institute for Applied Systems Analysis, *Demographic Research*, 22, pp. 383-472.

Lavy, V. (2015). “Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries.” *Econ J, 125: F397–F424. doi:10.1111/ecoj.12233*

Lazear, E.A. (2003). "Teacher Incentives". *Swedish Economic Policy Review*, 10(3), pp. 179-214.

Lee, J.W., Barro, R.J (2001) "Schooling Quality in a Cross Section of Countries", *Economica*, 38(272), pp. 465-88.

Lee, D.-W., Lee, T.-H. (1995). "Human Capital and Economic Growth: Tests Based on the International Evaluation of Educational Achievement". *Economics Letters*, 47(2), pp. 219-225.

Leuven, E., Oosterbeek, H., van Ophen, H. (2004). "Explaining International Differences in Male Skill Wage Differentials by Differences in Demand and Supply of Skill". *Economic Journal*, 114(495), pp. 466-486.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education,* 6(1), 83- 102.

Lutz, W., Goujon, A., K.C., S., Sanderson, W. (2007). "Reconstruction of Populations by Age, Sex and Level of Educational Attainment for 120 Countries for 1970-2000". International Institute for Applied Systems Analysis, *Interim Report IR-07-002*, Austria.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. New York: National Bureau of Economic Research Press.

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects.* Princeton, NJ: ETS Policy Information Center.

Mulligan, C.B. (1999). "Galton versus the Human Capital Approach to Inheritance". *Journal of Political Economy*, 107(6), S184-S224.

Mullis, I.V.S., Martin, M.O., & Foy, P. (2009). *TIMSS 2007 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O., & Gonzalez, E. J. (2004). International achievement in the processes of reading comprehension: Results from PIRLS 2001 in 35 countries. *Chestnut Hill, MA: Boston College*.

Murname, R.J., Willet, J.B., Duhaldeborde, Y., Tyler, J.H. (2000). "How Important Are the Cognitive Skills of Teenagers in Predicting Subsequent Earnings?". *Journal of Policy Analysis and Management*, 19(4), pp. 547-568.

Nelson R. and Phelps E., (1966). Investment in Humans, Technological Diffusion and Economic Growth. American Economic Review, 61, 69-75.

OECD (2010). *PISA 2009 Results: Overcoming Social Background: Equity in Learning Opportunities and Outcomes, (Volume II)*, PISA, OECD Publishing.

OECD (2011a), *Report on the Gender Initiative: Gender Equality in Education, Employment and Entrepreneurship*, OECD Publishing.

OECD (2013a). "Do immigrant students' reading skills depend on how long they have been in their new country?", *PISA in Focus*, n°29, June, Paris.

OECD (2013b). "What makes urban schools different?", *PISA in Focus*, n°28, Paris.

Pekkarinen, Tuomas (2012), "Gender Differences in Education", IZA Discussion Paper 6390.

Pritchett, L. (2001). "Where has all the education gone?". *World Bank Economic Review*, 15(3), pp. 367-391.

Psacharopoulos, G., Patrinos, H. (2004). "Returns to Investment in Education: A Further Update". *Education Economics*, 12(2), pp. 111-134.

Purves, A., & Elley, W. B. (1994). The role of the home and student differences in reading performance. *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*, 89-121.

Rindermann, H., & Ceci, S. J. (2009). Educational policy and country outcomes in international cognitive competence studies. *Perspectives on Psychological Science*, *4*(6), 551-568.

Robitaille, D.F., Garden, R.A. (Eds.) (1989). *The IEA Study of Mathematics II: Context and Outcomes of School Mathematics*. Oxford: Pergamon Press.

Ross, K.N., Postlethwaite, T.N. (1991). *Indicators of the Quality of Education: A Study of Zimbabwean Primary Schools*. Harare: Ministry of Education and Culture; Paris: International Institute for Educational Planning.

Saito, M., van Capelle, F. (2009). "Approaches to Monitoring the Quality of Education in Developing Countries – Searching for Better Research-Policy Linkages", Paper based on the

presentation during The International Symposium on Quality Education for All – Approaches to Monitoring and Improving the Quality of Education, Berlin, 11-12 May 2009.

Sakellariou, C. (2006). "Cognitive Ability and Returns to Schooling in Chile". Background Paper for Vegas, E., Petrow, J. (Eds.) (2008), *Raising Student Learning in Latin America. The Challenge for the 21st Century*. World Bank. Latin American Development Forum series.

Sandefur, J. (2016). Internationally Comparable Mathematics Scores for Fourteen African Countries. *CGD Working Paper.*

UNESCO (2000). *With Africa for Africa. Toward quality education for all*. UNESCO. Paris: Human Sciences Research Council.

UNESCO (2004). *EFA Global Monitoring Report 2005: The Quality Imperative*. Paris.

UNESCO. (2015). *Education for All 2000-2015: Achievements and Challenges*. Paris, UNESCO Publishing.

UNESCO (2017). *Exploring Commonalities and Differences in Regional and International Assessments.* UIS Information Paper. Paris.

Vandenbussche, J., Aghion, P., Meghir, C. (2006), "Growth, distance to frontier and composition of human capital", *Journal of Economic Growth*, 11(2), pp.97-127.

Vegas, E., Petrow, J. (2008). *Raising Student Learning in Latin America. The Challenge for the 21st Century*. World Bank, Latin American Development Forum series.

Von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, *28*(6), 389-406.

Von Davier, A. A., Holland, P. W., & Thayer, D. T. (2003). *The kernel method of test equating*. Springer Science & Business Media.

Wolf, A. (2002), *Myths about Education and Growth*. London, Penguin Books.

World Bank (1999), *Education Sector Strategy*, Human Development Network Series, World Bank, Washington, DC, 1999.

World Bank (2012), *World Development Report 2012: Gender Equality and Development*, Washington

World Bank (2017), *World Development Report 2018*: *Learning to Realize Education's Promise,* Washington.

Wu, M. (2010), "Comparing the Similarities and Differences of PISA 2003 and TIMSS", *OECD Education Working Papers*, No. 32, OECD Publishing

**Table 7:** Robustness check: Results for anchored value of USA mean score with alternative Sub-Samples of doubloon countries

| | Pseudo-linear linking | Pre-smoothed equipercentile linking |
|---|---|---|
| **Anchoring between PISA 2003 and TIMSS 2003 assessments** | | |
| 1. All doubloon countries | 500.28 | 481.10 |
| 2. Only first panel of doubloon countries | 510.53 | 481.10 |
| 3. Only second panel of doubloon countries | 488.84 | 481.10 |
| **Anchoring between PISA 2015 and TIMSS 2015 assessments** | | |
| 2. All doubloon countries | 497.40 | 501.67 |
| 3. Only first panel of doubloon countries | 498.52 | 499.60 |
| 4. Only second panel of doubloon countries | 491.66 | 506.00 |

Note: Results are based on mathematics for secondary level by comparing the anchored results of PISA 2003 and 2015 achievement scores for the USA using different samples of countries in the anchoring process. Two linking methods are presented: pseudo-linear and pre-smoothed equipercentile linking. See text for more information about these linking techniques.

**Table 8.0: Point Estimate Difference between and ISAT/RSAT and HLO (1995-2015)**

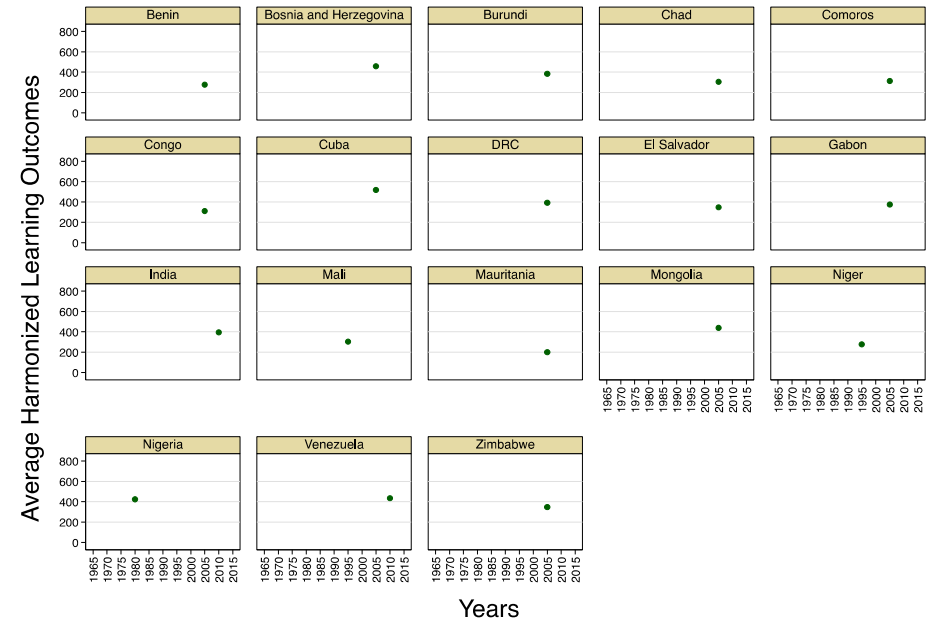
| Secondary | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2015 | 2010 | 2005 | 2000 | 1995 | Average |
| PISA | Math | -21.76 | -17.83 | -17.14 | -23.05 | | -19.94 |
| | Reading | 0.09 | 1.44 | -1.12 | -0.37 | | 0.01 |
| | Science | -27.53 | -23.64 | -21.39 | -23.94 | | -24.13 |
| TIMSS | Math | -1.59 | 1.81 | -1.159463 | 0.51 | -1.92 | -0.30 |
| | Reading | . | . | . | . | . | |
| | Science | -2.31 | 0.33 | -1.68 | 0.04 | -0.63 | -0.85 |
| Primary | | | | | | | |
| TIMSS | Math | -2.05 | 1.85 | -3.97 | . | 5.57 | 0.35 |
| | Reading | . | . | . | . | . | |
| | Science | -1.74 | 1.21 | -4.02 | . | 4.06 | -0.12 |
| SACMEQ | Math | . | . | 169.69 | 165.66 | . | 167.67 |
| | Reading | . | . | 201.02 | 194.57 | 194.63 | 196.74 |
| | Science | . | . | . | . | . | |
| LLECE | Math | 102.54 | . | 95.56 | . | . | 99.05 |
| | Reading | 54.21 | . | 51.87 | . | . | 53.04 |
| | Science | 50.68 | . | 55.05 | . | . | 52.86 |

**Table 8.1: Point Estimate Difference between and ISAT/RSAT and HLO (1995-2015)**

| Secondary | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2015 | 2010 | 2005 | 2000 | 1995 | Average |
| PISA | Reading | 0.00 | 0.00 | 0.00 | 0.00 | | 0.00 |
| | | (0.6) | (4.2) | (1.2) | (1.3) | | (1.8) |
| TIMSS | Math | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | | (1.2) | (1.3) | (1.2) | (0.3) | (2.8) | (1.4) |
| Primary | | | | | | | |
| TIMSS | Math | 0.00 | 0.00 | 0.00 | | 0.00 | 0.00 |
| | | (3.6) | (0.9) | (1.1) | | (4.2) | (2.5) |
| SACMEQ | Math | | | 0.00 | 0.00 | | 0.00 |
| | | | | (0.4) | (0.0) | | (0.2) |
| LLECE | Math | 0.00 | | 0.00 | | | 0.00 |
| | | (0.6) | | (1.5) | | | (1.0) |
| Average Rank Change | | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Average SD | | (1.5) | (2.1) | (1.1) | (0.5) | (3.5) | (1.4) |

