

« Cliometrics of Learning-Adjusted Years of Schooling: Evidence from a New Dataset »

Auteurs


Nadir ALTINOK, Claude DIEBOLT

Document de Travail n° 2023 – 02

Janvier 2023

Bureau d'Économie
Théorique et Appliquée
BETA

www.beta-umr7522.fr

 @beta_economics

Contact :
jaoulgrammare@beta-cnrs.unistra.fr

Cliometrics of Learning-Adjusted Years of Schooling: Evidence from a New Dataset

Nadir ALTINOK¹ and Claude DIEBOLT²

This version: January 1st, 2023

Abstract: Analyzing education does not only involve years of schooling, quality matters! This paper aims at providing better data on schooling with a focus on learning outcomes. It provides the largest dataset on learning outcomes, years of schooling and learning-adjusted year of schooling (LAYS) with comparable data between 1970 and 2020. The quantity dimension is measured by years of schooling and uses the latest data from Barro and Lee (2013), while the quality dimension is taken from linking standardized, psychometrically-robust international achievement tests and hybrid tests. The data are available for more than 120 countries between 1970 and 2020. Several findings can be highlighted. A global convergence on both learning outcomes and enrollment has occurred since 1970, but a breakdown can be found after 1990. A very low number of countries perform better over time regarding the quality of schooling, while most countries have a stable level of learning outcomes.

Keywords: Quality, Human Capital, Education, International, Achievement, Database, Cliometrics, PISA, TIMSS, SACMEQ, PASEC, LLECE, EGRA.

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

¹ Nadir ALTINOK, BETA/CNRS & University of Lorraine (France). Email: nadir.altinok@univ-lorraine.fr (corresponding author)

² Claude DIEBOLT, BETA/CNRS & University of Strasbourg (France). Email: cdiebolt@unistra.fr

*"It is the lack of relevant data more than the lack of relevant theory that is often the greater problem in research. In this way, cliometricians have made some of the greatest contributions to the fields of economics and history by discovering and compiling new data sets that can then be used by future researchers to better understand the evolution and growth of economies over time. The accumulation of the data is in itself monumental in many respects, but its usefulness has been expanded by the rapid growth of computing power. The ability to handle "big data" is not a cliometric issue by itself, but the construction of significant, important historical data sets, which can then be analyzed using the latest econometric techniques and computer programs, is very much a contribution of cliometrics. The marriage of cliometrics and big data is a natural one, and has been exploited by economic historians in new and creative ways."*³

1. Introduction

This paper summarizes the last decade of our scientific collaboration. We present a new data set on two dimensions of schooling, namely learning outcomes and years of schooling. Moreover, by combining the two dimensions, we propose a new measure for learning-adjusted years of schooling (LAYS). Compared to previous research, our dataset is novel in several ways. We combine several of the advantages of existing datasets in order to obtain a unique dataset offering multiple improvements, from the methodological point of view but also in terms of country and time coverage. First, we propose the first real panel dataset with 5-year intervals between 1970 and 2020 for more than 120 countries and for both dimensions of schooling (i.e., quality and quantity). To our knowledge, this is the largest existing dataset covering both dimensions of schooling. Second, our dataset aggregates all existing learning achievement tests in order to obtain comparable scores across countries, while most previous papers propose separate scores for each dimension. These scores are computed with a parsimonious methodology which combines multiple improvements derived from previous papers. For instance, we first focus on the actual results of student achievement tests, and only then do we use imputation methods for predicting results in a multiple dimension analysis (i.e., skill, year, grade, level of schooling). In a second step, additional data on schooling and literacy are used for the multiple imputation process. This makes our data close to the original results of achievement tests, but it also provides real panel data over time. Therefore, the data are more extensive than those in previous datasets and go back as far as 1970 for a large number of countries.

³ Diebolt C. (27 February, 2020a). "Building a bridge between theoretical models and history", Springer Nature Interview: <https://www.springernature.com/gp/researchers/the-source/blog/blogposts-life-in-research/claude-diebolt/17744496>. See also Diebolt, 2016, Diebolt and Hauptert, 2019a, 2019b, 2020, 2021, 2022b, 2022a.

Schooling is not the same as learning (World Bank, 2018, Filmer et al., 2020). We need better and more extensive data on education (World Bank, 2018). Most governments devote a significant share of their budgets to education. Total public expenditure on education accounted for 10.7% of total government expenditure in OECD countries in 2018 (OCDE, 2021). The proportion varies across countries, from less than 7% in Greece to 17% in Chile. In the meantime, schools are often too low-quality to generate human capital. Shortfalls in quality persist for many reasons (World Bank, 2018). First, pursuing good policies may not have positive impacts for policy actors. Second, the bureaucracies tasked with implementing policies often lack the capacity or the incentive to do so effectively. For instance, a recent study conducted by the World Bank in seven countries in Sub-Saharan Africa found that, on average, 3 in 10 fourth-grade teachers had not mastered the language curriculum they were teaching (Bold et al., 2017).

Better measurements of schooling highlight the political and bureaucratic failures that lead to the poor quality of schools. Information is thus an essential step which encourages citizens to demand more from their leaders and service providers (World Bank, 2018). In addition, a good measurement is essential to developing research and analysis to inform policies that improve human capital. Measures that only capture school quantity neglect the differences in the learning outcomes achieved (Hanushek and Woessmann, 2012). Indeed, we cannot reasonably accept that one year of schooling in Singapore or Japan results in the same increase in skills as one year of schooling in a low-income country where education spending is lower, schools are less efficient, and quality of teaching is *ceteris paribus* weaker. This intuition is confirmed by numerous empirical studies (Uwezo, 2014, ASER, 2021).

In parallel, several papers have proposed to contribute to the literature on learning outcomes. To our knowledge, the first paper to attempt to assess learning outcomes in an international setting is the work of Lee and Barro (2001). These authors used direct results from International Student Achievement Tests (ISATs) without any specific methodology for adjusting potential differences between all the series. Another method of anchoring was used by Hanushek and Kimko (2000). These authors adjusted ISATs between 1964 and 1995 using results from the National Assessment of Educational Progress (NAEP).⁴ Their methodology is

⁴ NAEP is the largest nationally representative system of continuing assessments in the United States. Since 1969, NAEP has been a common measure of student achievement across the country in mathematics, reading,

only based on United States scores, and the data are limited to the period 1964-1995. A research paper by Hanushek and Woessmann (2012) aimed at correcting some of these imperfections by using an approach that assumes stability over time of the variance of quality of student achievement in a restricted number of OECD countries. The authors suggest two criteria for a group of countries to serve as a standardization benchmark for performance variation over time. First, the countries have to be member states of the relatively homogenous and economically advanced group of OECD countries over the whole period of ISAT observations. Second, the countries should already have seen a substantial enrollment in secondary education in 1964. The authors suggest 13 countries that meet both of these measures of stability, named the “OECD Standardization Group” (OSG) of countries.⁵ Hanushek and Woessmann (2012, and 2015, hereafter HW) assume that cross-country variations among the OSG countries have not varied substantially since 1964. On this assumption, they build new indicators of student achievements and educational quality.

In a series of other papers, alternative datasets are proposed to augment HW’s sample by incorporating data on reading assessments and information from other sources, such as Regional Student Achievements Tests (RSATs) for countries that do not participate in ISATs (Altinok and Murseli, 2007, Angrist et al., 2013b, Altinok et al., 2014, Altinok et al., 2018). The main idea of this alternative methodology is to calculate an 'exchange rate' that can be used to adjust differences in difficulty and grading scales among different achievement tests. Our paper follows this path, with several improvements.

In their paper, Altinok et al. (2018) provide panel data with, for most countries, several observations that correspond to “Harmonized Learning Outcomes” (HLO).⁶ Using only learning outcomes may not be sufficient, since the stock of human capital is not the same across countries. By combining both quantity and quality of schooling, LAYS avoids the weaknesses of using either of these measures alone. Unlike the years of schooling measure

science and many other subjects. More information about NAEP can be obtained in National Academies of Sciences and Medicine (2017)

⁵ The OSG countries are Austria, Belgium, Canada, Denmark, France, Germany, Iceland, Japan, Norway, Sweden, Switzerland, the United Kingdom, and the United States.

⁶ In addition to this paper, other analyses are conducted, with roughly the same methodology (Altinok and Murseli, 2007, Angrist et al., 2013b, Altinok et al., 2014, Altinok et al., 2018, Angrist et al., 2021). In this latter paper, the methodology is somewhat different since the standardization procedure is based on a regression of the form $y_i = a + bx_i + \varepsilon_i$. The equation is estimated using data for all countries that took part in both assessments and is then used to make an estimation of Y for those countries that only participated in X. According to Kolen and Brennan (2014), this anchoring is close to linear equating.

alone, it keeps a focus on quality; and unlike the quality measure alone, it encourages schooling for all children, whether or not they will perform highly in achievement tests. A similar analysis is conducted by Glawe and Wagner (2022). These authors propose a new LAYS database over the period 1995-2015, using the World Bank (2018) methodology. Glawe and Wagner (2022) find that the average number of learning-adjusted years of schooling (LAYS) for the cohort of the population aged 25-29 is much lower than that of unadjusted years, which is around 9, ranging from 3 (for Morocco) to 15 (for South Korea). On average, mean school years are reduced by about 2.3 years when the quality of schooling is taken into account. One drawback of the study by Filmer et al. (2020) is the restriction of the data within a cross-country dimension. These authors compute average scores for the most recent years. While Glawe and Wagner (2022) innovate by providing a panel dataset on both quantity and quality of schooling, both the number of countries (35) and the time span (1995-2015) are very low.

An alternative analysis on the measure of human capital was conducted by Lim et al. (2018). These authors combine both education and health indicators in order to obtain a panel database on human capital for 195 countries between 1990 and 2016. Lim et al. (2018) estimate educational attainment using 2522 censuses and household surveys, while they base learning estimates on 1894 tests among school-age children. In the first step, the authors use the same methodology as Altinok et al. (2014), but they add different features. For instance, national assessments were included, such as the US National Assessment of Education Progress⁷ and the India National Achievement Survey (Sreekanth, 2015), along with representative studies measuring intelligence quotient (IQ) in school-aged children (Raven, 1936, Wechsler, 1949, Dunn, 1959).⁸ Another novelty is the methodology used to estimate test scores for all countries, years, and ages (5-year age groups from 5 to 19 years), where the authors used an imputation methodology – spatiotemporal Gaussian process regression – with per capita mean years of education as a predictor (Gakidou et al., 2010, Vos et al., 2017,

⁷ Institute for Education Sciences National Center for Education Statistics. National Assessment of Educational Progress- overview. <https://nces.ed.gov/nationsreportcard/about/> (accessed April 15, 2022).

⁸ IQ data included the Wechsler Intelligence Scale for Children (Wechsler, 1949), the Raven's Standard Progressive Matrices (Raven, 1936), and the Peabody Picture Vocabulary test (Dunn, 1959).

Lim et al., 2018).⁹ This imputation methodology was quite new in education and resulted in a full panel dataset between 1990 and 2016.

In our paper, we propose to extend the data on learning outcomes in order to provide broader information on three complementary indicators, namely years of schooling, learning outcomes and LAYS, from 1970 to 2020. To this end, we combine the advantages of various existing methodologies. First, we follow the initial anchoring methodology presented in Altinok et al. (2018) to obtain a global dataset on learning outcomes for most countries. Second, we use an imputation methodology, quite similar to the one used in Lim et al. (2018), in order to obtain a panel dataset for both dimensions of schooling – quality and quantity – between 1970 and 2020. In addition to this, our paper extends the Filmer et al. (2020) dataset on LAYS by combining both our learning-outcomes measures and the most recent data on years of schooling. In this way, our project seeks to fill two major gaps: to obtain a large panel database on learning outcomes and years of schooling over a 50-year period using alternative imputation techniques, but also to provide an alternative way to measure years of schooling by combining years of schooling with learning outcomes.

In section 2, we focus on student achievement tests and methodology. The final database is then presented with several analyses based on trends and the potential impact of learning on economic growth. We then highlight the potential limitations of the study and conclude with new perspectives.

2. Data and methodology

This section briefly describes the achievement tests we used to construct the dataset on learning achievement and additional data related to quantity of education, such as years of schooling.

The first set of achievement tests consists of international assessments which are already standardized; the second are regional standardized achievement tests; and the third are hybrid achievement tests. Each test covers different numbers of countries, from 10 to 79 for the Programme for International Student Assessment (PISA) conducted in 2018. By combining

⁹ This method is strengthened across time, space, and age, incorporates both data and model uncertainty, and produces a full-time series of estimates for all geographies with the use of covariate relationships and spatial and temporal patterns in residuals (Lim et al., 2018).

these assessments and making them comparable we are able to include 167 countries or territories covering almost the entire global population. It should be noted that although this is the largest dataset on learning and years of schooling, there are still availability issues. The main issue with data availability is related to learning achievement data. Indeed, data for learning outcomes is available at about 6.8 different points in time on average. For about 55 localities no data are available, but these localities are mostly islands with low populations.¹⁰ For more than two thirds of the countries, comparable data are obtained for almost all years between 1970 and 2020. For fewer than twenty countries/localities, only 1 observation is available.

In addition to learning outcomes, we also provide data for years of schooling. Data from Barro and Lee (2013) are used in their latest version (i.e., September 2021). A multiple imputation procedure is conducted to extend this dataset to 2020 and to cover more countries. As expected, data availability is much greater for the quantity of schooling. Indeed, data for years of schooling are provided for more than 200 countries or territories with an average of 10 different points in time. Meanwhile, data are lacking for only 19 countries and territories. Most of them are islands with very low populations.

Our final database includes mean scores for all 167 countries and territories from 1970 to 2020.¹¹ Compared to previous datasets, our methodology allows us to extend the database over a 50-year interval. Cognitive skills are disaggregated by subject (mathematics, reading and science), schooling level (primary and secondary), and gender (male and female). Since data for years of schooling are disaggregated by schooling level and gender, our Learning-Adjusted Years of Schooling are also available for each schooling level and each gender. Below, we briefly present the methodology used to anchor these in order to provide a global snapshot of learning outcomes over the world.

¹⁰ However, we failed to collect comparable data for the following countries with significant populations: Central African Republic, Comoros, Equatorial Guinea, Eritrea, Guinea, North Korea, Libyan Arab Jamahiriya, South Sudan, Sri Lanka, Suriname, Turkmenistan and Uzbekistan.

¹¹ Data for 2020 is extrapolated by using the latest results, provided mostly in either 2018 or 2019. In that sense, the potential effects of Covid-19 on learning achievement are not included in our analysis.

2.1. A snapshot of learning achievement tests

The best way to evaluate learning outcomes is to focus on student learning achievement tests. A variety of tests have been conducted in recent years. Student achievement tests come in two varieties. The first one focuses on curricula and measures the academic achievement level of students at different grades of primary and/or secondary schools. The second set of tests focuses on students' command of the basic and applied skills that can be identified with a broad concept of literacy, instead an academic achievement in the strict sense.

We identify eight international achievement tests in which more than 160 countries have participated. These groups can be divided into three subgroups. As well as well-known international assessments, a growing number of regional assessments across different regions have been observed since the 1990s. Moreover, some hybrid tests, which are not focused on a single region, have been conducted over the last two decades. Detailed information on these assessments is provided in Table A.1. Only a short presentation of the various existing learning assessments is given below. More information can be obtained in Appendix A.

The *International Association for the Evaluation of Educational Achievement* (IEA) was the first body to measure individual learning achievement for international comparative purposes in the early 1960s. The surveys include the highly regarded "Trends in International Mathematics and Science Study" (TIMSS) and "Progress in International Reading Literacy Study" (PIRLS). The TIMSS test aims at evaluating skills of students in grades 4 and 8¹² in mathematics and science, while PIRLS is based on a reading test in grade 4. Several rounds of TIMSS and PIRLS have been conducted to date.¹³ Another well-known international assessment is PISA (Programme for International Student Assessment). The Organisation for Economic Co-operation and Development (OECD) launched its PISA in 1997. More generally, PISA has assessed the skills of 15-year-olds every three years since 2000 in countries that

¹² A grade consists of a specific stage of instruction in initial education usually covered during an academic year. Students in the same grade are usually of similar age. It is also referred to as a 'class', 'cohort' or 'year' (Glossary of UIS website available at: <http://glossary.uis.unesco.org/glossary/en/home>).

¹³ TIMSS was conducted in 1995, 1999, 2003, 2007, 2011, 2015 and 2019. PIRLS was conducted in 2001, 2006, 2011, 2016 and 2021. In our current version (version 1.0), we failed to include PIRLS 2021 due to the unavailability of results.

together account for almost 90% of the global economy – i.e., the majority of world GDP. Seven rounds of PISA are available to date (2000, 2003, 2006, 2009, 2012, 2015, 2018).¹⁴

Three major regional assessments have been conducted in Africa and Latin America. The *Southern and Eastern Africa Consortium for Monitoring Educational Quality* (SACMEQ) emerged from a very extensive national investigation of the quality of primary education in 15 African countries in 1995-1999, 2000-2002 and 2007.¹⁵ Following a different approach, surveys under the *Programme d'Analyse des Systèmes Educatifs* (Program of Analysis of Education Systems, PASEC) of the Conference of Ministers of Education of French-Speaking Countries (CONFEMEN) have been conducted in the French-speaking countries of sub-Saharan Africa since 1993. PASEC was transformed and standardized in 2012 and two more recent rounds have since been conducted (2014 and 2019). Finally, 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 established in 1994 and is coordinated by the UNESCO Regional Bureau for Education in Latin America and the Caribbean. Assessments conducted by the LLECE focused on learning achievements in reading, mathematics and science¹⁶ in grades 3 and 4¹⁷ in 13 countries of the subcontinent in 1998, and for grade 3 and 6 pupils in 2006, 2013 and 2019.

Hybrid assessments can be considered as a mix of international and regional achievement tests. Wagner (2011) argues that EGRA represent a hybrid type of assessment. The Early Grade Reading Assessment (EGRA) is an individually administered oral assessment of the most basic foundation skills for literacy acquisition in early grades. The assessment lasts for about 15 minutes per child. We compile and include data on the proportion of pupils with a

¹⁴ Two other international assessments are available. Drawing on the experience of the National Assessment of Educational Progress (NAEP), the International Assessment of Educational Progress (IAEP) comprises two surveys first conducted in 1988 and 1991. Under a joint UNESCO and UNICEF project, learning achievements have been assessed as part of the *Monitoring Learning Achievement* (MLA) program on a vast geographical scale in more than 72 countries (Chinapah et al., 2000). This program of assessment is flexible and ranges from early childhood to basic and secondary education to nonformal adult literacy. However, not all the data have been published. Supplementing national reports, a separate report on MLA I was drafted for 11 African countries (Botswana, Madagascar, Malawi, Mali, Morocco, Mauritius, Niger, Senegal, Tunisia, Uganda and Zambia; see Chinapah et al., 2000). As the microdata from IAEP and MLA are not available, we preferred not to include these assessments in our database.

¹⁵ A fourth round of SACMEQ was conducted in 2012-2014. However, the results were not officially released, due to methodological issues.

¹⁶ Science skill was included only after the second round.

¹⁷ A grade is a stage of instruction usually equivalent to one complete year. Hence, grade 3 represents the third year of compulsory schooling – i.e., of primary education in most countries.

0-score in the “Oral Reading Fluency” test provided in almost all EGRA tests. We believe this is the best comparable measure since it is not related to language complexity, which may differ across countries. Similarly to EGRA, ASER is a quick oral assessment on early grades. It has been conducted in rural localities of several countries, including India and Pakistan. We thus include results from ASER studies in both Pakistan and India for multiple rounds in order to obtain comparable results over time, although these studies began quite recently.¹⁸

All achievement tests undertaken and the main information concerning them are summarized in Table A.2 and presented in Appendix A. The methodology used to adjust them in order to yield comparable indicators is presented below.

2.2. Methodology

As presented in section 1, several previous studies have proposed a dataset on learning outcomes and/or LAYS. While some can be considered as “orthodox” papers, others use novel ways of predicting results and covering more countries and years. These “heterodox” studies suffer, however, from robustness and validity issues since most of the data provided are not based on original results. Our priority is to focus on original data based on student achievement tests, as “orthodox” papers have done. Next, we use additional data on literacy, adult learning outcomes and other proxies to improve data availability, similarly to “heterodox” papers. Therefore, our analysis can be considered as being midway between a conservative and a statistical approach. In order to obtain quality-adjusted years of schooling, we follow a four-step process. First, we compile and impute proxies for learning outcomes and years of schooling in order to obtain a large panel dataset on education variables between 1970 and 2020. Second, we compile and anchor all the student achievement tests presented in the previous section. Third, we use the global dataset on education variables compiled in step 1 to fill in missing values for our standardized measure for learning outcomes. We use a multiple imputation technique. Last, we combine quantity and quality of schooling and obtain a panel dataset on learning-adjusted years of schooling. Since we have disaggregated data for each gender, our dataset includes comparable data for both male and female populations. Below we present the details of the methodology.

¹⁸ ASER conducted its first study in 2005 in India, while Pakistan began in 2012. As these studies were only conducted in rural areas, we have been unable to extrapolate the results for the entire country and therefore considered them as representative of the whole.

Step 1. Preparation of background data for quantity of education. In order to obtain comparable data for a large number of countries, we employ a multiple imputation technique on several proxies of learning outcomes, but also variables which evaluate quantity of schooling. Four sets of variables are included in the "quantity dataset": proxies of quality of schooling with available data from 1970 to 2020; other variables without large data availability; variables related to quantity of schooling and the new literacy rates provided by Le Nestour et al. (2022). The list of all variables is provided in Table B.1 and the correlation matrix can be found in Table B.2. In particular, data for years of schooling are updated throughout this imputation process. Indeed, Barro and Lee (2013) propose a global dataset on years of schooling over the period 1950-2015 in their latest update (i.e., the version of September 2021). Data are not available for all countries and for 2020. Multiple imputation has been shown to reduce bias and increase efficiency compared to listwise deletion. We use the "Amelia II" program within R Statistics in order to obtain an imputation of the missing values. Amelia II imputes missing values using a bootstrapping approach called the EMB (expectation-maximization with bootstrapping) algorithm (Honaker et al., 2011). Multiple imputation involves imputing m values for each missing cell in a data matrix and creating m "completed" datasets. In order to obtain more precise predictions, we perform multiple imputation in several steps. Across these completed datasets, the observed values are the same, but the missing values are filled in with a sample of values from the predictive distribution of the missing data (Honaker, King and Blackwell, 2011). The imputation model in Amelia assumes that the complete data are multivariate normal. If we denote the $(n \times k)$ dataset as D (with observed part D^{obs} and unobserved part D^{mis}), then this assumption is

$$(1) D \sim \mathcal{N}_k(\mu, \Sigma),$$

which states that D has a multivariate normal distribution with mean vector μ and covariance matrix Σ . The essential problem of imputation is that we only observe D^{obs} , not the entirety of D . Amelia assumes that the data are *missing at random* (MAR). This assumption means that the pattern of missingness only depends on the observed data D^{obs} , not the unobserved data D^{mis} . In order to make the MAR assumption more plausible, additional variables can be included in the dataset D . This auxiliary information is useful when it helps predict the value of the missing data. In Figure B.1, we present a schematic diagram of our approach to multiple imputations with the EMB algorithm.

Within each step, 10 different imputations are conducted. Results are obtained by computing the average value of each variable across these different imputations. Ultimately, we obtain a global dataset on 18 education variables which are considered important to both quantity and quality of schooling (see Table B.1. for the list of all variables).

Step 2. Equating achievement tests. 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, the assessments are not directly comparable since their difficulty and content may differ. The method to link achievement tests is quite similar to the one used in previous papers (Altinok and Murseli, 2007, Angrist et al., 2013a, Altinok et al., 2014, Altinok et al., 2018). We present below the general methodology, while a more detailed description can be found in Altinok et al. (2018). When building globally comparable education quality estimates, we rely on classical test theory (Holland and Hoskens, 2003). Specifically, we use pseudo-linear linking (hereafter, the “exchange rate” approach) and equipercentile linking.

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 which took an ISAT and an RSAT at the same point in time. Since ISATs and RSATs are psychometrically robust, sample-based tests 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 by using the 'exchange rate'. Table B.2 provides the list of countries that overlap in assessments. This enables the inclusion of Regional Standardized Achievement Tests (RSATs) from Latin America and sub-Saharan Africa, and thus an international comparison. This is a significant addition, since many developing countries have participated in RSATs (LLECE, SERCE, PASEC, and SACMEQ) and HSATs (EGRA, ASER), but rarely or never in ISATs (PISA, TIMMS, PIRLS). The transformation of regional scores into an internationally comparable value is more 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. Student achievement tests are computed as averages taken over different achievement tests administered in the same

year or nearby years, after adjusting their results for differences in difficulty. In parallel, following the approach of Hanushek and Kimko (2000), the adjustment of early international student achievement tests is done with the anchoring of each test to the results of the NAEP.¹⁹

The standardization process is mainly based on the average scores obtained in each test by the set of countries that participate in both of them in order to calculate an 'exchange rate' that can be used to anchor all tests with each other, regardless of the difficulty and the grades. That is, given two tests X and Y , the score of country i in test X , x_i , is converted to the scale of test Y using

$$(2) \ y_i = x_i \times e$$

with

$$(3) \ e = \frac{\mu(y)}{\mu(x)} = \frac{\frac{1}{n} \sum_{i \in X \cap Y} y_i}{\frac{1}{n} \sum_{i \in X \cap Y} x_i}$$

where the average scores for the two assessments, $\mu(x)$ and $\mu(y)$, are calculated over the n countries that have taken part in both of them.²⁰ By using the results of countries taking part in several achievement tests at the same time, we obtain “exchange rates” between achievement tests. Our reference tests are mostly TIMSS and PISA tests. All other tests are linked to these reference tests. Following the original idea of Angrist et al. (2021), we construct the exchange rate over the entire sample, and not only over two tests. For instance, while in previous papers authors used only data for PISA 2003 and TIMSS 2003 for the adjustment of PISA data, we now use all possible combinations between PISA and TIMSS. This approach improves the likelihood of capturing test-specific rather than country-specific differences. When fixing the exchange rate, we assume that the relationship between tests remains constant across rounds. For instance, in previous versions, only countries which simultaneously take part in both TIMSS 2003 and PISA 2003 in mathematics were chosen for the computation of the exchange rate. In this paper, all achievement tests which have been

¹⁹ Since the U.S. participated in NAEP and various international achievement tests at every interval. We adjust old IEA studies by trend on NAEP results. 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. This equating relative to the NAEP results was first used in Hanushek and Kimko (2000).