**HLO data availability by Available Scores Averaged across Subjects and Level and across 5-year Intervals**

**Figure 12.0 – One Score Available from 1965-2015, 18 Countries**



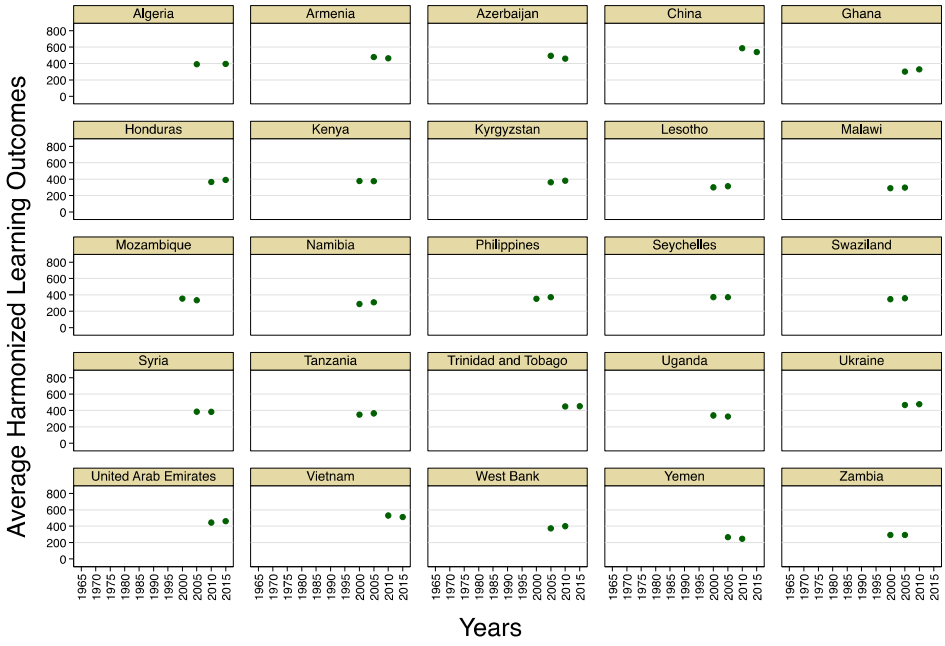



**Figure 12.2 – Three Scores from 1965-2015, 30 Countries**

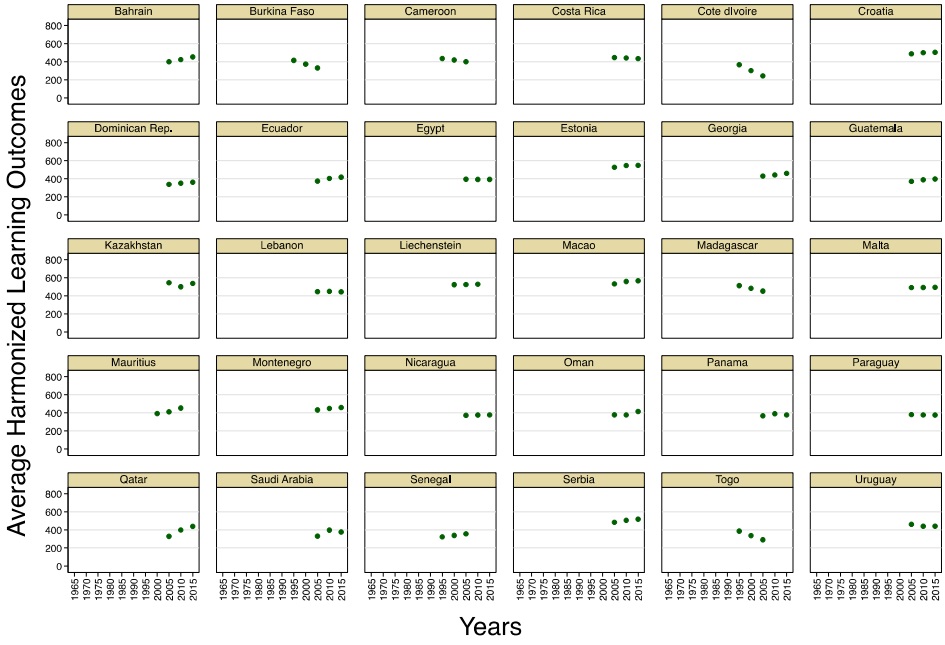



**Figure 12.3 – Four Scores from 1965-2015, 17 Countries**

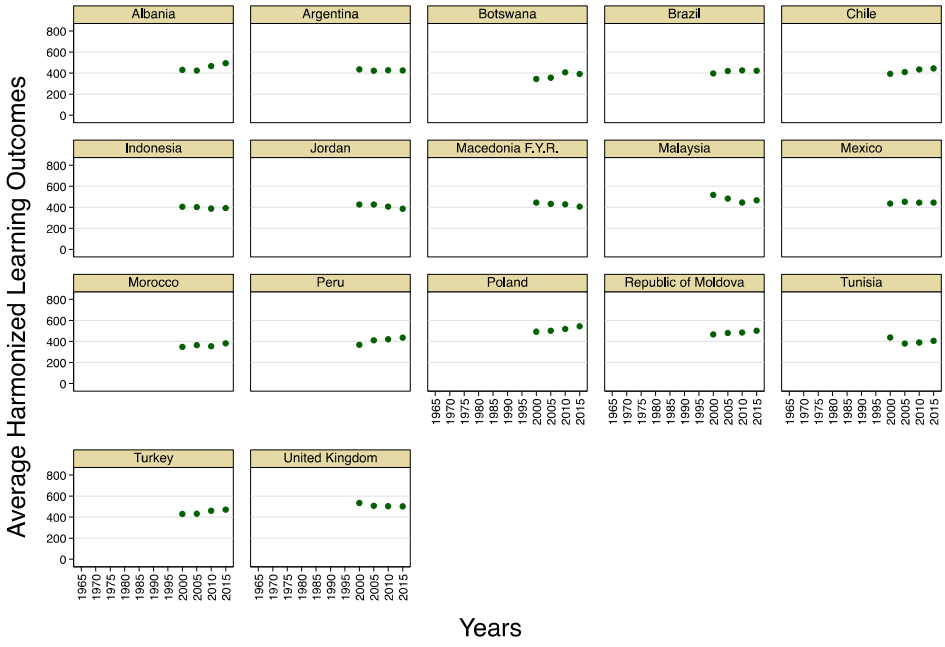


**Figure 12.4 – Five Scores from 1965-2015, 27 Countries**

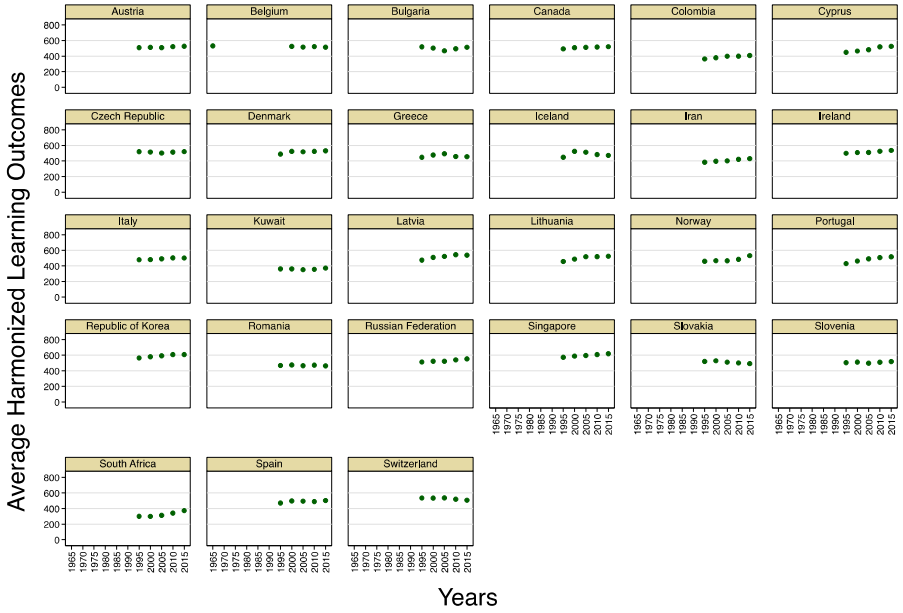



**Figure 12.5 – Eight Scores from 1965-2015, 6 Countries**

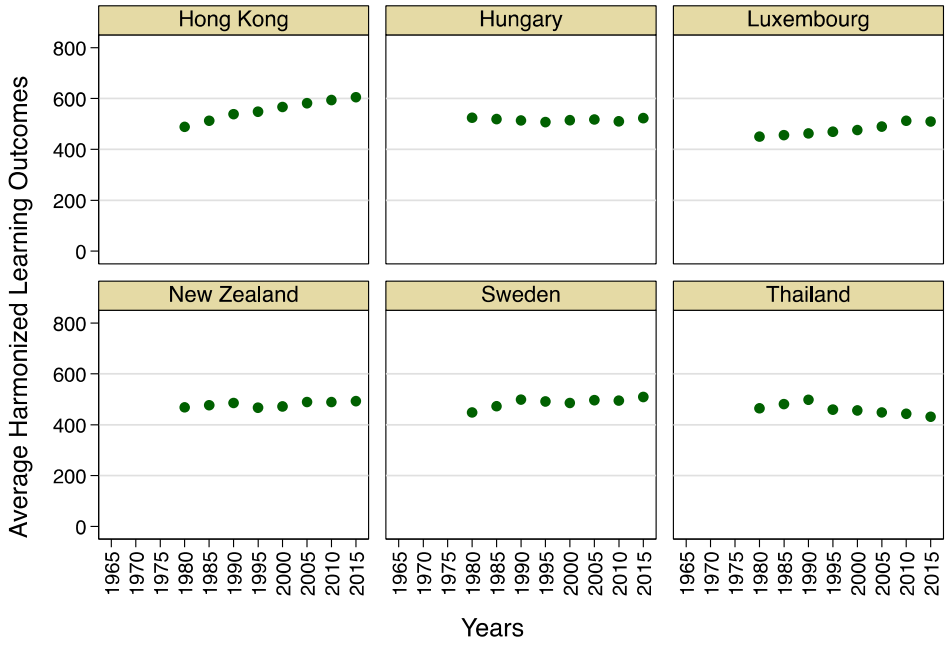


**Figure 12.6 – Eleven Scores from 1965-2015, 8 Countries**

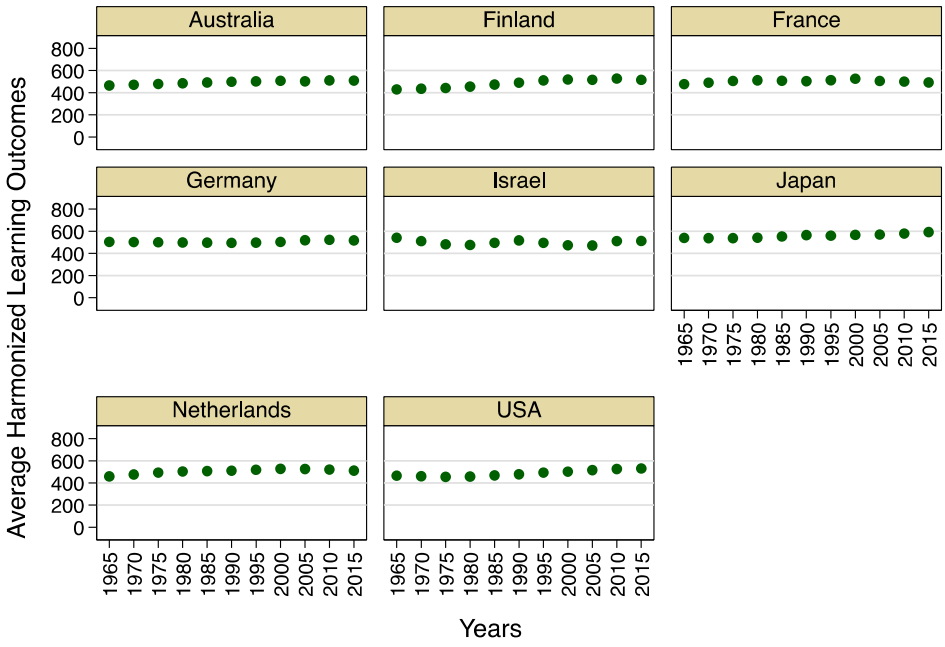
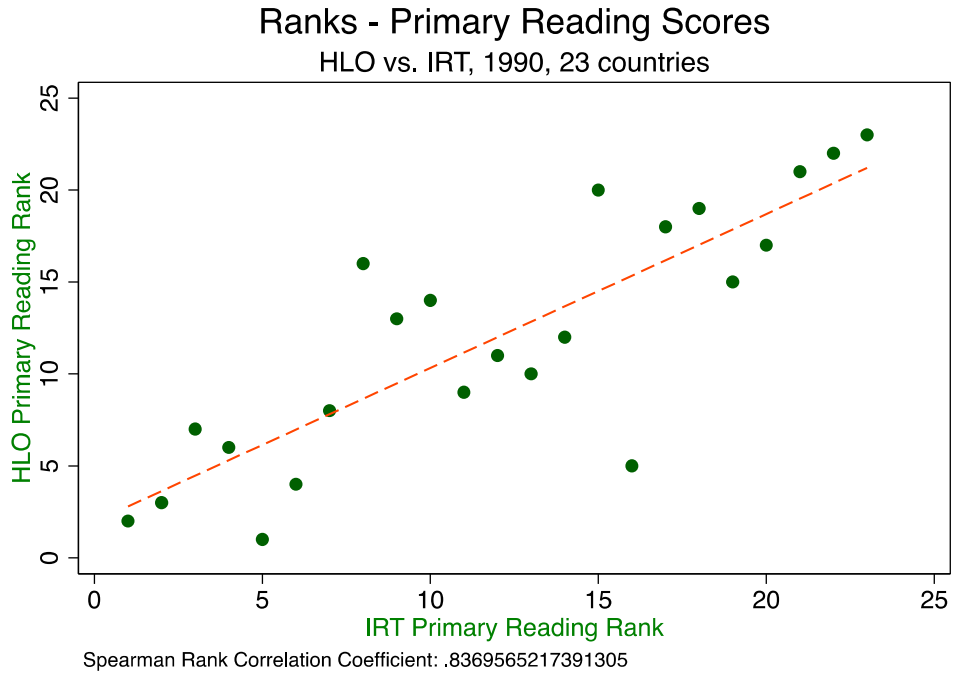
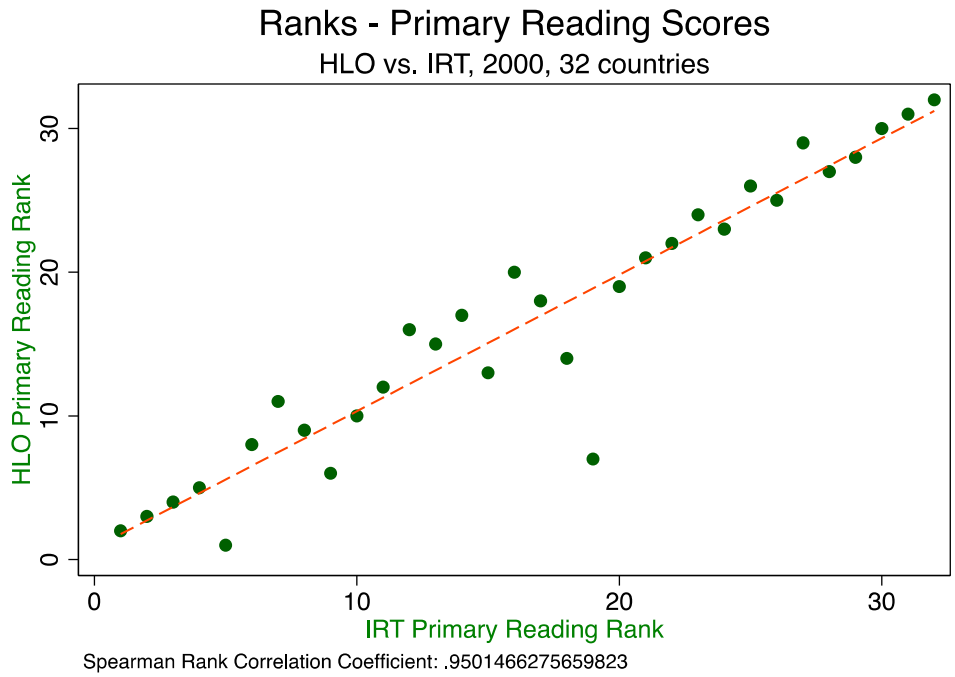
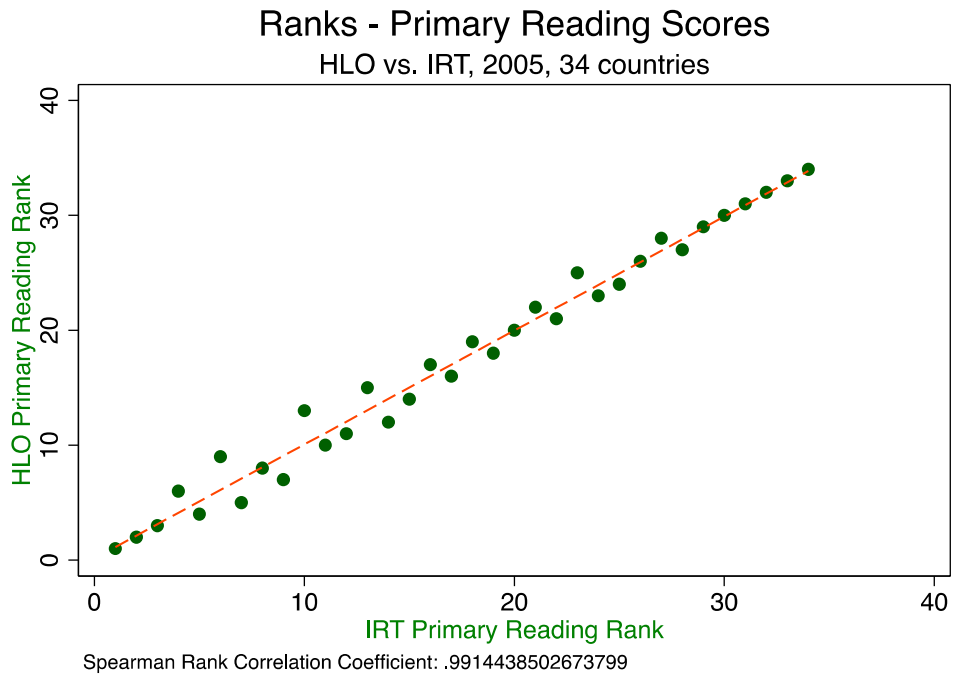