²⁰ The list of countries used for the equating process is presented in Table B.2.

conducted at roughly the same time are included in the computation of the exchange rate.²¹ This assumption allows us to keep the largest number of countries participating in a given pair of tests linked to each other. As indicated in Angrist et al. (2021), one advantage of this approach is that it means that any changes in test scores over the interval are due to actual progress in learning rather than changes in equating functions between tests. Every new round of tests allows the inclusion of more estimates by enabling the construction of a more robust equating procedure. We discuss this point in section 4.

Step 3. Grouping and predicting scores. The result of step 2 is a global database of learning outcomes, at different levels (primary, lower secondary), different skills (mathematics, sciences and reading) and different years (between 1970 and 2020). Table B.4 provides the data available for each dimension. As in previous papers, one disadvantage of the available data is the lack of a comparable dataset across skills and levels. Data are missing for several years and countries (Table B.5). Indeed, while data for 1970 are more focused on secondary schools and developed economies, most African countries do not have an evaluation of the performance of lower secondary schools. By using this multi-dimensional dataset, we employ a multiple imputation process in order to take into account all this information and to obtain a global dataset on learning outcomes. We proceed to a parsimonious imputation approach in order to get as close as possible to the original results, and with less biased estimations too.

We adapt the imputation procedure in order to obtain the most complete data but also to avoid bias due to prediction. Since most achievement tests were conducted after 1995, we first restrict our imputation procedure to the panel dataset over the years 1995-2020. After having imputed data for this restricted subsample, we use the Amelia package to predict a larger timespan, namely 1980-2020. We follow a multi-step procedure to fill in the missing values. The main idea of our methodology is to first focus on the portion of the data with few missing values, and then use the new imputed data to improve the multiple imputation method for countries and years with few observations. Learning outcomes scores recorded

²¹ For instance, in order to convert PISA scores on the TIMSS scale on mathematics, we use all available combinations across the two assessments: TIMSS 2003/PISA 2003, TIMSS 2007/PISA 2006, TIMSS 2011/PISA 2012, TIMSS 2015/PISA 2015, TIMSS 2019/PISA 2018. The adjustment process for PISA 2000 is based separately with TIMSS 1999 since data from the OECD are not fully comparable between PISA 2000 and PISA 2003 for both math and science.

over time within a cross-sectional unit are observed to vary smoothly over time. In such cases, knowing the observed values of observations that are close in time to any missing value may greatly help the imputation of such values. We therefore use this option in Amelia in order to build a general model of patterns within variables across time by creating a sequence of polynomials of the time index with a second-order level. These polynomials can be interacted with the cross-section unit to allow the patterns over time to vary between cross-sectional units. We believe that this is a reasonable setting since we do not think that all countries have the same patterns over time in all skills and levels. We therefore impute with trends specific to each country by using a specific option available in Amelia (“intercs”). An additional way of handling time-series information is to include lags and leads of the variables of interest. Since the measure of learning outcomes we are using can be considered as a "global stock of human capital", using *lags* and *leads* may increase the accuracy of the EMB multiple-imputation procedure.

Step 4. Adjusting years of schooling with learning outcomes. The learning-adjusted years of schooling (LAYS) are obtained by using equation (4). Filmer et al. (2020) propose combining HLOs with data on years of schooling to construct learning-adjusted years of schooling (LAYS). In that way, LAYS for country i are calculated as

$$(4) \text{LAYS}_i = S_i \times Q_i^b$$

where S_i is the average number of years of schooling of the population cohort and Q_i^b a measure of achievement, relative to a benchmark level b . In their work, this benchmark corresponds to the best performing country or to a maximum benchmark such as 625 points, i.e., the threshold level for reaching the Advanced International Benchmark set by TIMSS.²² Filmer et al. (2020) define the measure of relative learning as:

$$(5) Q_i^b = \frac{L_i}{L_b}$$

where L_i and L_b are the measures of average learning-per-year in countries i and b respectively. L can be thought of as a measure of the learning “productivity” of schooling in each country, while Q is a relative productivity (i.e., productivity in country i relative to that in

²² The Trends on International Mathematics and Science Study (TIMSS) is an assessment conducted by the IEA.

country *b*). We define our benchmark as 700 points, which is the upper bound threshold of the highest benchmark in the PISA test.

3. Results

3.1. The database

In this section we present the coverage and detail of the database. Table 2 presents coverage for country-year observations by region for LAYS and both quantity and quality schooling. The database includes 1,462 observations for LAYS, 1,495 observations for quality of schooling and 2,229 observations for quantity of schooling. The disaggregation of the schooling quality data by schooling level, subject taught and year are shown in Table 3. Most data come from reading scores with 1,426 country-year observations, followed by math scores with 1,294, and lastly by science scores with 1,068. Data for primary schools are more available (1,301 versus 967 for secondary level). Data relative to years of schooling are available for most countries. This shows that our availability of LAYS variables depends mostly on the qualitative dimension of schooling. Across all years, 1,462 observations are provided for the quality-adjusted years of schooling variable, which represents almost two-thirds of observations for years of schooling. Despite the methodology used to expand the data for most countries and years, missing values are problematic, especially for years before 1990 (Table B.4). Indeed, data are missing for learning achievement for about 42% of countries between 1970 and 1985, while this rate drops to approximately 30% between 1985 and 2015. The expansion of the data using multiple imputation results in more comparable data for the period 1970-1990, since about 70% of the dataset on learning outcomes is predicted for these years (Table B.5). Thanks to the expansion of learning achievement tests since 2000, only one-third of the dataset is imputed in 2005, and less than 10% in 2010 and 2015.

Our LAYS indicator is available for most countries around the world, and especially for Sub-Saharan African countries with approximately 350 observations (Table 2). On average, LAYS are equal to 4.62 years. The highest level of LAYS is found in North America with 8.87 years, while South Asian countries have the lowest value (approximately 1.9 years). In comparison with standard years of schooling, LAYS provides a slightly different picture. Figure 1 provides a decomposition of LAYS with its quantitative dimension. The adjustment from traditional years of schooling to LAYS leads to a lower value for all regions. For instance, while Latin

American countries have on average 9 years of schooling, the adjustment for quality of schooling tends to reduce this level by about 3 years. This represents a reduction of approximately 40%. This transformation is greatest for Sub-Saharan Africa (48%), while it is lowest for North America (13%). This adjustment is carried out for a selection of countries across all regions (Figure 2). In a country like Mauritius, years of schooling decrease from 10 to less than 6 when quality of schooling is taken into account. The difference is even larger for countries like Chad (from 8 to 4 learning-adjusted years of schooling).

Figure 3.1 presents our measure of schooling quality, while Figure 3.2 focuses on the LAYS indicator for 168 countries and territories for the year 2015. A few clear trends emerge when our results are compared across regions. As expected, developed economies have a higher quantity and quality of schooling than developing economies; Sub-Saharan Africa and South Asia lag behind all other regions. It is interesting to highlight the highest and lowest performers across regions. Table 4 provides a ranking of countries within regions for the three indicators (quality of schooling, quantity of schooling and LAYS). Some countries perform very well in both dimensions of schooling, such as Singapore, while others like Estonia rank differently. In Asia, the top-performing countries are Singapore, South Korea and Japan in all indicators. The US and Canada perform quite highly in years of schooling, although the quality of schooling is intermediate. In Latin America, the top performers are Cuba, Trinidad and Tobago and Chile, while countries like Honduras and Haiti perform the worst. Israel, Malta and the UAE are the highest performing in the MENA region, whereas countries like Morocco or Yemen have the lowest performance. In Sub-Saharan Africa, Botswana, Mauritius and South Africa perform the best, while Niger is ranked 148th, the lowest rank in the world for 2015.

3.2. Comparison with alternative measures

In addition to our measure, alternative indicators have been proposed for learning outcomes and LAYS. First, we compare our indicator of quality of schooling with other leading measures of human capital (Table 5). Our database complements alternative measures of human capital. Figure 4 shows direct comparisons to learning data used in growth regressions by Hanushek and Woessmann (2012). Strong and significant correlations are found (0.9074), indicating high consistency.

In Table 5, we also compare our data to learning outcomes from other datasets, such as the Harmonized Learning Outcomes, published in 2021, or the dataset published by Altinok et al. (2018). Both datasets are very similar to the methodology used in our paper and show high consistency, since correlations are close to 0.90. Another dataset, published by Lim et al. (2018), confirms the high accuracy of our data in comparably measuring learning outcomes in a panel dataset. Lim et al. (2018) propose a panel dataset on learning outcomes between 1990 and 2010.

In the second part of Table 5, we also compare our LAYS indicators with the one provided by the World Bank. Correlation is again very high for the year 2020. The World Bank does not provide a panel dataset on LAYS, but one for only two different years. We therefore use only our values for 2020 in the comparison. Figure 5 offers a quick overview of the high consistency of our results with those provided by the World Bank. Correlation is equal to 0.88, indicating that the two datasets overlap very well for 2020.

Correlation is just as high as for other alternative datasets. In all cases, these comparisons indicate that even as we expand both country and period coverage, we maintain high levels of consistency with alternative measures where there is an overlap.

3.3. Conditional convergence of schooling indicators over time

Since comparable data are available for more than 100 countries for both quantity and quality of schooling, we can compare LAYS trends over time between 1970 and 2020. In addition, it is also important to test for the stability of learning outcomes over time. To gauge the degree of stability of country performance over time, we estimate country-specific trends as follows. Given an educational indicator, x , let

$$(6) \Delta x_n = \frac{x_n - x_{n-1}}{t_n - t_{n-1}}$$

be its average annual variation between observations $n - 1$ and n , dated at t_{n-1} and t_n respectively. It has become common to gauge improvements in education outcomes over time in terms of a fraction of a standard deviation in test scores. The method can use the standard deviation of a particular year, generally the start or end of a series for a particular unit such as a country (UNESCO, 2019). We therefore adapt equation (6) in order to express the variation according to the standard deviation of the start of each series:

$$(7) \Delta x_{n/SD} = \frac{x_n - x_{n-1}}{(t_n - t_{n-1}) \times \sigma_x}$$

Hanushek and Woessmann (2017) use a projection analysis of learning outcomes and attempt to answer the question, “how fast does any [education] reform achieve its results”. These authors assume that an improvement of 0.5 standard deviations over a long period is possible, in the context of an “aggressive reform plan”. A standard deviation here is the standard deviation of student scores across many countries in almost all achievement tests, such as PISA or TIMSS. In TIMSS and other tests, the standard deviation is set equal to 100 points, based on actual standard deviations among countries participating in the initial year of the test. In addition to the study of Hanushek and Woessmann (2007), Mourshed et al. (2010) use trends from twelve countries or regions to arrive at an improvement of 0.115 standard deviations in ten years, or 0.012 per year. Gustafsson (2014) find that a feasible policy target could be premised on a test score improvement of 0.06 standard deviations a year. This figure is obtained from trends seen in several testing programs in the years 2000 to 2009, and specifically the trends of fast-improving countries. More recently, UNESCO (2019) studied historical trends from PIRLS, PISA and LLECE and found annual improvements in learning outcomes of between 0.01 and 0.06 standard deviations a year.

In experimental studies, improvements appear to be larger than 0.06 standard deviation, and are quite often as high as 0.2 of a standard deviation following an intervention of one or two years (McEwan, 2015). This represents a larger improvement than the findings from the study of UNESCO (2019). One reason for this difference is the fact that results from experimental research are difficult to replicate successfully across an entire system. For instance, while teachers’ unions may not oppose a small research intervention, they might want to alter or even stop the same intervention when expanded (UNESCO, 2019).

The results provided in Table 6 show that the increase in learning outcomes is somewhat lower than that seen in previous findings. Indeed, we find an average improvement of about 0.08 standard deviation for a decade if we aggregate data for all countries. However, the improvement is greatest in Latin America and MENA countries, whereas it is lowest in East Asia and Pacific and South Asia. When we use both quantity and quality of schooling, improvements in LAYS are quite similar to improvements in years of schooling. On average,

countries improved the level of LAYS by about 0.25 standard deviations (per decade) between 1970 and 2020. Again, this improvement is highest in MENA and Latin America.

We can test for a potential convergence of each indicator over time. Since some specific policies can improve learning outcomes or enrollment over a short-term period, we conduct a separate analysis between short-term (20 years, between 2000 and 2020) and long-term (50 years, between 1970 and 2020) trends. Comparisons between the two periods can highlight some specific recent improvements or declines. Developing countries tend to have more scope for rapid improvements, given their distance from what one might think of as natural ceilings for cognitive skills or LAYS (Gustafsson, 2014). While a convergence of years of schooling occurs across countries over a 50-year period, results are more contrasted for learning outcomes and LAYS (Figure 6). As we can see in the bottom-right part of Figure 6, an inverted-U shape is obtained when we compare initial level of LAYS and trends. This may be explained by differences between the economic level of countries. Figures 7 and 8 distinguish between OECD21²³ and developing countries. While a clear convergence is found for OECD21 countries (Figure 7, correlation equal to -0.74), results are more contrasted for developing countries. Indeed, some countries grow quickly because they have low initial level of LAYS, like Botswana or Peru. However, a significant number of countries seem to be trapped in slow improvements, despite an initial low level. This is especially the case of several Sub-Saharan African countries (for example Mozambique, Senegal or Burundi). Conversely, a number of countries with an already high initial level increase their performance on LAYS. This is the case of the Russian Federation, Kazakhstan or Armenia. Some countries appear clearly to be low performers. For instance, France seems to be far from the convergence of countries, since its improvement is lower than other countries such as Italy or Finland. On the contrary, Japan and the USA show larger improvements than expected.

We can thus obtain two main findings from the trends analysis. First, there is no global convergence across countries over time for learning outcomes. Results for years of schooling are more contrasted. Some countries seem to be in a trap while others perform better than expected where convergence is the case. Indeed, we find an inverse U-shaped relationship between LAYS and trends, indicating that specific outliers can be identified. In the meantime,

²³ The list of OECD 21 countries is provided in Table 7.

convergence seems to occur within the OECD 21 countries in the long-run and for both quantity and quality of schooling (Figure 7). Countries with low initial learning outcomes see a higher increase than others (such as Norway or Portugal). Results for LAYS are quite similar to those for learning outcomes and years of schoolings. More work should be done on this topic of convergence, but current results seem to confirm a possible conditional convergence, though not an absolute convergence among countries. Some initial factors such as quality of institutions may be an important condition for school expansion, both in quality and quantity (Glawe and Wagner, 2022).

The second most important result concerns the potential stability of learning outcomes over time. We believe that country performance is not stable in the long run, a hypothesis put forward in Hanushek and Kimko (2000) and Hanushek and Woessmann (2015). Short-term trends are higher for some countries than long-term trends, indicating that specific policies have been successful in the short run. For instance, countries like Portugal or Denmark experienced a significant increase in their learning outcomes over time. For developing countries, results are more contrasted because they have had to increase attainment levels in both primary and secondary schools since 1970. Therefore, a stagnation of learning outcomes cannot be considered as a bad performance, if enrollment increased in the given country at the same time. In Figure 8, some of the high-performing countries are Albania, Peru, Turkey and Vietnam. These countries increased the level of learning outcomes to a high level (i.e., higher than 0.1 standard deviation per decade). It should be noted that although 0.1 SD may seem to be a low performance, it indicates growth of about 1 point each year and thus 10 points for a decade. Over a 50-year period, this means that the performance of the given country increased by about 50 points, i.e., half of a standard deviation. Therefore, an increase of 0.13 SD in Norway can be translated as a global increase of 65 PISA-equivalent score points. Conversely, the slowdown found in France in the short-term trends is greater than it would seem. Indeed, the negative trend of about -0.075 SD is equivalent to a decrease of 15 PISA-equivalent score points in only 20 years. Given the fact that a year of schooling is roughly equivalent to 15-20 points on the PISA scale (Avvisati, 2021), this would mean that France experienced a dramatic decrease in learning outcomes between 2000 and 2020, equivalent to a loss of about 1 year of schooling.

The ranking analysis in Table 7 confirms our rejection of the hypothesis of the stability of learning outcomes. If learning outcomes were stable, the ranking of OECD 21 countries should be similar across years. This is only the case for Japan, while the average range of the ranking is equal to 8, which is quite close to half of 21, the number of countries. Our findings thus confirm the results of De la Fuente and Doménech (2021). We have fairly clear indications of sharp positive or negative trend rises, raising increasing doubts about the validity of the constant quality assumption for learning outcomes in the long run that has often been used in the growth literature.

3.4. The importance of schooling for economic growth

We wonder to what extent schooling variables are associated with economic growth. Since our database includes a large number of countries, we are able to run a regression analysis with about 101 countries. A prior analysis was conducted by Hanushek and Woessmann (2012) with 50 countries. The study of Altinok and Aydemir (2017) used the dataset from Altinok et al. (2014) in order to extend the number of countries to 84 countries. Our database thus has the largest number of developing countries that can potentially benefit the most from schooling accumulation. Although the recent work of Angrist et al. (2021) includes a large number of countries, their time interval is restricted to 2000-2010, while our dataset covers the period 1970-2020.

The full sample of countries for which we have data for all variables is 126. We exclude countries for which more than 25 percent of GDP is derived from rents from natural resources, such as the Democratic Republic of the Congo or Venezuela.²⁴

We use a simple growth model based on the intuition of Nelson and Phelps (1966): a country's growth rate (g) is a function of the skills of workers (H) and other factors (X). These factors include initial levels of income and the investment rate.²⁵ Skills are often referred to simply as the workers' human capital stock. Our specification assumes that H is a one-dimensional index and that growth rates are linear in these inputs:

²⁴ We also exclude from our regression some African countries for which data availability is poor for learning outcomes, such as Madagascar or Burkina Faso.

²⁵ We include the log of initial GDP per capita in 1970 or the nearest year, and the log of the gross fixed capital formation expressed as a % of GDP. Our explanatory variables are included as the average value across the period 1970-2020.

$$(8) g = \gamma H + \beta X + \varepsilon$$

Thus, in our model, it is the *level* of cognitive skills which explains the *variation* in economic output (i.e., GDP per capita).²⁶ The most important specification issue in this framework is the nature of the skills (H) and where they might come from. In the educational production function literature (Hanushek, 2002), skills are explained by many factors such as family inputs (F), the quantity and quality of inputs provided by schools (qS), individual ability (A), and other relevant factors (Z) which include labor market experience, health, and other specific characteristics:

$$(9) H = \alpha F + \beta(qS) + \gamma A + \delta Z + v$$

Human capital, however, is a latent variable that cannot be directly observed. Therefore, we need a correct measure of human capital in order to test its impact on economic growth. The main existing theoretical and empirical work on growth begins by taking the quantity of schooling of workers (S) as a direct measure of H . Following Hanushek and Kimko (2000), we focus on the cognitive skills component of human capital and evaluate H with our learning outcomes variable. In addition to this, we also test the relationship between *LAYS* and economic growth.

Figure 9 presents an added-variable plot of the relation between the average value of each schooling variable and economic growth over the period 1970-2020 (conditional on initial per capita income). Added-variable plots depict the association between two variables after the influences of other control variables (in our case the initial per capita income) are taken out (see also Hanushek and Woessmann, 2008). The plots indicate a positive association between each schooling component and economic growth. We further explore this relationship in our regression model (Table 8).

In columns (1)-(4), we include all countries in the sample, while in columns (5)-(8) we restrict the estimation to OECD countries and then to non-OECD countries (columns (09)-(12)). In all estimations, we include the initial level of GDP per capita and the average value of physical capital over the period 1970-2020. The negative coefficient associated with the initial level of GDP per capita supports the hypothesis of income convergence postulated by neoclassical

²⁶ It should be noted that the form of this relationship has been the subject of considerable debate. Our model can be considered as fitting with both basic endogenous growth models such as that of Lucas (1988) and Aghion et al. (1998) and neoclassical growth models such as that of Mankiw et al. (1992).

growth theory. The relationship between standard years of schooling and economic growth is positive and significant (column 1). We further include only learning outcomes (column 2) or those in association with years of schooling (column 3). All estimations support the hypothesis that education is positively and significantly associated with economic growth over the period 1970-2020. In addition, we test the relationship between the LAYS indicator and economic growth (column 4). Results show a positive and significant effect. Although LAYS may capture important aspects of both quantity and quality of schooling, it seems that the explanatory power of the model with LAYS is lower than that of the model with schooling variables included separately (column (4) versus column (3) respectively).

One possible robustness test of this relationship is a focus on OECD countries (columns 5 to 8). The results show that both schooling variables seem to be positively correlated with economic growth, although years of schooling is not significant at the 10% level when learning outcomes are included in the model (column 7). Developing countries may benefit more from investment in education. The results tend to confirm this hypothesis (columns 9 to 12) with higher values for the coefficients associated with schooling variables.

While the results are robust across various specifications and subsamples, reverse causality and endogeneity bias may potentially be driving the results. Reverse causality would arise if higher economic growth enables countries to develop better education systems that yield higher test performances. The presence of other factors, such as institutions or access to natural resources, which affect growth and are also correlated with cognitive skills, will lead to an endogeneity bias in our estimations. One way to address reverse causality bias is the use of initial values for our schooling variables (Table C.1). The results remain quite similar across different specifications. Obviously, a more detailed analysis of the impact of schooling on economic growth should be conducted with this new dataset. For instance, an estimation with a system GMM estimator, as proposed by Arellano and Bover (1995) and Blundell and Bond (1998), may help to correct endogeneity bias.

4. Limitations of the study

The database obtained using the methodology presented in the previous section may include multiple estimation biases, since some assumptions may be not valid. Below we briefly discuss the potential limitations of our study. Further discussion on these limitations can be

found in Altinok (2017) and Altinok et al. (2018). In order to validate our anchoring methodology based on “exchange rates”, we make some assumptions to keep the results of countries’ student achievement tests intact. These assumptions are mainly based on the fact that we suppose that the populations tested and instruments used are similar across assessments. More generally, we can consider at least four strong differences between achievement tests which may explain why comparability between these assessments should be treated with caution.

A. Differences in score distributions across assessments. First, although achievement tests are meant to be representative of studies of the whole population, there is no reason why the original distributions of scores in each assessment should coincide among themselves. For instance, it may be possible for the distribution of scores in the anchored test for a doubleton country to be different from the distribution of scores of the same country in the reference test. When we use the “exchange rates” methodology (also called “pseudo-linear linking”), we are assuming that the distribution of scores across assessments is similar. Each assessment uses its own psychometric methodology and hence the items included within each test are different for each assessment. This means that the degree of difficulty of items may also differ and thus that the distribution of scores may not be exactly the same between assessments. For instance, the items included in the SACMEQ study may be easier than those in TIMSS and therefore the distribution of scores may be more positively skewed for TIMSS results and negatively skewed for SACMEQ results. This difference may lead to different results for countries that took part in several assessments simultaneously. In order to verify the accuracy of this assumption, we compared the normality of score distributions for each assessment by focusing on “doubleton countries” (Table 8)²⁷. In theory, the distributions of scores for these countries should be similar in order to proceed to pseudo-linear linking. We computed four different measures to test this normality (mean, standard deviation, skewness and kurtosis). The mean is usually used to test the central tendency for quantitative variables, while the standard deviation (SD) is the most widely used measure of dispersion. Normality is generally evaluated with two additional statistics that are known as skewness and kurtosis.

²⁷ We use the methodology provided in Altinok (2017) for the robustness section.

Skewness is a measure of whether a distribution trails off in one direction or another.²⁸ Kurtosis measures the thickness of the tails of a distribution.²⁹ As shown in Table 9, there are some differences between anchored and reference assessments (respectively Assessment 1 and Assessment 2). For instance, while the skewness is positive and close to 1 in SACMEQ, it is close to 0.2 in the TIMSS assessment. The kurtosis comparison allows us to measure the thickness of the tails of a distribution. If our adjusting methodology does not take into account the variability of the distribution across assessments, the kurtosis may be similar for all countries used as “doubloon countries”. If we focus again on SACMEQ countries, it appears that kurtosis is very different as compared to the TIMSS assessment. Indeed, while kurtosis is close to a normal distribution in TIMSS, its value is higher than 5 in the SACMEQ assessment, indicating that the main scores are concentrated in the middle and thus fail to capture very high and very low skill levels. However, the comparison between other anchors does not show very strong differences and hence allows us to adopt a linking approach based on “exchange rates”. In general, we find that our methodology is well suited to all assessments with the exception of SSA studies, such as earlier PASEC assessments and the SACMEQ study. One potential solution to this issue is to use either equipercntile or presmoothed equipercntile linking methods which take into account the distribution of results from each assessment. Instead of using only mean scores, these linking methodologies match each percentile from anchored and reference tests and thus provide a one-to-one percentile matching which avoids the potential difference in the distribution of scores. We preferred to use the presmoothed equipercntile approach for the adjustment of benchmarks, but not for mean scores. Indeed, although this adjustment takes into account the distribution of scores, it does not keep original scores from achievement tests intact. A method which uses percentiles may convert specific thresholds better, but not trends over time based on mean scores. This is why we prefer to use the “exchange rate” approach to compute the adjusted mean scores and the presmoothed equipercntile methodology for benchmarks.

²⁸ A normal distribution has skewness of 0. If the skewness is greater than 0, the distribution is negatively skewed.