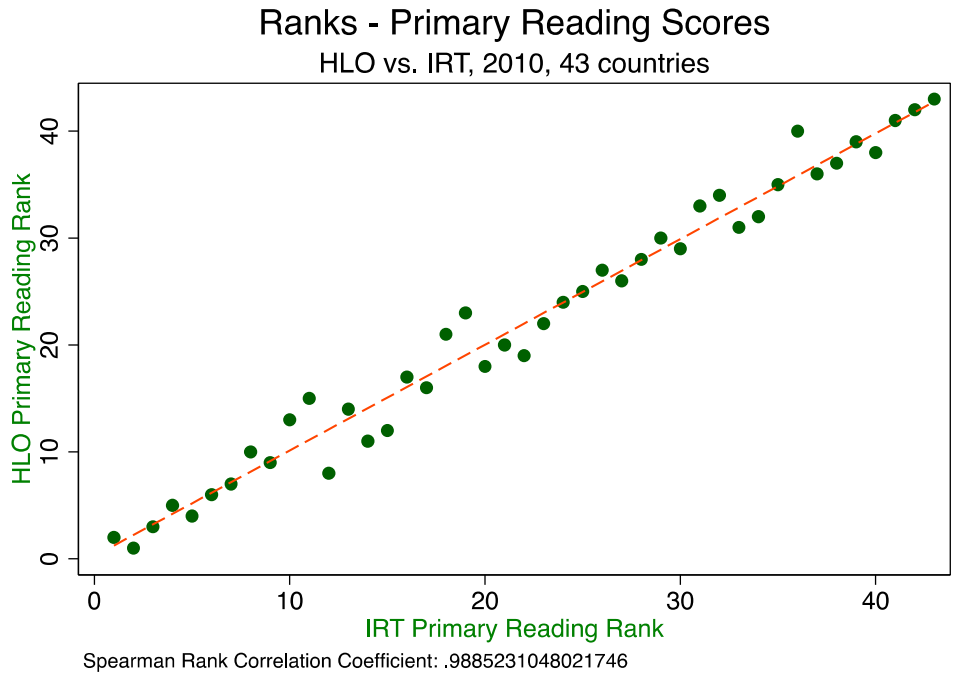
**Figure 13.0-13.3 – Robustness Test – IRT vs. HLO for Primary Reading Ranks**






Spearman Rank Correlation Coefficient: .9914438502673799

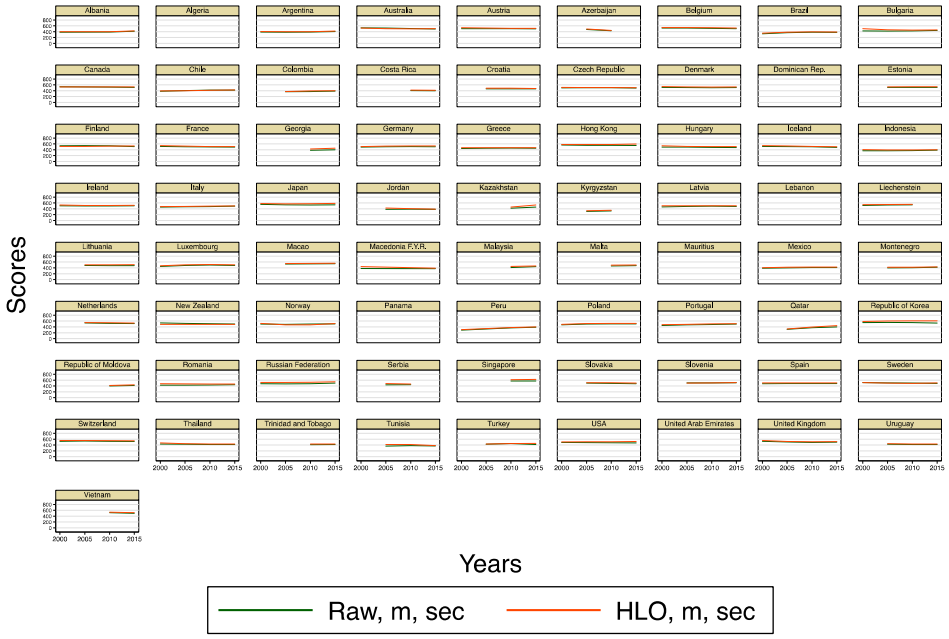



Spearman Rank Correlation Coefficient: .9885231048021746

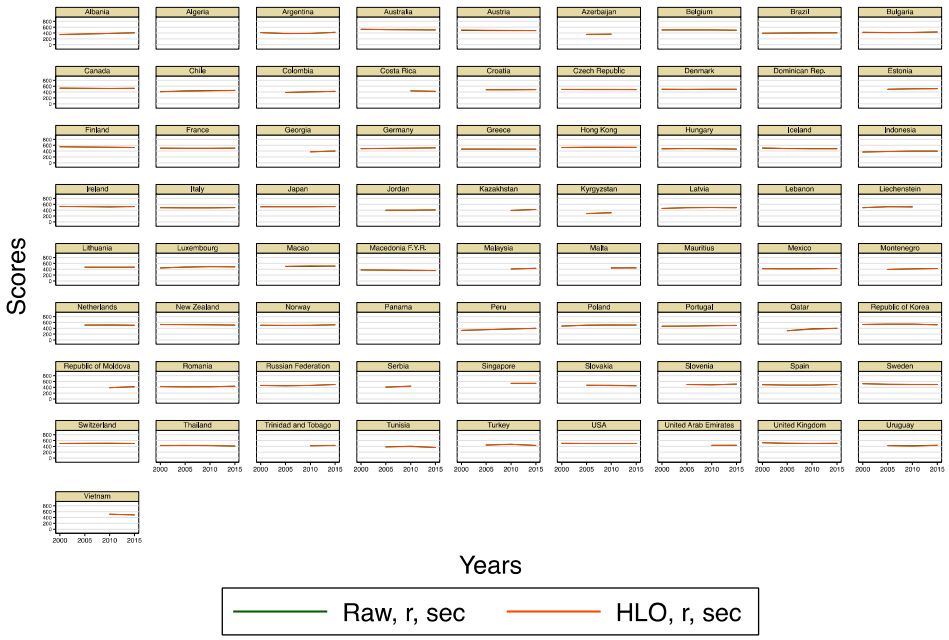
## Figures 14.0-14.9: Comparison to Raw ISAT and RSAT scores

*Note: where country cells are empty there is only one data point so a comparison over time is not possible*

### Figure 14.0: Raw PISA vs. HLO Secondary Math Scores

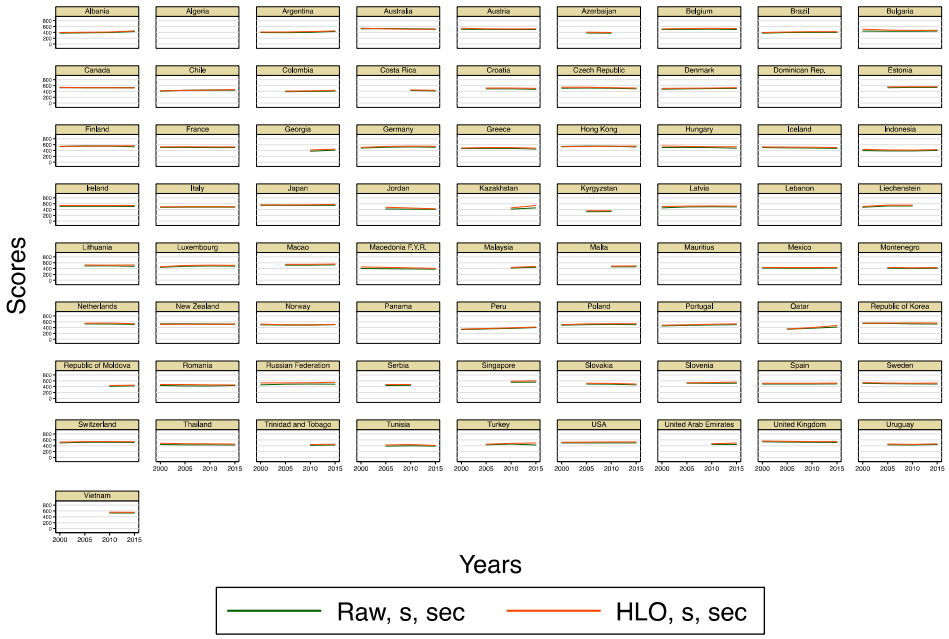



### Figure 14.1: Raw PISA vs. HLO Secondary Reading Scores

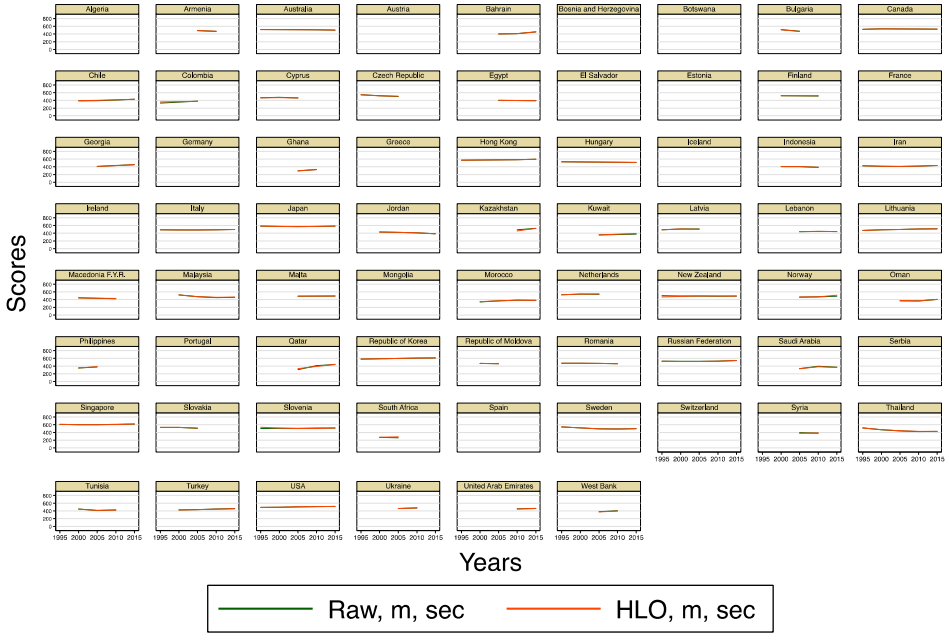



G h b

## Figure 14.2: Raw PISA vs. HLO Secondary Science Scores




## Figure 14.3: Raw TIMSS vs. HLO Secondary Math Scores



## Figure 14.4: Raw TIMSS vs. HLO Secondary Science Scores

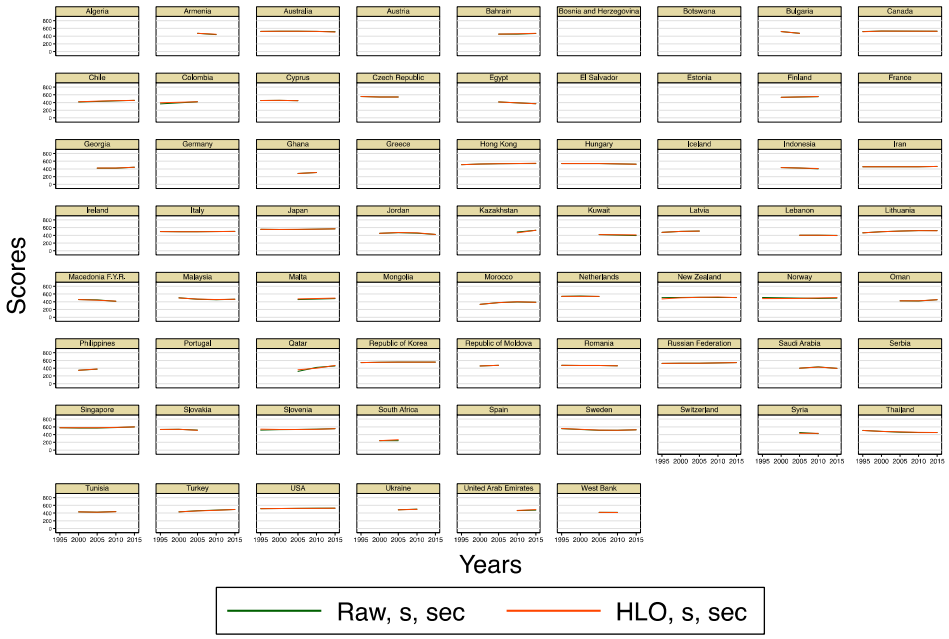



**Table 8.0: Point Estimate Difference between and ISAT/RSAT and HLO (1995-2015)**

## 4.5: Raw TIMSS vs. HLO Primary Math Scores

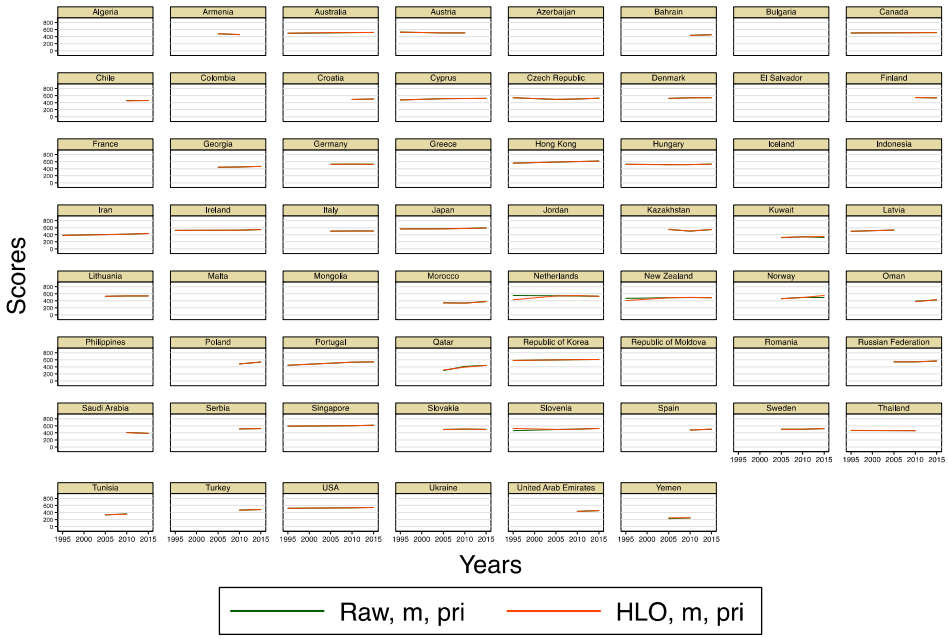


**Figure 14.6: Raw SACMEQ vs. HLO Primary Math Scores**

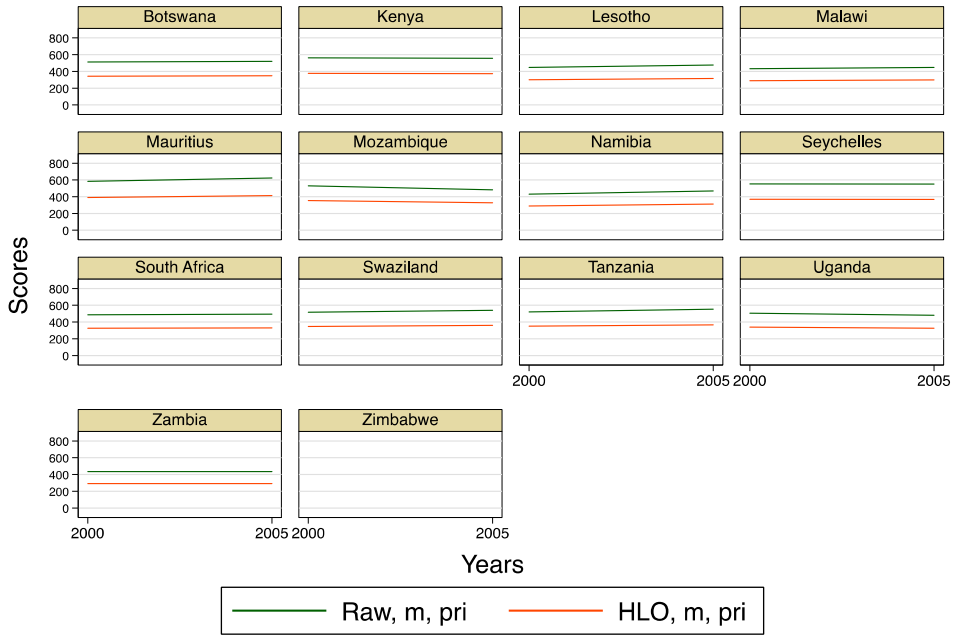



**Figure 14.7: Raw SACMEQ vs. HLO Primary Reading Scores**

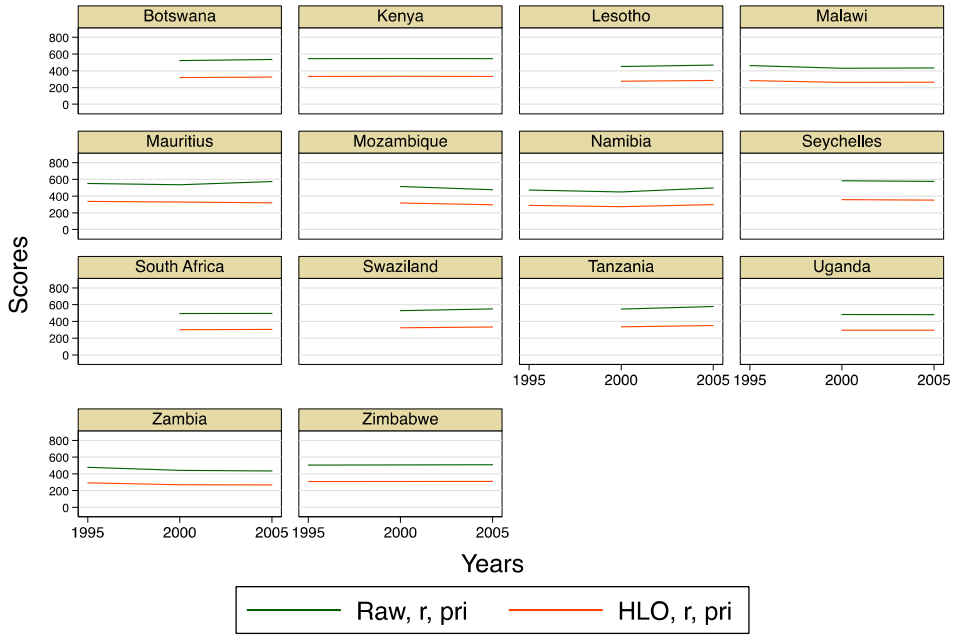


**Figure 14.8: Raw LLECE vs. HLO Math Reading Scores**

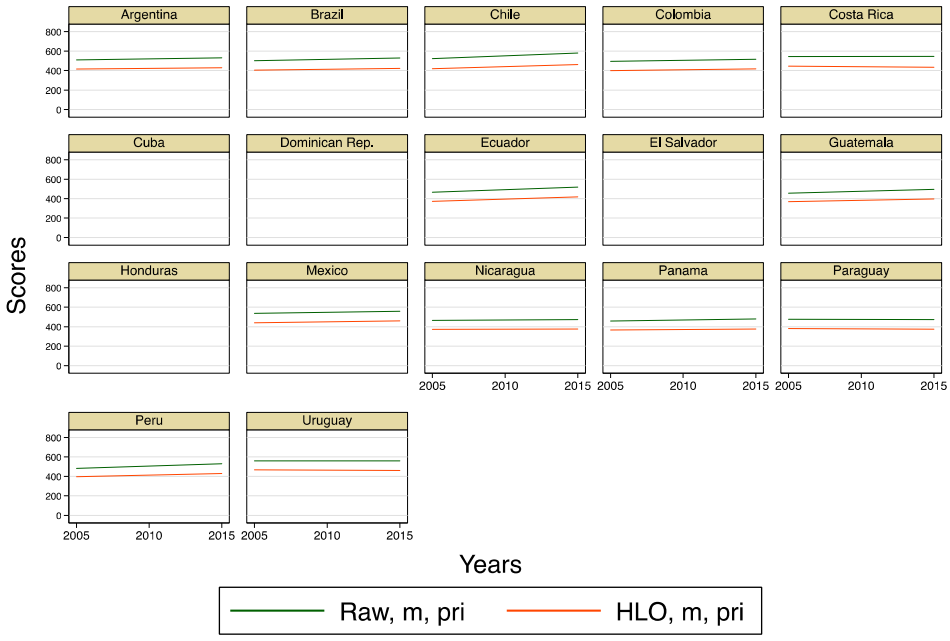
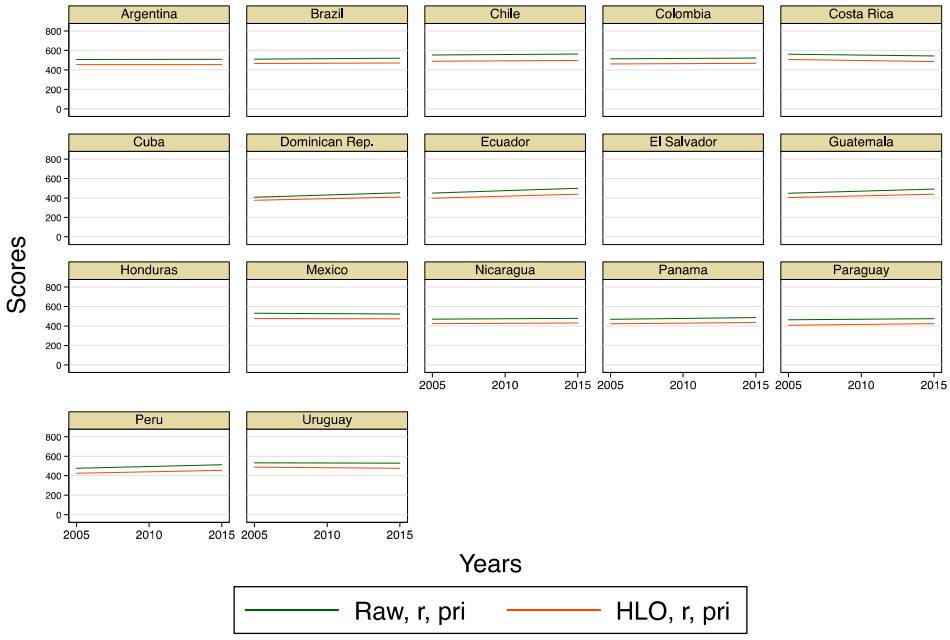
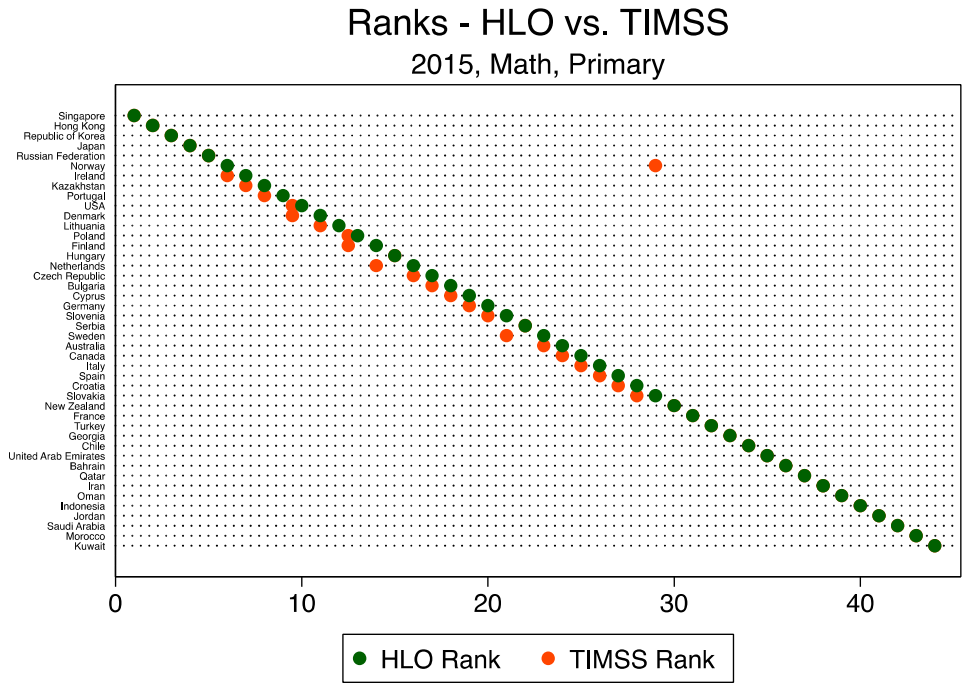
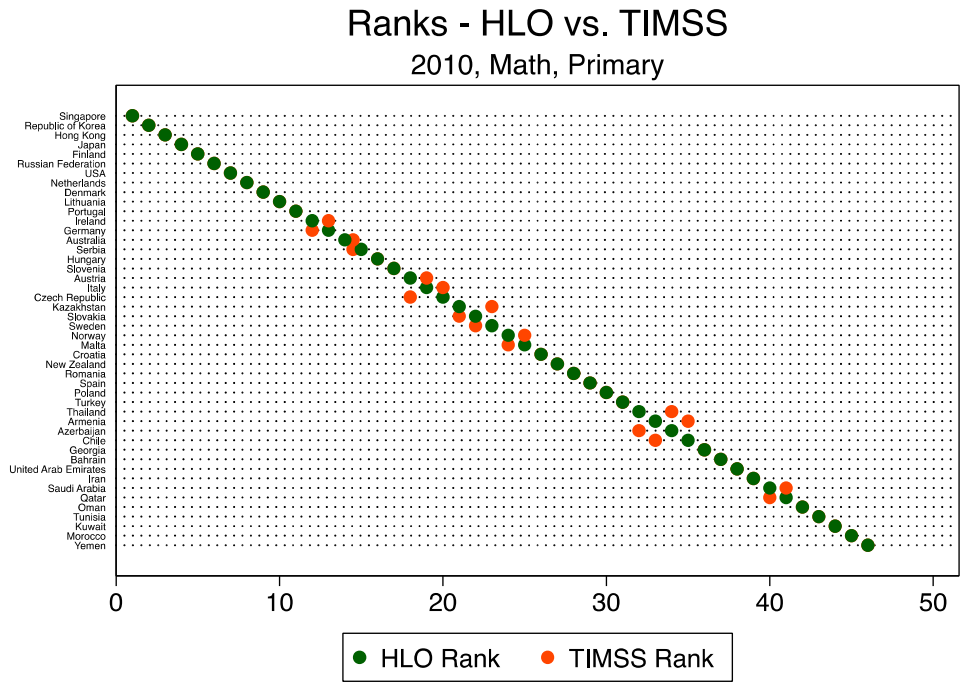