²⁹ A normal distribution will have a kurtosis of 3.00. A value less than 3.00 means that the tails are too thick (hence, too flat in the middle), and a value of greater value than 3.00 means that the tails are too thin (hence, too peaked in the middle).

B. Estimation bias may also occur when tested populations differ across assessments used for the linking. The most evident difference can be obtained between PISA and other assessments. While PISA is an assessment based on the age of the student, the other tests focus on the grade tested. This distinction can lead to strong differences in countries where repetition and/ or drop-out rates are high. The focus on a single grade may exclude a proportion of students who repeated classes, while assessments based on the age of students may include these groups. Since we consider that populations are similar and comparable across assessments, this difference may lead to estimation bias. Wu (2010) showed that differences between TIMSS and PISA are not significantly large and comparisons can be made. It is possible to assess to what extent our results may be distorted by this difference by comparing results between TIMSS and PISA for countries that took part in both assessments. In Table 10, we compare the original results for countries that took part in both PISA and TIMSS assessments, in both math and science. We ran a linear regression to test to what extent results in PISA can explain student performances in the TIMSS assessment. Dummy variables were included for both skills and years to control for potential external factors related to these variables. We computed the mean grade tested in each double country and each assessment. Therefore, only countries which took part at both assessments in roughly the same year were included. This yielded about 300 observations in our estimations. In column 1, we regressed the PISA results on the TIMSS results. While the R squared is very high (approximately 0.8), we find that the PISA results are underestimated compared to the TIMSS results, regardless of the grade difference. The most interesting test is to control for grade difference and hence to test to what extent this grade difference may impact the results of the linking process. When we include both dummies for grade difference (column 2), the overall difference between PISA and TIMSS remains quite similar. However, the dummy for a 2-year difference is not significant, which means that the differences found between PISA and TIMSS are not fully due to grade differences. Despite the fact that a significant and very high amplitude effect is found on the 3-year difference in grades tested, this concerns only two countries (Malta and New Zealand). The correlation is very high, suggesting that the anchoring between the two assessments may be considered as valid. However, for specific countries, we observe diverging results. This is especially the case for the Russian Federation and Kazakhstan, where TIMSS results appear to be overestimated. When estimations are done separately for samples with 2 years of difference (column 3) or 3

years (column 4), the results remain quite similar. In order to find which countries diverge between PISA and TIMSS, in Figure 10 we plotted the residuals obtained by using the specification in column 2. For a small number of countries, we detect significant differences between PISA and TIMSS scores. This is the case of South Korea, Bulgaria and Kazakhstan, where student performance appears to be higher in TIMSS than in PISA. Conversely, we find the opposite results for Qatar, Norway and the Netherlands. For most countries, the difference is lower than 40 points in residuals, suggesting that the comparison between the two assessments is valid.

C. The content tested may also vary among assessments. While in assessments such as PISA and PASEC III, items are more focused on competency skills, in all the other tests, items are mostly based on the common curricula of the countries. This distinction may lead to significant differences in countries that are more based on content knowledge rather than competence knowledge. This is especially true of most developing countries but may also include some developed economies. It is possible to test for this difference by focusing on countries that took part simultaneously in TIMSS and PISA assessments with approximately the same grade. Although grades are not exactly the same, we selected countries that took part in both assessments and where the mean grade tested in PISA was grade 9. This represents a lower number of countries than the number of double countries. It is clear that our estimations are not robust since other factors may explain the differences found between the two results, but this analysis presents at least some robustness which is often lacking in previous studies. The results are provided in columns 7 and 8 of Table 9. If we focus on the mean scores, PISA scores are significantly different to TIMSS scores by about 0.8 score points, which is very low. As expected, the restriction to countries where the difference between grades tested is the lowest reduces the difference between PISA and TIMSS scores. When the estimation is made for countries with higher grade differences, the coefficient is higher, but still with a small amplitude (column 8). We can then conclude that characteristics directly related to assessments may not bias the estimation results, at least when we compare the TIMSS and PISA assessments. For the other assessments, since we do not have enough double countries, the estimations cannot be performed. But we can reasonably assume that the differences are greater since the education systems of these countries are still developing and thus any difference between assessments may lead to performance

divergences across them.

D. Hypothesis of absence of country-specific factors. Our methodology supposes that linking equations computed for the doubloon countries are mainly due to the differences between assessments and are independent of country-specific factors. We thus consider these coefficients as “exchange rates”. For instance, when we anchor SACMEQ and PIRLS assessments using South Africa as a doubloon country, we are assuming that the differences in score distribution are only due to the specific characteristics of these two assessments and are independent of the education system in South Africa. Obviously, by using only one or two doubloon countries, our methodology includes a severe estimation bias, since within-country specific factors may explain differences found between anchored and reference assessments. When the number of doubloon countries is high, this bias may be lowered. Since we adapted our methodology in this paper in order to include all assessments for the computation of the “exchange rate”, the bias relative to potential unobserved country-fixed effects may be significantly reduced. Moreover, Altinok (2017) tested this validity hypothesis by comparing linking equations between different rounds of these assessments. This analysis gives us two main results: first, it is important to highlight that country-specific factors are included in the linking process. These are not fully explained by the achievement tests themselves. This means that our estimation strategy is biased due to these country-specific factors. The second main result relates to the number of doubloon countries. The increase in these countries is very sharp in order to reduce the bias related to country-specific factors in the linking process. This is why we adapted our methodology in the current paper by increasing the number of countries in the computation of “exchange rates”.

5. Conclusion

We need better and more complete data to measure both quality and quantity of schooling. The purpose of this paper is to propose a new and more complete dataset on education. This dataset is not only based on learning outcomes, but also on the main measures surrounding the quantity of schooling. Based on previous works, our aim is to provide several improvements.

Three groups of datasets currently exist. First, “orthodox” papers focus only on achievement tests without making any adjustments. Although the dataset included in such a group may be considered as robust, it has the main disadvantage of being available only for recent years and mostly for developed economies. Another group of studies uses specific potential adjustments and thus extends the coverage of countries and the timespan. These studies have the main advantage of providing both better data and more data. A last group of studies, which we consider as “heterodox”, take the risk of introducing several imputation methods. The main advantage of such studies is that they provide the largest datasets and thus cover almost all countries and several decades. However, since they are based on specific methods of prediction, the initial aim of providing better data with better coverage can be called into question.

In this paper, we propose to combine all these three groups of papers in order to keep the original scores intact as much as possible, but also to expand the coverage of countries and the timespan. By using the results from original student achievement tests, our dataset retains the main advantages of the “orthodox” papers. In addition, we expand this dataset using modern imputation methods through literacy studies and other proxies which can be correlated with learning outcomes.

We thus obtain a dataset with a panel dimension in terms of both quantity and quality of schooling for a large number of countries (i.e., more than 120) and over a 50 year span (i.e., between 1970 and 2020). While our dataset provides mean scores for learning outcomes as previous works have already proposed, we combine these with years of schooling in order to obtain a new indicator called “learning-adjusted years of schooling” (LAYS). Previous papers have already proposed a measure such as this (Filmer et al., 2020). Our main innovation is to obtain a panel dataset for a large number of countries. This dataset may be useful for further works which include the educational dimension in their analysis. Alongside the traditional dimension related to years of schooling, it is also important to include the potential differences related to quality of schooling.

As highlighted in the paper, our dataset is not perfect. Several biases can be found. Probably the greatest drawback of our work is the lack of quality of higher education. Indeed, student achievement tests are only conducted in primary and secondary schools, while innovation processes and sources of growth can be identified in the quality of higher

education for developed countries. Recent works have attempted to include higher education in the human capital dimension (Demirgüç-Kunt and Torre, 2022). One possible extension of our work would be to include this level in the dataset on LAYS. This is our ambition for the years to come.

Figure 1. Average years of schooling of the cohort of 15-64-year-olds, unadjusted and adjusted for learning (using the LAYS adjustment).

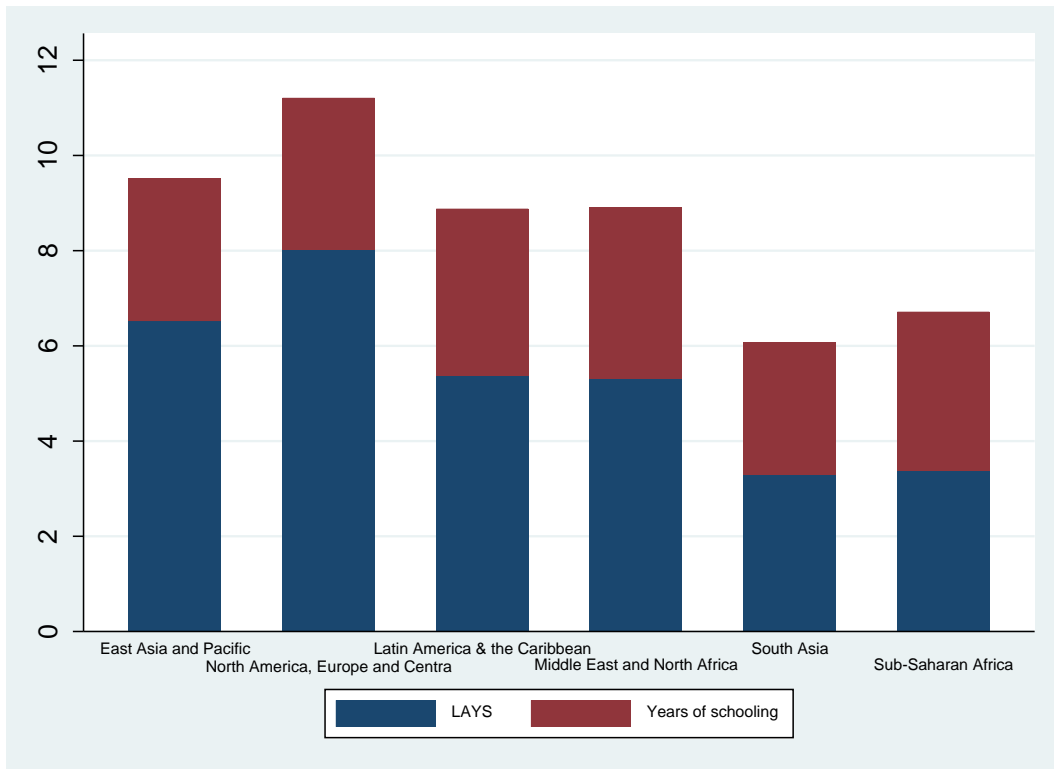


Figure 2. Average years of schooling of the cohort of 15-64-year-olds, unadjusted and adjusted for learning (using the LAYS adjustment), Selection of countries