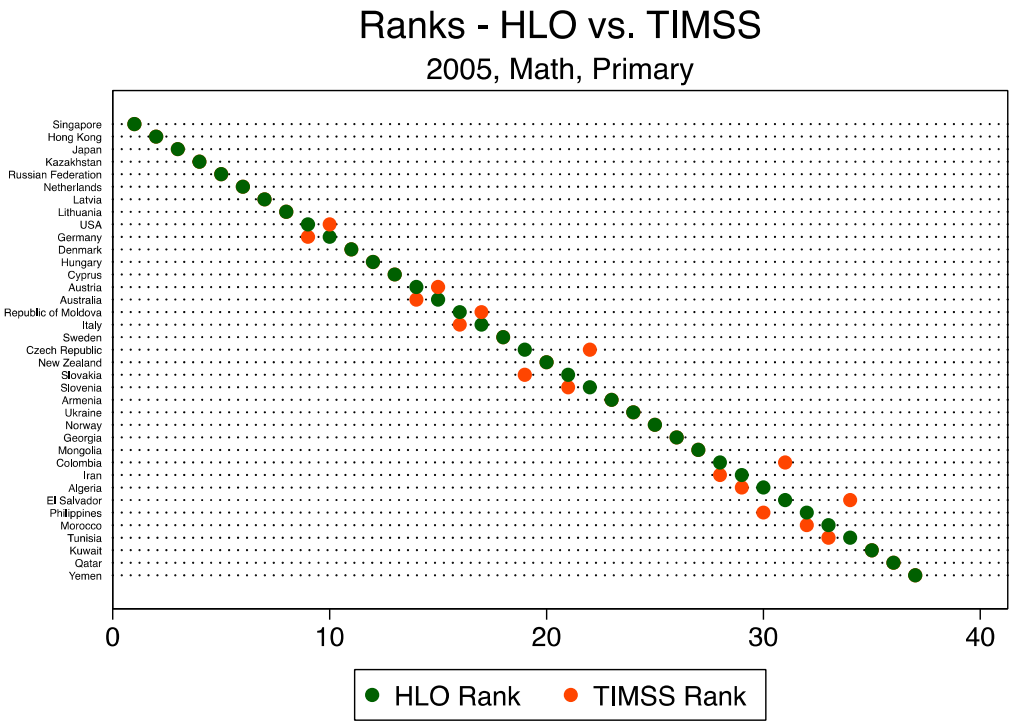

**Figure 14.9: Raw LLECE vs. HLO Reading Scores**

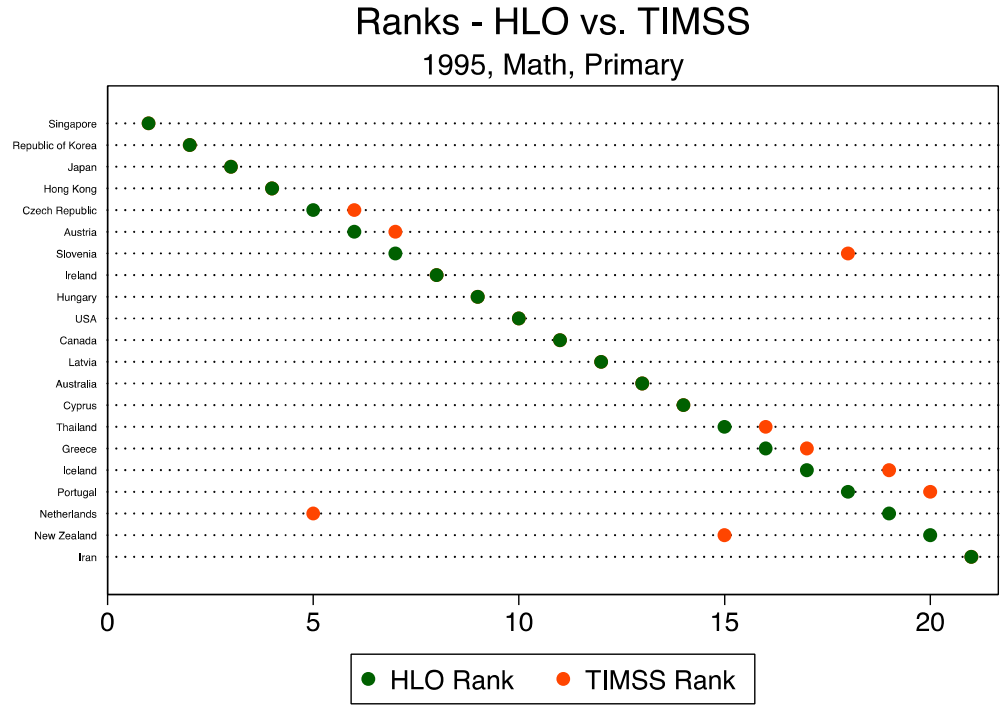


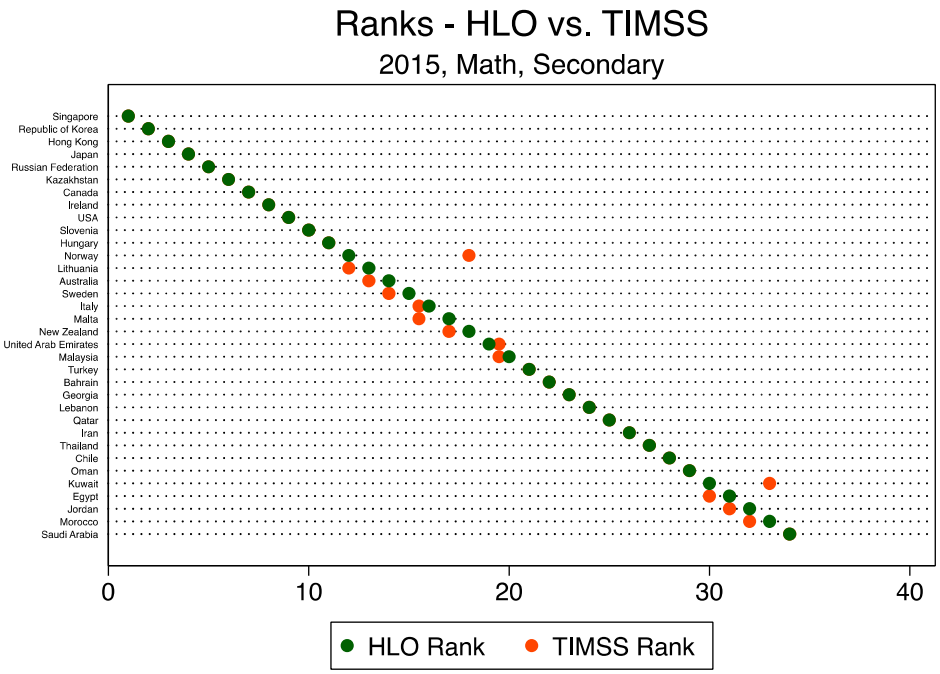
## Figures 15.0-15.16: Rank Comparison of ISAT and RSAT scores

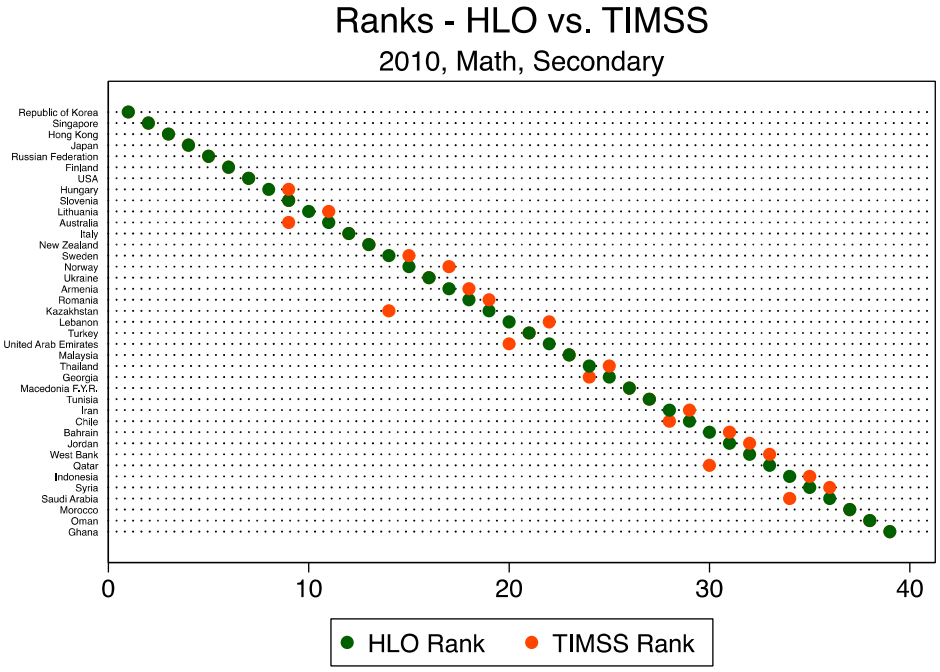


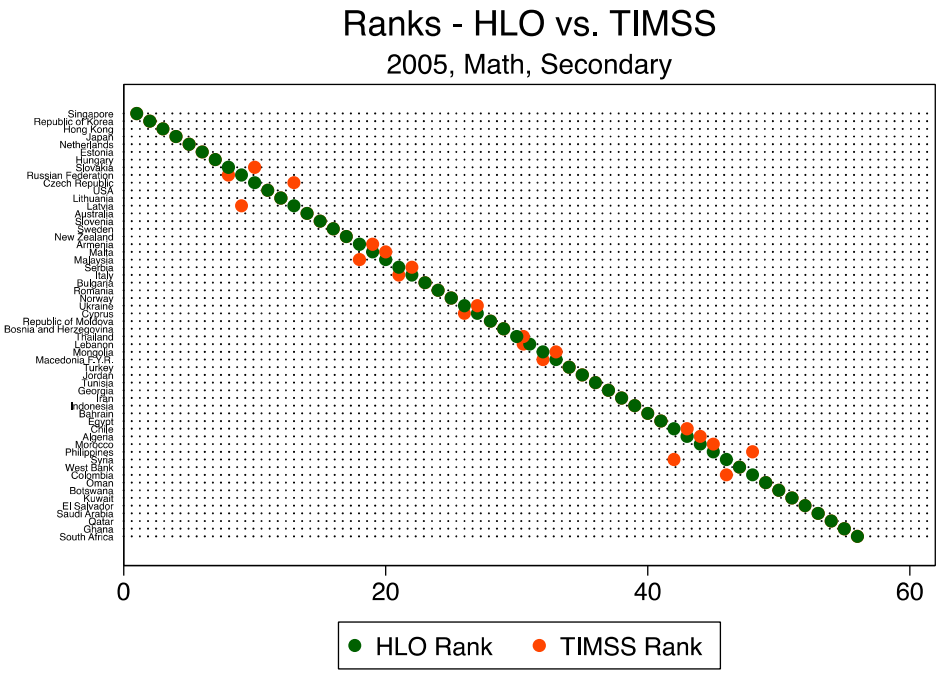



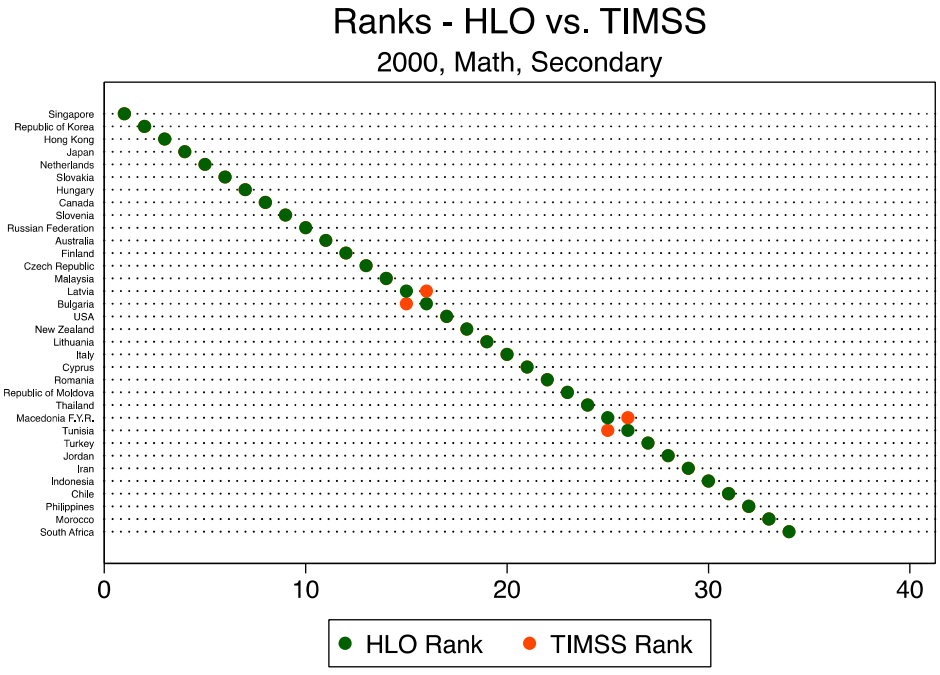

Ranks - HLO vs. TIMSS
2005, Math, Primary
Singapore
Hong Kong
Japan
Kazakhstan
Russian Federation
Netherlands
Latvia
Lithuania
USA
Germany
Denmark
Hungary
Cyprus
Austria
Australia
Republic of Moldova
Italy
Sweden
Czech Republic
New Zealand
Slovakia
Slovenia
Armenia
Ukraine
Norway
Georgia
Mongolia
Colombia
Iran
Algeria
El Salvador
Philippines
Morocco
Tunisia
Kuwait
Qatar
Yemen
0
10
20
30
40
HLO Rank
TIMSS Rank

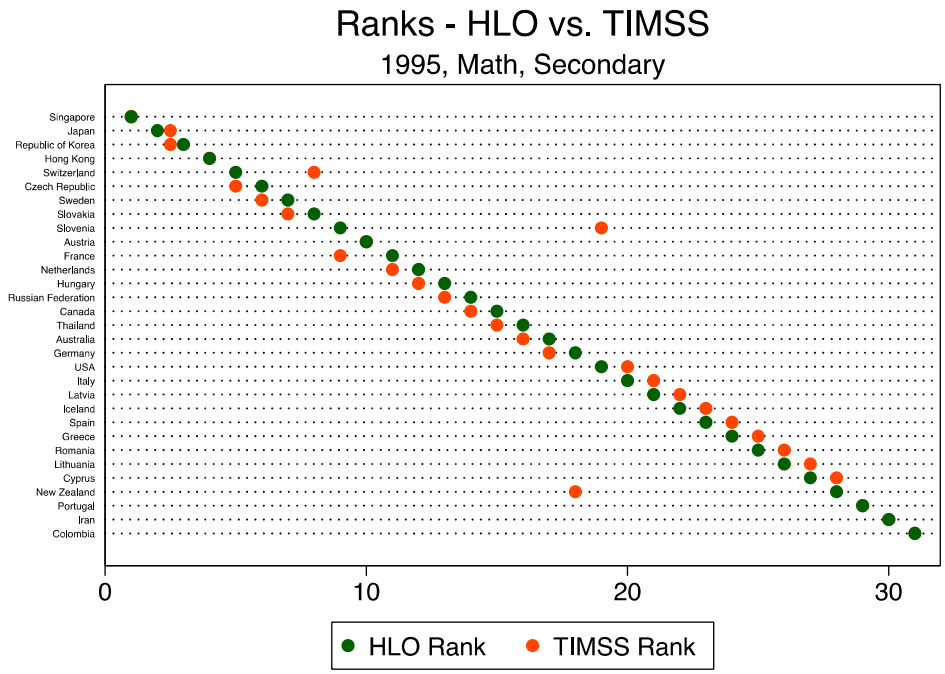

Ranks - HLO vs. TIMSS
1995, Math, Primary
Singapore
Republic of Korea
Japan
Hong Kong
Czech Republic
Austria
Slovenia
Ireland
Hungary
USA
Canada
Latvia
Australia
Cyprus
Thailand
Greece
Iceland
Portugal
Netherlands
New Zealand
Iran
0
5
10
15
20
HLO Rank
TIMSS Rank

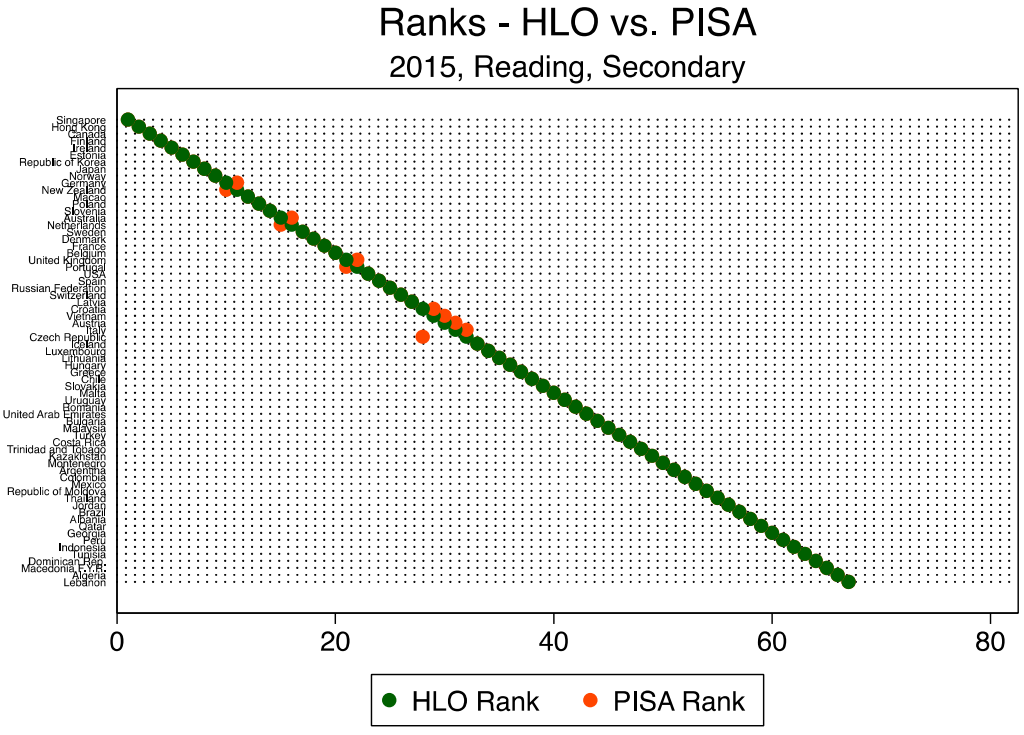

Ranks - HLO vs. TIMSS
2015, Math, Secondary
Singapore
Republic of Korea
Hong Kong
Japan
Russian Federation
Kazakhstan
Canada
Ireland
USA
Slovenia
Hungary
Norway
Lithuania
Australia
Sweden
Italy
Malta
New Zealand
United Arab Emirates
Malaysia
Turkey
Bahrain
Georgia
Lebanon
Qatar
Iran
Thailand
Chile
Oman
Kuwait
Egypt
Jordan
Morocco
Saudi Arabia
0
10
20
30
40
HLO Rank
TIMSS Rank

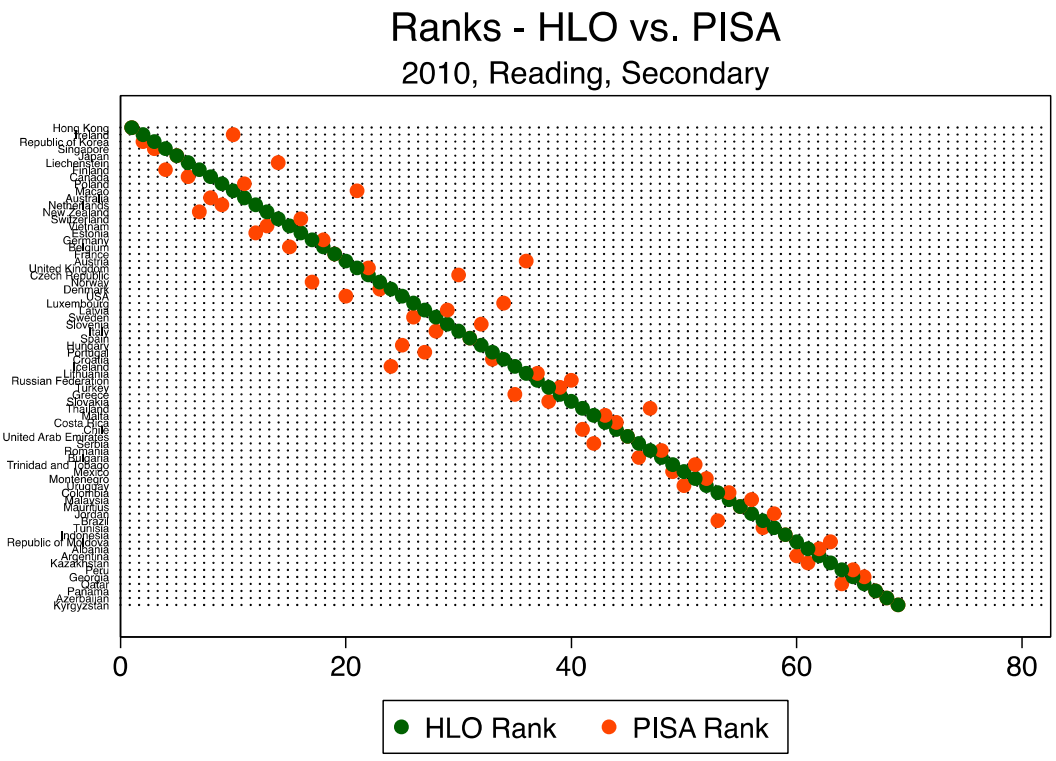

Ranks - HLO vs. TIMSS
2010, Math, Secondary
Republic of Korea
Singapore
Hong Kong
Japan
Russian Federation
Finland
USA
Hungary
Slovenia
Lithuania
Australia
Italy
New Zealand
Sweden
Norway
Ukraine
Armenia
Romania
Kazakhstan
Lebanon
Turkey
United Arab Emirates
Malaysia
Thailand
Georgia
Macedonia, F.Y.R.
Tunisia
Iran
Chile
Bahrain
Jordan
West Bank
Qatar
Indonesia
Syria
Saudi Arabia
Morocco
Oman
Ghana
0
10
20
30
40
HLO Rank
TIMSS Rank

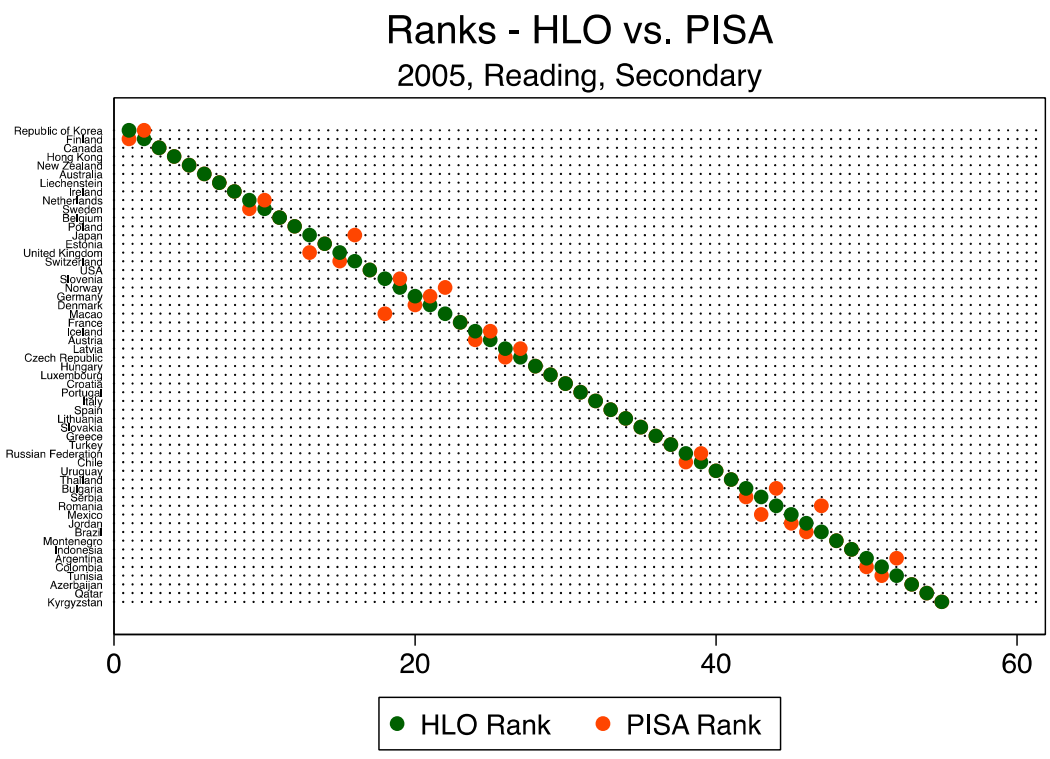

Ranks - HLO vs. TIMSS
2005, Math, Secondary
Singapore
Republic of Korea
Hong Kong
Japan
Netherlands
Estonia
Hungary
Slovakia
Russian Federation
Czech Republic
USA
Lithuania
Latvia
Australia
Slovenia
Sweden
New Zealand
Armenia
Malta
Malaysia
Serbia
Italy
Bulgaria
Romania
Norway
Ukraine
Cyprus
Republic of Moldova
Bosnia and Herzegovina
Thailand
Lebanon
Mongolia
Macedonia F.Y.R.
Turkey
Jordan
Tunisia
Georgia
Iran
Indonesia
Bahrain
Egypt
Chile
Algeria
Morocco
Philippines
Syria
West Bank
Colombia
Oman
Botswana
Kuwait
El Salvador
Saudi Arabia
Qatar
Ghana
South Africa
0
20
40
60
HLO Rank
TIMSS Rank