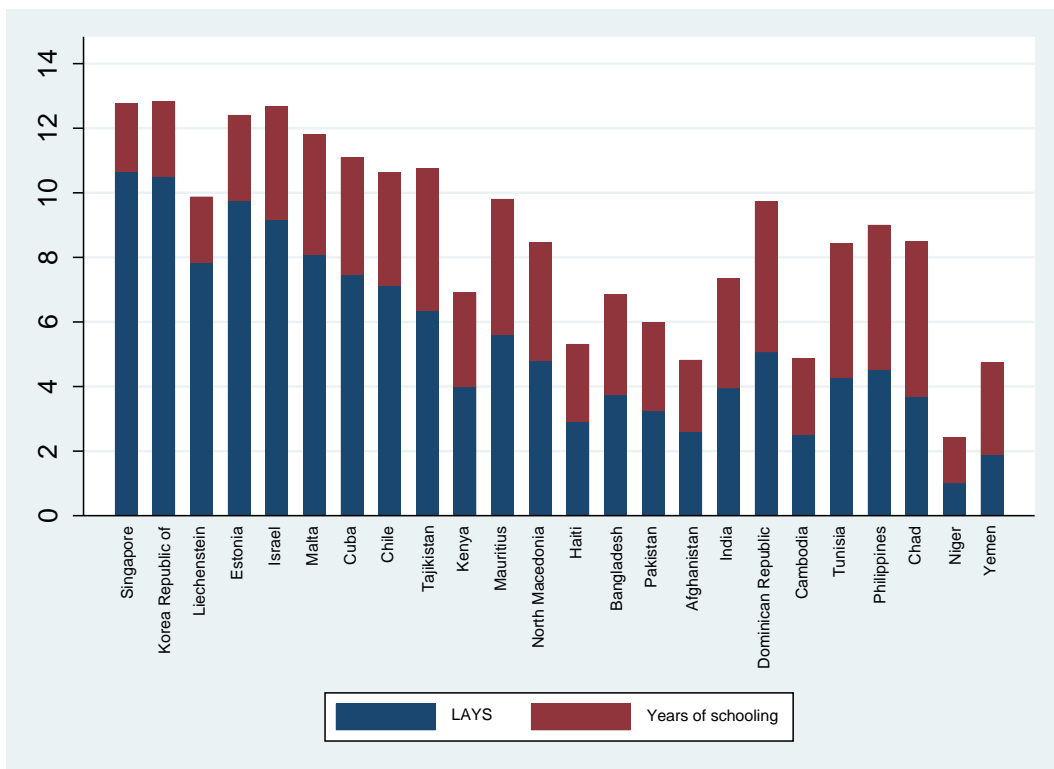


Figure 3.1 Learning Outcomes, both education levels, all skills, 2015

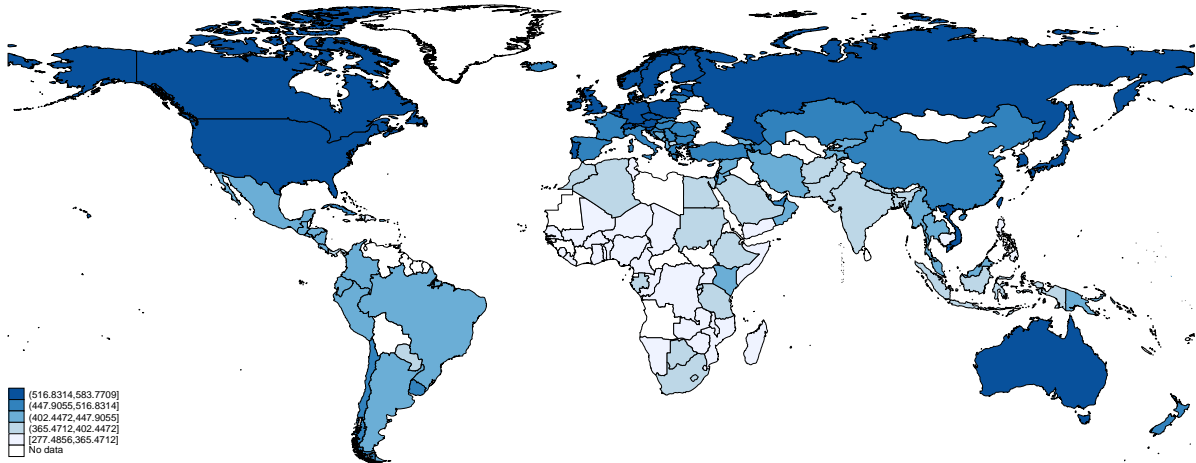


Figure 3.2 Learning-Adjusted Years of Schooling, 2015

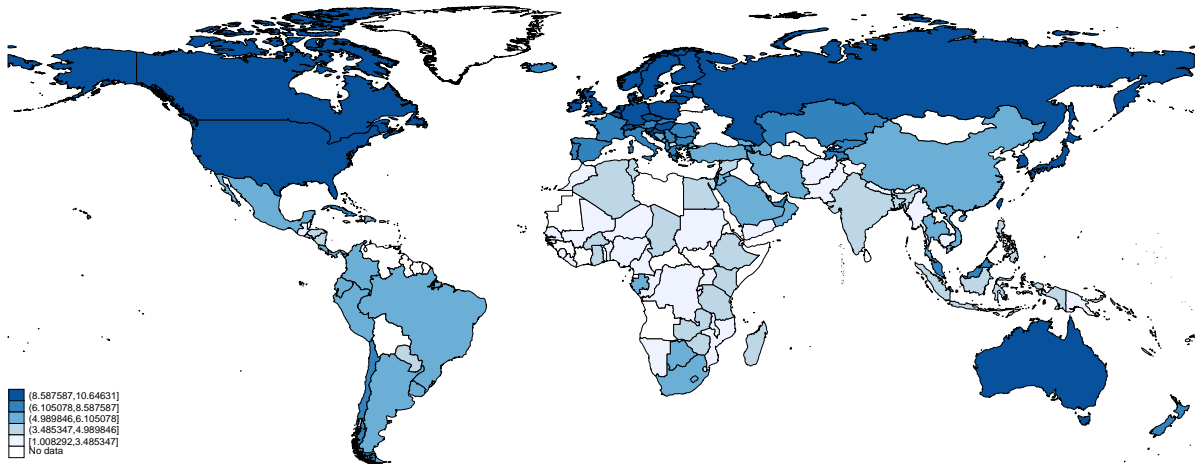
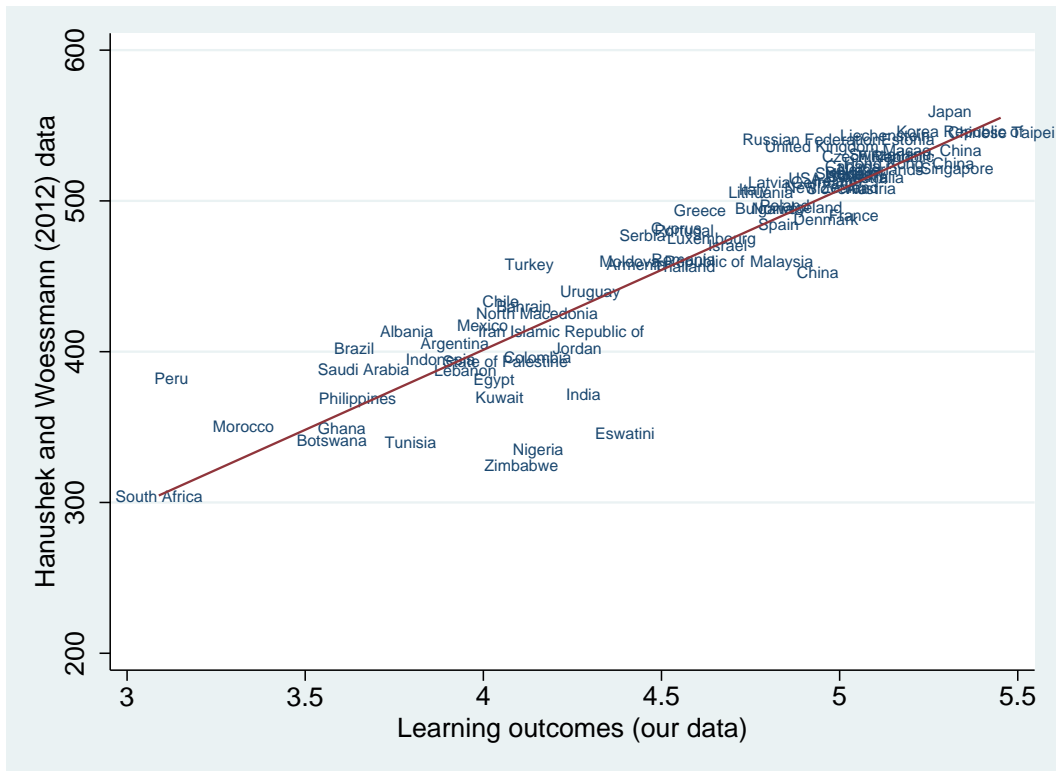


Figure 4. Comparison to Hanushek and Woessmann (2012)



Note: Hanushek and Woessmann (2012) provide estimates of comparable learning measures of human capital, based on an expansion of the original work done by Hanushek and Kimko (2000). We compare these measures for the set of countries included in their growth regressions. Since their measures deal with means for the period 1970-2010, we use the global average for our indicator for the period 1970-2020.