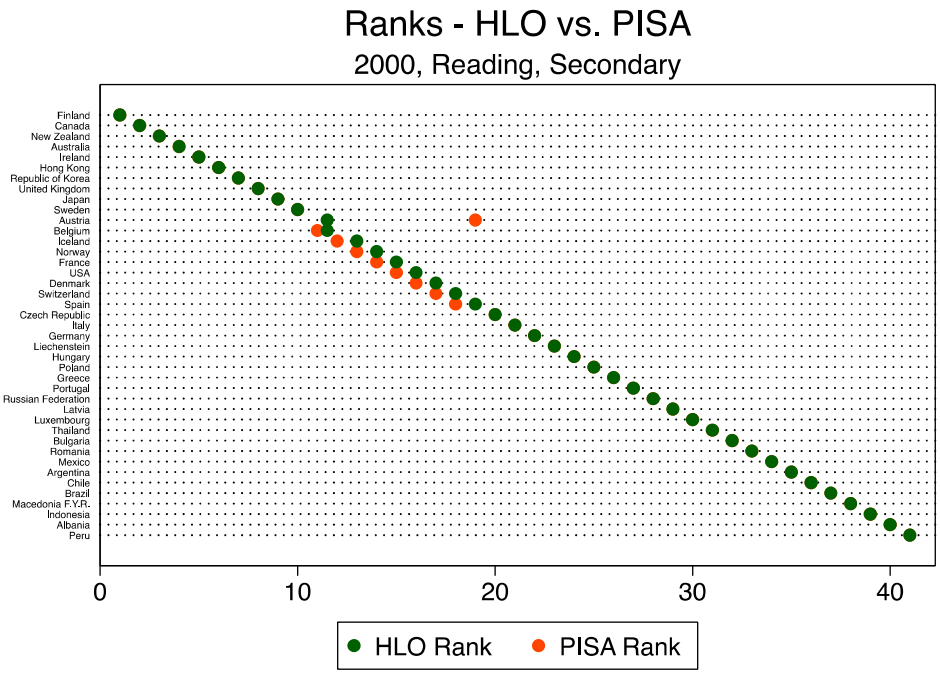

Ranks - HLO vs. TIMSS
2000, Math, Secondary
Singapore
Republic of Korea
Hong Kong
Japan
Netherlands
Slovakia
Hungary
Canada
Slovenia
Russian Federation
Australia
Finland
Czech Republic
Malaysia
Latvia
Bulgaria
USA
New Zealand
Lithuania
Italy
Cyprus
Romania
Republic of Moldova
Thailand
Macedonia F.Y.R.
Tunisia
Turkey
Jordan
Iran
Indonesia
Chile
Philippines
Morocco
South Africa
0
10
20
30
40
HLO Rank
TIMSS Rank

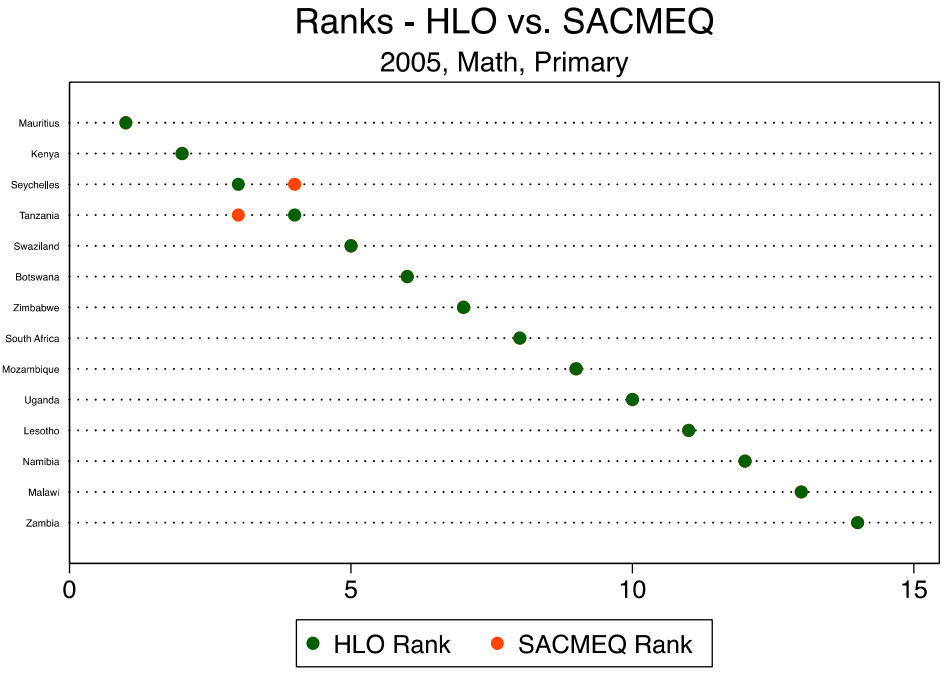

Ranks - HLO vs. TIMSS
1995, Math, Secondary
Singapore
Japan
Republic of Korea
Hong Kong
Switzerland
Czech Republic
Sweden
Slovakia
Slovenia
Austria
France
Netherlands
Hungary
Russian Federation
Canada
Thailand
Australia
Germany
USA
Italy
Latvia
Iceland
Spain
Greece
Romania
Lithuania
Cyprus
New Zealand
Portugal
Iran
Colombia
0
10
20
30
HLO Rank
TIMSS Rank

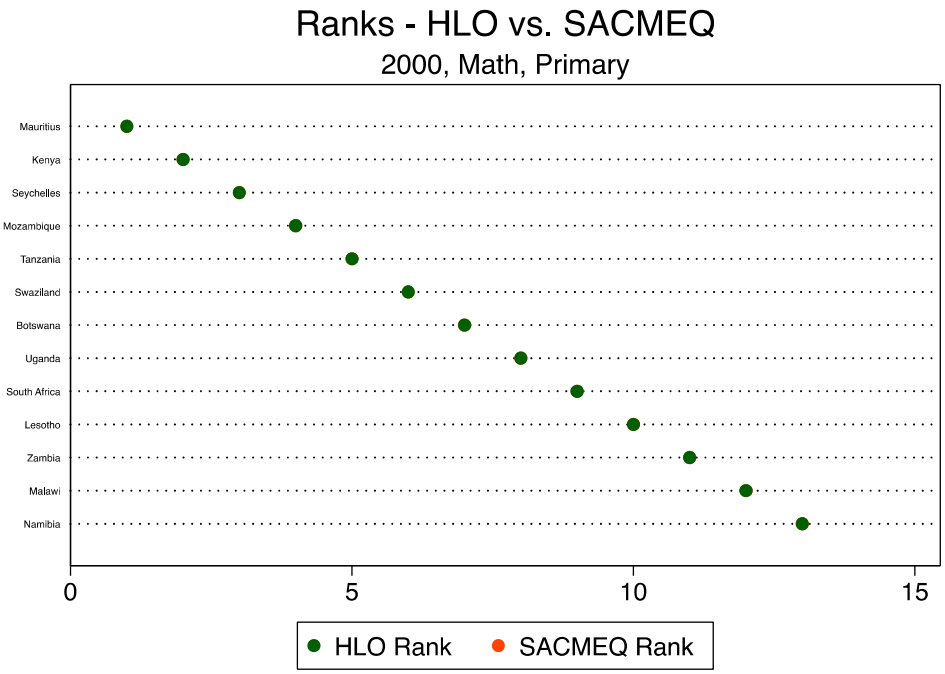

Ranks - HLO vs. PISA
2015, Reading, Secondary
0
20
40
60
80
HLO Rank
PISA Rank

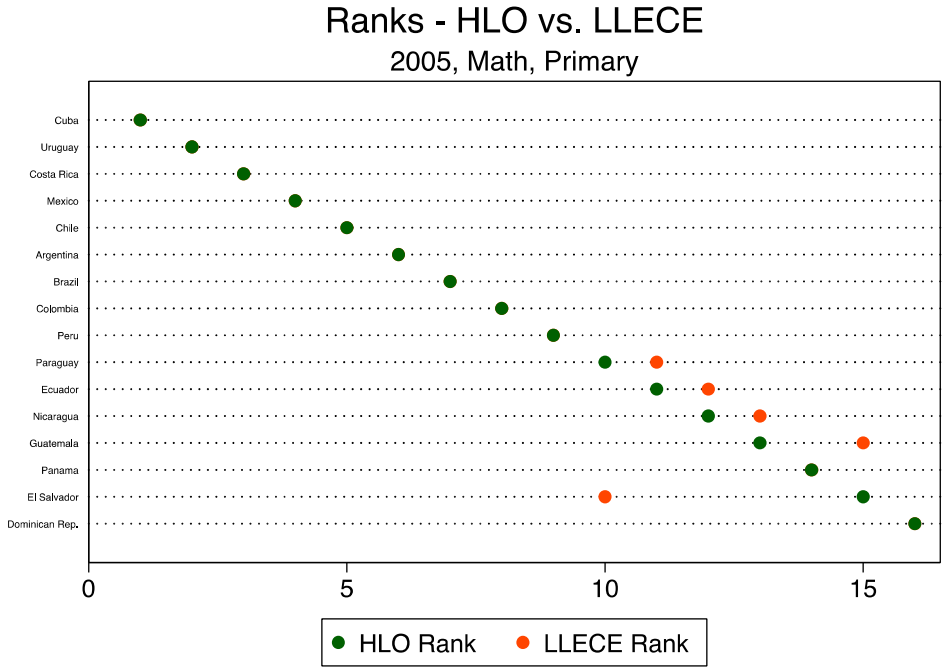

Ranks - HLO vs. PISA
2010, Reading, Secondary
0
20
40
60
80
HLO Rank
PISA Rank


Ranks - HLO vs. PISA
2005, Reading, Secondary
Republic of Korea
Finland
Canada
Hong Kong
New Zealand
Australia
Liechtenstein
Ireland
Netherlands
Sweden
Belgium
Poland
Japan
Estonia
United Kingdom
Switzerland
USA
Slovenia
Norway
Germany
Denmark
Macao
France
Iceland
Austria
Latvia
Czech Republic
Hungary
Luxembourg
Croatia
Portugal
Italy
Spain
Lithuania
Slovakia
Greece
Turkey
Russian Federation
Chile
Uruguay
Thailand
Bulgaria
Serbia
Romania
Mexico
Jordan
Brazil
Montenegro
Indonesia
Argentina
Colombia
Tunisia
Azerbaijan
Qatar
Kyrgyzstan
0
20
40
60
HLO Rank
PISA Rank


Ranks - HLO vs. PISA
2000, Reading, Secondary
Finland
Canada
New Zealand
Australia
Ireland
Hong Kong
Republic of Korea
United Kingdom
Japan
Sweden
Austria
Belgium
Iceland
Norway
France
USA
Denmark
Switzerland
Spain
Czech Republic
Italy
Germany
Liechenstein
Hungary
Poland
Greece
Portugal
Russian Federation
Latvia
Luxembourg
Thailand
Bulgaria
Romania
Mexico
Argentina
Chile
Brazil
Macedonia F.Y.R.
Indonesia
Albania
Peru
0
10
20
30
40
HLO Rank
PISA Rank


Ranks - HLO vs. SACMEQ
2005, Math, Primary
Mauritius
Kenya
Seychelles
Tanzania
Swaziland
Botswana
Zimbabwe
South Africa
Mozambique
Uganda
Lesotho
Namibia
Malawi
Zambia
0
5
10
15
HLO Rank
SACMEQ Rank


Ranks - HLO vs. SACMEQ
2000, Math, Primary
Mauritius
Kenya
Seychelles
Mozambique
Tanzania
Swaziland
Botswana
Uganda
South Africa
Lesotho
Zambia
Malawi
Namibia
0
5
10
15
HLO Rank
SACMEQ Rank


Ranks - HLO vs. LLECE
2005, Math, Primary
Cuba
Uruguay
Costa Rica
Mexico
Chile
Argentina
Brazil
Colombia
Peru
Paraguay
Ecuador
Nicaragua
Guatemala
Panama
El Salvador
Dominican Rep.
0
5
10
15
HLO Rank
LLECE Rank

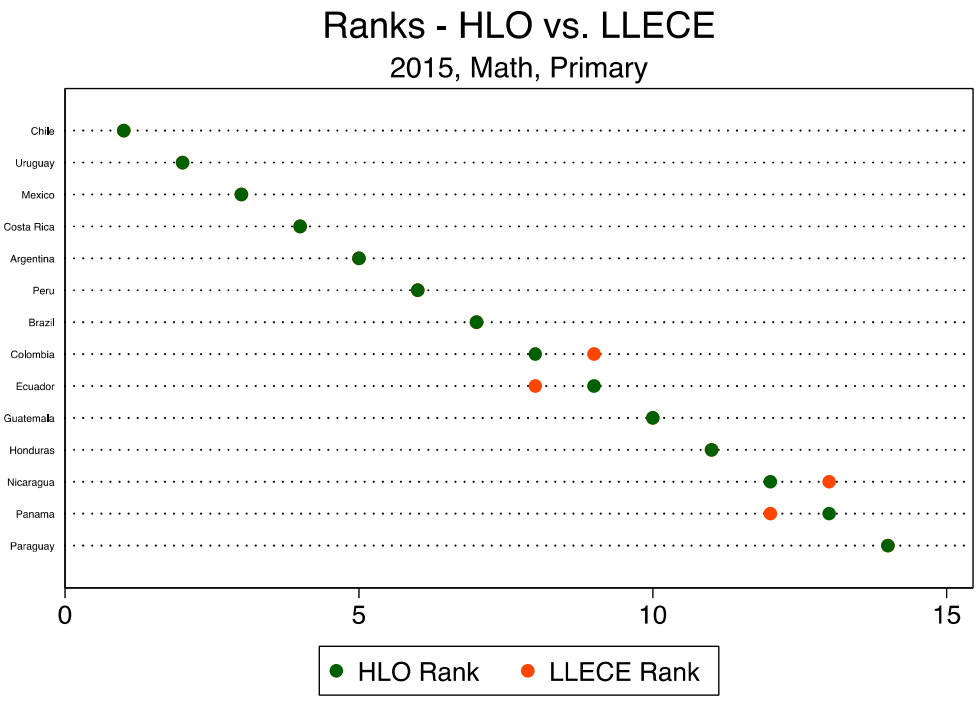

Ranks - HLO vs. LLECE
2015, Math, Primary
Chile
Uruguay
Mexico
Costa Rica
Argentina
Peru
Brazil
Colombia
Ecuador
Guatemala
Honduras
Nicaragua
Panama
Paraguay
0
5
10
15
HLO Rank
LLECE Rank

**Table 2. List of countries used for the linking between assessments**

| Linking number | Anchored assessment | Reference assessment | List of countries used for linking |
|---|---|---|---|
| 0 | FIMS, SIMS, FISS, SISS, IAEP, first wave of TIMSS & PIRLS | NAEP (different years and grades) | USA |
| 1 | LLECE I, grades 3-4, math | TIMSS 1995, grade 8 | Colombia |
| 2 | LLECE I, grades 3-4, reading | PIRLS 2001, Grade 4, reading | Argentina, Colombia |
| 3 | LLECE III, grade 6, math | TIMSS 2011, grade 4, math | Chile, Honduras |
| 4 | LLECE III, grade 6, reading | PIRLS 2011, grade 4, reading | Colombia, Honduras |
| 5 | SACMEQ II, grade 6, math | TIMSS 2003, Grade 8, math | Botswana, South Africa |
| 6 | SACMEQ III, grade 6, reading | PIRLS 2006, grade 4, reading | South Africa |
| 7 | PISA 2000, 15 years old pupils, math | TIMSS 1999, grade 8, math | Australia, Bulgaria, Canada, Chile, Czech Republic, Finland, Hong-Kong China, Hungary, Indonesia, Israel, Italy, Japan, Korea, Latvia, Netherlands, New Zealand, Romania, Russian Federation, Thailand, Macedonia, USA |
| 8 | PISA 2003, 15 years old pupils, math | TIMSS 2003, grade 8, math | Australia, Hong-Kong China, Hungary, Indonesia, Italy, Japan, Korea, Latvia, Netherlands, New Zealand, Norway, Russian Federation, Slovakia, Sweden, Tunisia, USA |
| 9 | PISA 2006, 15 years old pupils, math | TIMSS 2007, grade 8, math | Australia, Bulgaria, Chinese Taipei, Colombia, Czech Republic, Hong-Kong China, Hungary, Indonesia, Israel, Italy, Japan, Jordan, Korea, Lithuania, Norway, Qatar, Romania, Russian Federation, Serbia, Slovenia, Sweden, Thailand, Tunisia, Turkey, USA |
| 10 | PISA 2012, 15 years old pupils, math | TIMSS 2011, grade 8, math | Australia, Chile, Chinese Taipei, Finland, Hong-Kong China, Hungary, Indonesia, Israel, Italy, Japan, Kazakhstan, Jordan, Korea, Lithuania, Malaysia, New Zealand, Norway, Qatar, Romania, Russian Federation, Singapore, Slovenia, Sweden, Thailand, UAE, Tunisia, Turkey, USA |
| 11 | PISA 2015, 15 years old pupils, math | TIMSS 2015, grade 8, math | Australia, Canada, Chile, Chinese Taipei, Georgia, Hong-Kong China, Hungary, Ireland, Italy, Japan, Kazakhstan, Jordan, Korea, Lebanon, Lithuania, Malaysia, Malta, New Zealand, Norway, Qatar, Russian Federation, Singapore, Slovenia, Sweden, Thailand, UAE, Turkey, USA, Buenos Aires (Argentina). |
| 12 | PASEC I & II, grade 5, math | SACMEQ III, math | Mauritius (+ linking n°5) |
| 13 | PASEC I & II, grade 5, reading | SACMEQ III, reading | Mauritius (+ linking n°6) |

## Table 3.1. Description of the international anchored benchmarks: Minimum International Benchmark

| Skill | Assessment used for the definition | Lower score limit | What students can typically do |
|---|---|---|---|
| **Primary education** | | | |
| Math | TIMSS | 400 | Students have some basic mathematical knowledge. Students can add and subtract whole numbers. They have some recognition of parallel and perpendicular lines, familiar geometric shapes, and coordinate maps. They can read and complete simple bar graphs and tables. |
| Science | TIMSS | 400 | Students show some elementary knowledge of life, physical, and earth sciences. Students demonstrate knowledge of some simple facts related to human health, ecosystems, and the behavioral and physical characteristics of animals. They also demonstrate some basic knowledge of energy and the physical properties of matter. Students interpret simple diagrams, complete simple tables, and provide short written responses to questions requiring factual information. |
| Reading | PIRLS | 400 | When reading Literary Texts, students can locate and retrieve an explicitly stated detail. When reading Informational Texts, students can locate and reproduce explicitly stated information that is at the beginning of the text |
| **Secondary education** | | | |
| Math | PISA | 420 | At this level, students can interpret and recognize situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions to solve problems involving whole numbers. They are capable of making literal interpretations of the results. |
| Science | PISA | 409 | At this level, students have adequate scientific knowledge to provide possible explanations in familiar contexts or draw conclusions based on simple investigations. They are capable of direct reasoning and making literal interpretations of the results of scientific inquiry or technological problem solving. |
| Reading | PISA | 410 | Some tasks at this level require the reader to locate one or more pieces of information, which may need to be inferred and may need to meet several conditions. Others require recognizing the main idea in a text, understanding relationships, or construing meaning within a limited part of the text when the information is not prominent and the reader must make low level inferences. Tasks at this level may involve comparisons or contrasts based on a single feature in the text. Typical reflective tasks at this level require readers to make a comparison or several connections between the text and outside knowledge, by drawing on personal experience and attitudes. |

Note: * Lower bounds for each benchmark are original values. Adjusted values may differ, especially for PISA benchmarks.

## Table 3.2. Description of the international anchored benchmarks: Intermediate International Benchmark

| Skill | Assessment used for the definition | Lower score limit* | What students can typically do |
|---|---|---|---|
| **Primary education** | | | |
| Math | TIMSS | 475 | Students can apply basic mathematical knowledge in straightforward situations. Students at this level demonstrate an understanding of whole numbers and some understanding of fractions. Students can visualize three-dimensional shapes from two-dimensional representations. They can interpret bar graphs, pictographs, and tables to solve simple problems. |
| Science | TIMSS | 475 | Students have basic knowledge and understanding of practical situations in the sciences. Students recognize some basic information related to characteristics of living things, their reproduction and life cycles, and their interactions with the environment, and show some understanding of human biology and health. They also show some knowledge of properties of matter and light, electricity and energy, and forces and motion. Students know some basic facts about the solar system and show an initial understanding of Earth's physical characteristics and resources. They demonstrate ability to interpret information in pictorial diagrams and apply factual knowledge to practical situations. |
| Reading | PIRLS | 475 | When reading Literary Texts, students can:<br>• Retrieve and reproduce explicitly stated actions, events, and feelings<br>• Make straightforward inferences about the attributes, feelings, and motivations of main characters<br>• Interpret obvious reasons and causes and give simple explanations<br>• Begin to recognize language features and style<br>When reading Informational Texts, students can:<br>• Locate and reproduce two or three pieces of information from within the text<br>• Use subheadings, text boxes, and illustrations to locate parts of the text |
| **Secondary education** | | | |
| Math | PISA | 482 | Students can execute clearly described procedures, including those that require sequential decisions. Their interpretations are sufficiently sound to be a base for building a simple model or for selecting and applying simple problem solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They typically show some ability to handle percentages, fractions and decimal numbers, and to work with proportional relationships. Their solutions reflect that they have engaged in basic interpretation and reasoning. |
| Science | PISA | 484 | Students can identify clearly described scientific issues in a range of contexts. They can select facts and knowledge to explain phenomena and apply simple models or inquiry strategies. Students at this level can interpret and use scientific concepts from different disciplines and can apply them directly. They can develop short statements using facts and make decisions based on scientific knowledge. |
| Reading | PISA | 480 | Tasks at this level require the reader to locate, and in some cases recognize the relationship between, several pieces of information that must meet multiple conditions. Interpretative tasks at this level require the reader to integrate several parts of a text in order to identify a main idea, understand a relationship or construe the meaning of a word or phrase. They need to take into account many features in comparing, contrasting or categorizing. Often the required information is not prominent or there is much competing information; or there are other text obstacles, such as ideas that are contrary to expectation or negatively worded. Reflective tasks at this level may require connections, comparisons, and explanations, or they may require the reader to evaluate a feature of the text. Some reflective tasks require readers to demonstrate a fine understanding of the text in relation to familiar, everyday knowledge. Other tasks do not require detailed text comprehension but require the reader to draw on less common knowledge. |

Note: * Lower bounds for each benchmark are original values. Adjusted values may differ, especially for PISA benchmarks.

## Table 3.3. Description of the international anchored benchmarks: Advanced International Benchmark

| Skill | Assessment used for the definition | Lower score limit* | What students can typically do |
|---|---|---|---|
| **Primary education** | | | |
| Math | TIMSS | 625 | Students can apply their understanding and knowledge in a variety of relatively complex situations and explain their reasoning. They can solve a variety of multi-step word problems involving whole numbers, including proportions. Students at this level show an increasing understanding of fractions and decimals. Students can apply geometric knowledge of a range of two- and three-dimensional shapes in a variety of situations. They can draw a conclusion from data in a Table and justify their conclusion. |
| Science | TIMSS | 625 | Students apply knowledge and understanding of scientific processes and relationships and show some knowledge of the process of scientific inquiry. Students communicate their understanding of characteristics and life processes of organisms, reproduction and development, ecosystems and organisms' interactions with the environment, and factors relating to human health. They demonstrate understanding of properties of light and relationships among physical properties of materials, apply and communicate their understanding of electricity and energy in practical contexts, and demonstrate an understanding of magnetic and gravitational forces and motion. Students communicate their understanding of the solar system and of Earth's structure, physical characteristics, resources, processes, cycles, and history. They have a beginning ability to interpret results in the context of a simple experiment, reason and draw conclusions from descriptions and diagrams, and evaluate and support an argument. |
| Reading | PIRLS | 625 | When reading Literary Texts, students can:<br>• Integrate ideas and evidence across a text to appreciate overall themes<br>• Interpret story events and character actions to provide reasons, motivations, feelings, and character traits with full text-based support<br>When reading Informational Texts, students can:<br>• Distinguish and interpret complex information from different parts of text, and provide full text-based support<br>• Integrate information across a text to provide explanations, interpret significance, and sequence activities<br>• Evaluate visual and textual features to explain their function |
| **Secondary education** | | | |
| Math | PISA | 633 | Students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare, and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well-developed thinking and reasoning skills, appropriate linked representations, symbolic and formal characterizations, and insight pertaining to these situations.<br>They begin to reflect on their work and can formulate and communicate their interpretations and reasoning. |
| Science | PISA | 607 | Students can identify the scientific components of many complex life situations, apply both scientific concepts and knowledge about science to these situations, and can compare, select and evaluate appropriate scientific evidence for responding to life situations. Students at this level can use well-developed inquiry abilities, link knowledge appropriately, and bring critical insights to situations. They<br>can construct explanations based on evidence and arguments based on their critical analysis. |
| Reading | PISA | 607 | Tasks at this level that involve retrieving information require the reader to locate and organize several pieces of deeply embedded information, inferring which information in the text is relevant. Reflective tasks require critical evaluation or hypothesis, drawing on specialized knowledge. Both interpretative and reflective tasks require a full and detailed understanding of a text whose content or form is unfamiliar. For all aspects of reading, tasks at this level typically involve dealing with concepts that are contrary to expectations. |

Note: * Lower bounds for each benchmark are original values. Adjusted values may differ, especially for PISA benchmarks.