Figure 5. Comparison to LAYS from the World Bank HCI project, year 2020

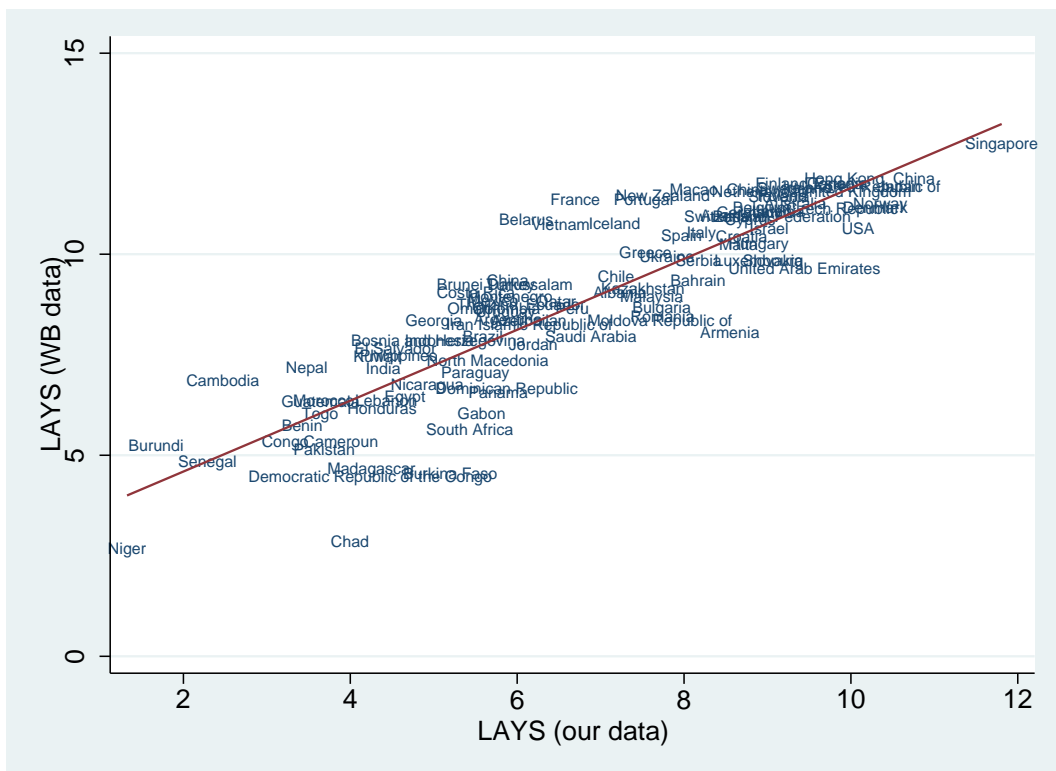


Figure 6. Global trends on education indicators – All countries

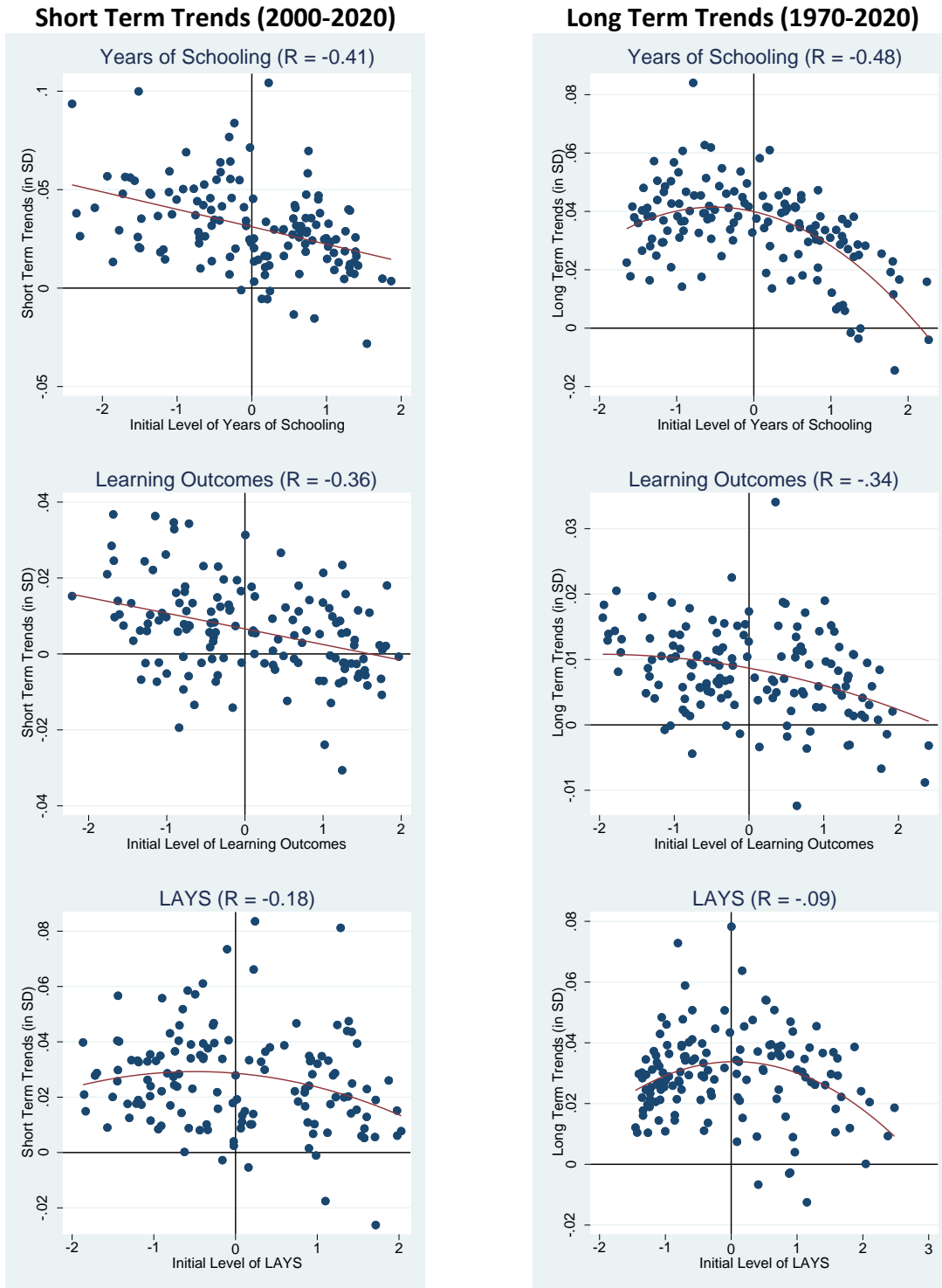


Figure 7. Global trends on education indicators – OECD 21 Countries

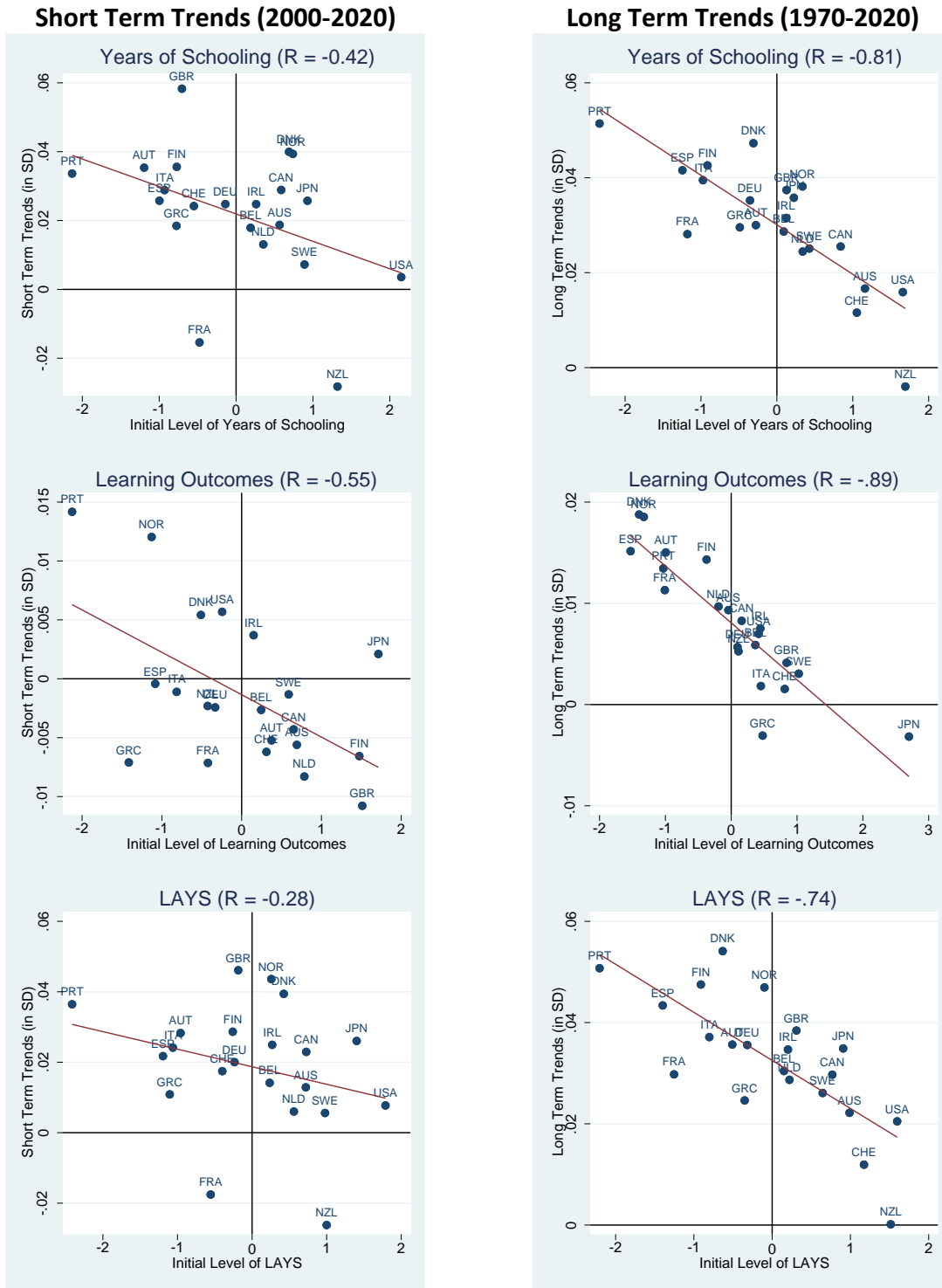


Figure 8. Global trends on education indicators – Developing Countries

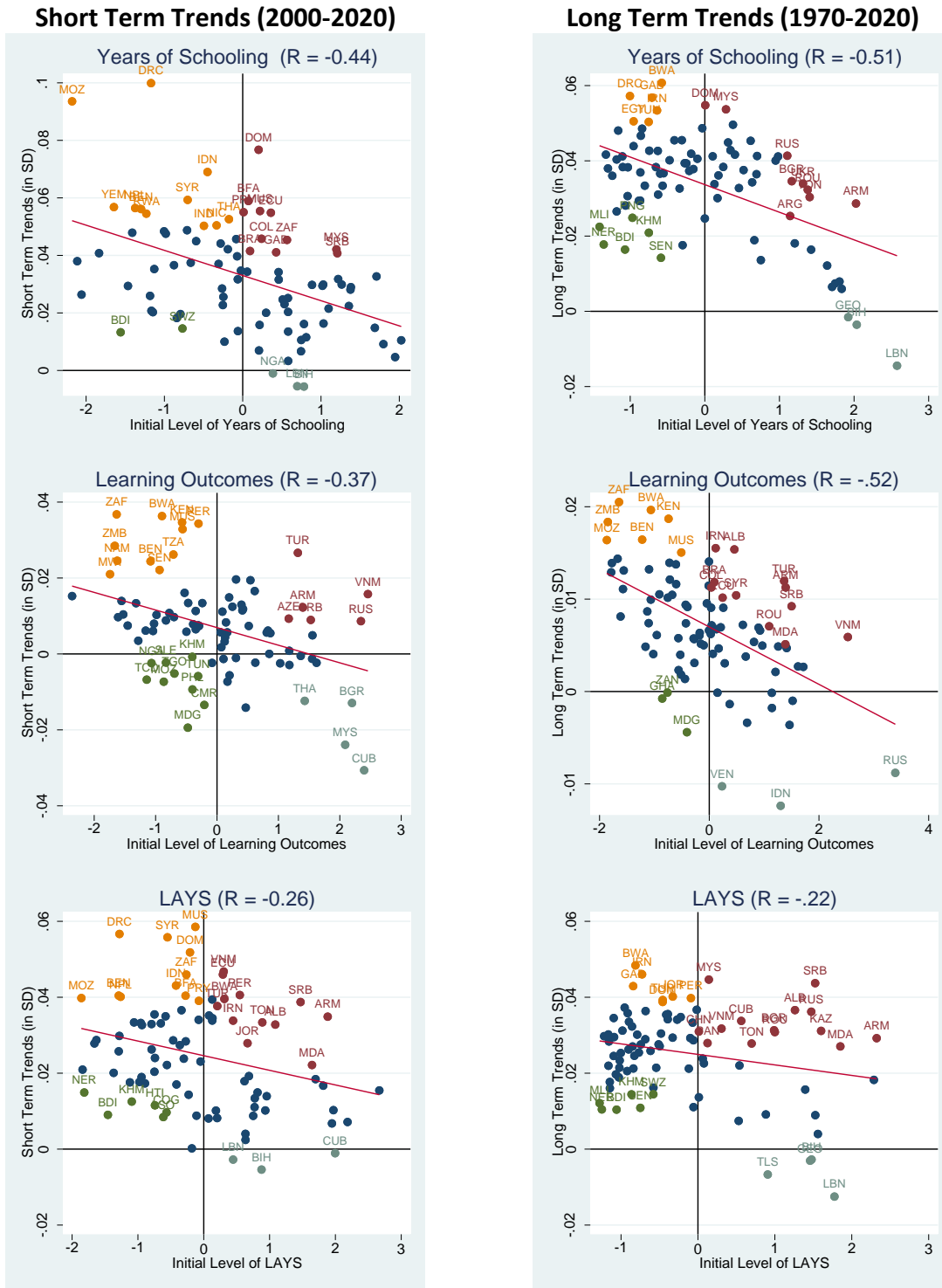
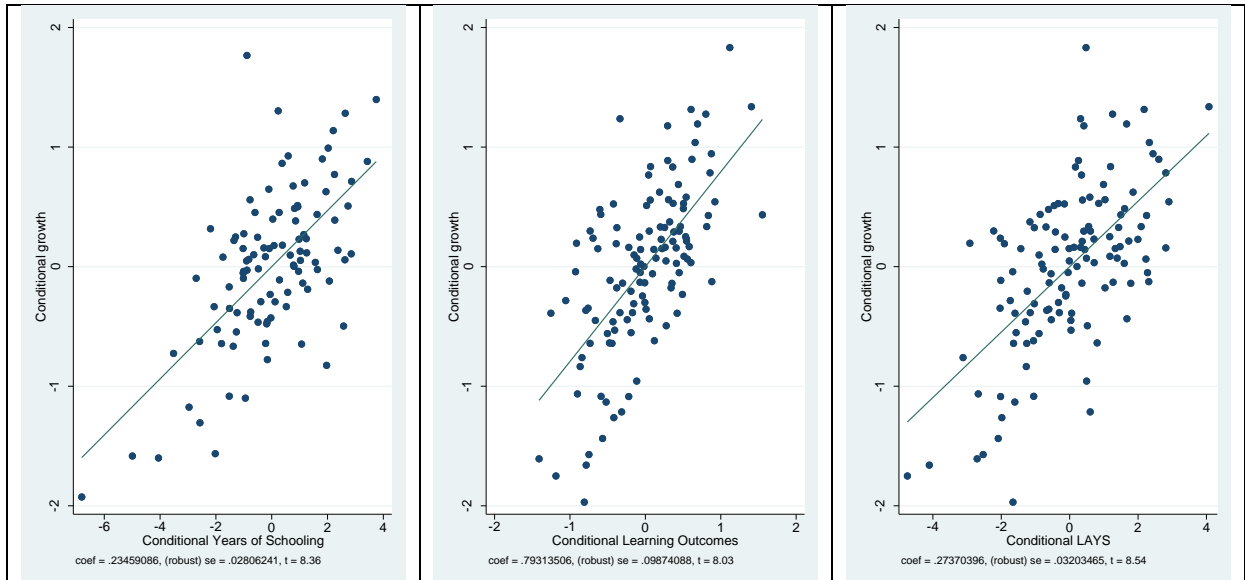
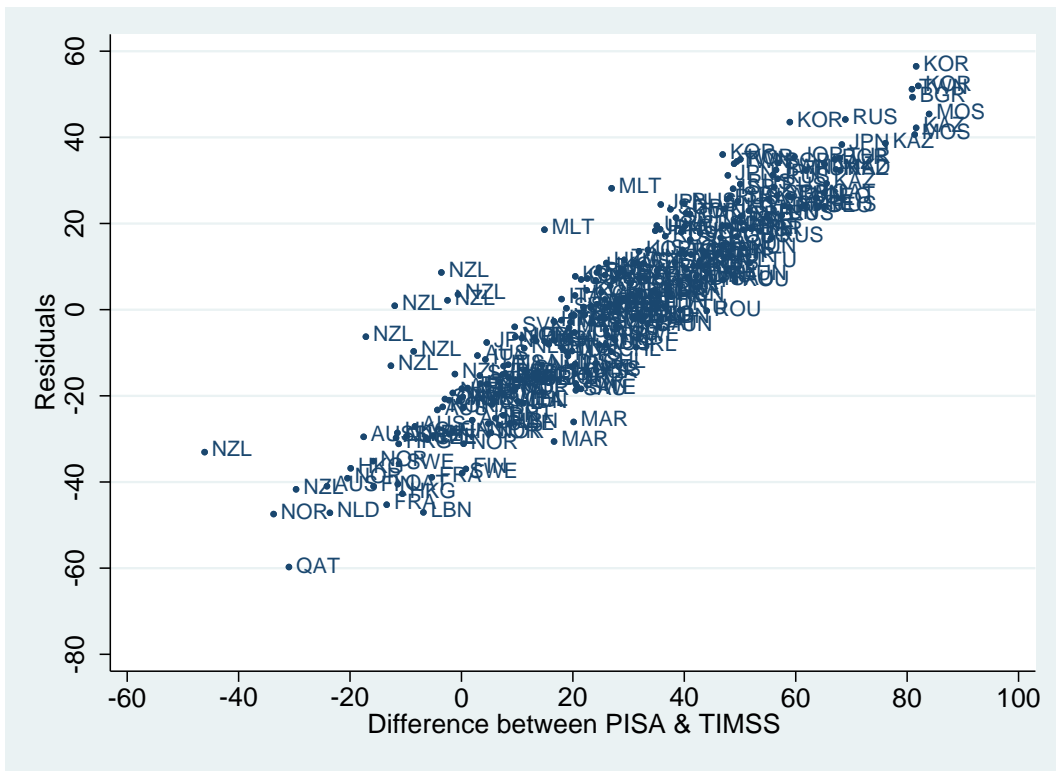


Figure 9. Added variable plot.



Data source: see main text. Notes: Added-variable plot of a regression of per capita (real) GDP growth on the average schooling variable over the period 1970-2020 (conditional on the initial per capita income).

Figure 10. Comparison of original scores between PISA and TIMSS assessments



Note: Both mathematics and science skills are included. All available years are included (1999-2000, 2003, 2006-2007, 2011, 2015 and 2018-19). We group TIMSS and PISA results for similar years or when there is only one year difference. Hence, we directly compare TIMSS 1999 and PISA 2000 results. Results from PISA 2009 were not included. Residuals are obtained from Table 8, column (2).

Table 1. Review of existing datasets on learning outcomes

Authors	Nature	Count ries	Period	Panel data	Methodology – advantages/limits
Hanushek & Kimko (2000)	LO	50	1965-2000	No	Score adjustment is mainly based on the results of the USA on NAEP. No panel data available, imputation of scores for about 20 countries. No disaggregation of data, nor specific benchmarks.
Lee and Barro (2001)	LO	58	1965-1991	Yes	No specific adjustment of scores across tests. Data not available. Inclusion of achievement tests with severe biases like IAEP or MLA. Data not available.
Coulombe and Tremblay (2006)	LIT	14	1960-95	Yes	Use disaggregated results by age group in order to construct synthetic time series of scores using IALS data.
Altinok & Murseli (2007)	LO	104	1965-2003	No	Adjustment similar to Hanushek and Kimko (2000) for most tests, extension made with regional tests with an adjustment based on an "exchange rate" across tests. Panel dimension is not really present. Disaggregation across gender is available. No measure of benchmarks.
Hanushek & Woessmann (2012)	LO	77	1965-2007	No	Methodology partly similar to Hanushek and Kimko (2000) with an approach based on the cross-country variance in mean scores across a group of 13 advanced OECD countries.
Angrist et al. (2013)	LO	128	1970-2010	Yes	The methodology is quite similar to Altinok & Murseli (2007). No disaggregation of data, nor specific benchmarks. Data hosted by the World Bank.
Altinok et al. (2014)	LO	151	1965-2012	Yes	The methodology is quite similar to Altinok & Murseli (2007). Provides a panel dataset for a low number of countries between 1965 and 2012. Disaggregation across gender and type of location available. Three different benchmarks available (minimum, medium and advanced).
Feenstra et al. (2015)	LAYS	145	1950-2019	Yes	Use a combination of different sources of data relative to years of schooling with additional data on rates of return. Rates of return are from Psacharopoulos (1994) and follow the implementation of Caselli (2005). Data from years of schooling are from Barro and Lee (2013), De la Fuente and Doménech (2006), Cohen and Soto (2007) and Cohen and Laker (2013).
Kaarsen (2017)	LAYS	78	1995-2011	No	Author uses results from TIMSS achievement test for years 1995, 1999, 2003, 2007 and 2011. Kaarsen (2017) estimates the quality of education by assuming that the test score is determined by a production function.
Lim et al. (2018)	LO	186	1990-2016	Yes	The methodology uses the same approach as Altinok & Murseli (2007) but adds a prediction of scores in order to obtain a full panel dataset between 1990 and 2016. Disaggregation for each gender available. No data for benchmark.
Altinok et al. (2018)	LO	131	1965-2015	Yes	The methodology uses the same approach as Altinok et al. (2014), adds more countries (like India and China) and more disaggregation between subsamples, like gender or type of location. A larger panel dataset is provided between 1970 and 2015 than the one presented in Altinok et al. (2014) and Angrist et al. (2013).
Schwerdt & Wiederhold (2019)	LIT	31	1970-2014	Yes	Use disaggregated data by age group in order to construct synthetic time series of scores, as has been done by Coulombe & Tremblay (2006). Authors use results from PIAAC and IALS for the estimation of adult literacy across time. Not real panel data. Disaggregation across gender and type of location available. Five different benchmarks available.
Filmer et al. (2020)	LAYS	174	2015-20	No	Use learning outcomes from TIMSS and PISA assessments and years of schooling from Barro and Lee (2017) to construct a hybrid indicator for education, namely the Learning-Adjusted Years of Schooling for about 174 countries. Only two years available (2015 and 2020). No disaggregation of data, nor specific benchmarks.
Angrist et al. (2021)	LO	163	2000-16	Yes	The methodology is similar to the approach of Altinok et al. (2014). The adjustment used is linear, instead of the "exchange rate" used previously. Comparability over time compromised with the use of linear methodology for adjustment. No real panel data provided. Disaggregation for each gender provided. No data for benchmark.
Le Nestour et al. (2022)	LIT	58	1961-2004	Yes	Use disaggregated data from demographic surveys (DHS and MICS) by age group and schooling level in order to construct synthetic time series of literacy scores for populations with 5 years of schooling. Disaggregation for each gender available. No data for benchmark.
Glawe and Wagner (2022)	LAYS	33	1995-2015	Yes	Authors mainly use original TIMSS data for learning outcomes and extended Cohen and Laker (2014) for years of schooling. The database includes almost exclusively either OECD countries or oil countries. Moreover, data are not available and no specific discussion about comparability over time. No disaggregation of data, nor specific benchmarks.
Our data: Altinok (2022)	LO + LAYS	165	1970-2020	Yes	Use a hybrid approach by first focusing on original results from learning student achievement tests and then using an imputation methodology in order to expand the dataset to almost all countries in the world, over a 50-year period (1970-2020).

Table 2. Summary statistics

	Obs	Mean	SD	Min	Max
LAYS					
World	1462	4.62	2.57	0.13	11.81
East Asia and Pacific	213	5.36	2.59	0.79	11.81
Europe and Central Asia	414	6.97	1.61	1.67	10.35
Latin America and the Caribbean	218	4.07	1.54	0.63	7.85
Middle East and North Africa	190	3.88	1.99	0.4	9.44
North America	22	8.87	1.03	6.62	10.09
South Asia	52	1.89	1.05	0.24	4.4
Sub-Saharan Africa	353	2.3	1.29	0.13	5.76
Learning outcomes					
World	1493	419.42	77.77	244.71	594.32
East Asia and Pacific	214	463.57	72.01	322.4	594.32
Europe and Central Asia	442	491.64	38.28	373.09	575.25
Latin America and the Caribbean	218	401.96	35.8	300.67	517.4
Middle East and North Africa	190	393.9	46.74	250.27	505.58
North America	22	517.63	14.46	489.18	536.73
South Asia	52	362.7	14.83	321.94	391.85
Sub-Saharan Africa	355	329.5	33.81	244.71	418.48
Years of schooling					
World	2229	7.35	2.89	0.07	14.35
East Asia and Pacific	362	7.84	2.6	1.38	13.91
Europe and Central Asia	583	9.37	1.89	2.58	13.76
Latin America and the Caribbean	418	7.62	2.04	1.22	12.05
Middle East and North Africa	231	6.5	2.87	0.07	14.35
North America	33	10.24	3	2.53	13.3
South Asia	88	4.87	2.79	0.52	11.58
Sub-Saharan Africa	514	5.11	2.57	0.16	10.74

Table 3. Country-year observations by year

Year	LAYS	Years of sch.	%	Total	Reading	Math	Science	Primary	Secon.	P+M	P+R	P+S	S+M	S+S	S+R
1970	117	203	58	119	113	103	84	106	77	82	60	98	75	77	58
1975	122	203	60	124	118	104	84	111	77	83	59	104	75	77	58
1980	124	203	61	126	120	105	84	113	79	83	59	106	77	77	58
1985	124	203	61	126	120	104	86	113	79	83	60	106	75	79	58
1990	145	202	72	148	138	125	104	130	95	96	72	118	93	95	73
1995	146	202	72	149	139	127	104	130	95	101	74	118	93	95	73
2000	144	203	71	147	137	125	104	128	95	96	71	116	93	95	73
2005	142	203	70	145	134	125	104	126	95	96	71	112	93	95	73
2010	144	203	71	147	138	125	105	128	96	96	71	116	94	96	74
2015	147	202	73	151	144	126	105	132	96	96	71	120	94	96	75
2020	107	202	53	111	125	125	104	84	83	80	67	63	83	75	69
Total	1462	2229	66	1493	1426	1294	1068	1301	967	992	735	1177	945	957	742

Table 4. Ranking of countries for each region and around the world, 2015

Ranking	EAP	W	Western countries	W	LAC	W	MENA	W	SSA	W
Learning outcomes										
1	Singapore	1	Liechtenstein	6	Cuba	49	Israel	36	Kenya	87
2	Korea, Rep.	2	Estonia	8	Chile	50	Malta	47	Mauritius	91
3	Japan	3	Finland	9	Trinidad & T.	56	Bahrain	55	Rwanda	95
Years of schooling										
1	Korea, Rep.	6	USA	1	Trinidad & T.	41	Israel	10	Botswana	55
2	Japan	7	Czech Rep.	2	Cuba	45	Malta	27	South Africa	57
3	Singapore	8	Canada	3	Chile	53	UAE	32	Mauritius	64
Learning-Adjusted Years of Schooling										
1	Singapore	1	USA	5	Cuba	48	Israel	19	Botswana	71
2	Korea, Rep.	2	Canada	7	Trinidad & T.	50	Malta	35	Mauritius	75
3	Japan	3	Estonia	8	Chile	53	UAE	46	South Africa	79

Notes: Ranking is presented within regions in rows and for the world in columns "W". For instance, Singapore is ranked first in learning outcomes in both the World and within the East Asia and Pacific region. Korea is ranked second for learning outcomes, 6th for years of schooling and 2nd for LAYS. Countries from South Asia are not shown since there are only five countries in this area. Western countries include North America, Europe and Central Asia. Full results are provided in Table A.

Table 5. Relationship with alternative learning human capital measures

	Pearson Coefficient	p-value	Observations
Correlation with quality of schooling			
Harmonized Learning Outcomes – HLO (2021)	0.9063	<.001	429
Altinok, Angrist, Patrinos – AAP (2018)	0.8948	<.001	558
Lim et al. – IHME (2018)	0.9309	<.001	850
Hanushek and Woessmann – HW (2012)	0.9077	<.001	77
Correlation with LAYS			
LAYS, Human Capital Index, World Bank, 2020	0.8835	<.001	106

Notes: IHME = Institute for Health Metrics and Evaluation, University of Washington. See Lim et al. (2018) for more information. Comparisons are made using average scores across subjects and schooling levels for HLO and IHME. Average across subjects, schooling levels and years are used for the comparison HW scores.

Table 6. Trends of Schooling Indicators (1970-2020 and 2000-2020)

Region	Initial level (2000)			20 Year Trends (annual improvement in SD)		
	Quality	Quantity	LAYS	Quality	Quantity	LAYS
East Asia and Pacific	452.5	8.2	5.4156	0.0019	0.0243	0.0282
N. America, Europe and Central	490.9	10.0	7.2779	0.0033	0.0196	0.0201
Latin America & the Caribbean	405.8	8.1	4.5130	0.0061	0.0236	0.0257
Middle East and North Africa	396.5	7.1	4.2043	0.0082	0.0474	0.0345
South Asia	361.7	5.3	2.1013	0.0112	0.0525	0.0347
Sub-Saharan Africa	334.4	5.4	2.4727	0.0114	0.0324	0.0252
Total	421.6	7.8	4.9464	0.0063	0.0282	0.0255

Region	Initial level (1970)			50-Year Trends (annual improvement in SD)		
	Quality	Quantity	LAYS	Quality	Quantity	LAYS
East Asia and Pacific	430.0	6.1	3.7630	0.0069	0.0236	0.0331
N. America, Europe and Central	470.9	7.2	5.5282	0.0065	0.0266	0.0295
Latin America & the Caribbean	380.1	5.9	2.6824	0.0070	0.0245	0.0308
Middle East and North Africa	362.4	3.6	2.1068	0.0113	0.0423	0.0361
South Asia	350.7	2.6	0.6295	0.0068	0.0382	0.0288
Sub-Saharan Africa	310.8	3.2	1.3018	0.0096	0.0273	0.0219
Total	397.5	5.3	3.2931	0.0080	0.0280	0.0292

Table 7. Ranking for Learning Outcomes, OECD21 countries

	1970	1980	1990	2000	2010	2020	Average rank	Range max - min rank
Japan	1	1	1	1	1	1	1	0
Finland	15	9	3	3	2	2	5.7	13
Ireland	6	16	13	11	11	3	10	13
Sweden	2	15	9	7	16	4	8.8	14
United Kingdom	3	2	2	2	12	5	4.3	10
USA	7	10	12	12	9	6	9.3	6
Canada	11	6	5	5	5	7	6.5	6
Norway	19	17	14	19	19	8	16	11
Australia	14	13	11	6	8	9	10.2	8
Denmark	21	21	20	16	10	10	16.3	11
Belgium	9	4	6	10	6	11	7.7	7
Netherlands	13	5	7	4	4	12	7.5	9
Austria	16	14	4	8	14	13	11.5	12
Switzerland	4	3	8	9	3	14	6.8	11
Germany	12	12	10	13	7	15	11.5	8
Portugal	18	19	21	21	15	16	18.3	6
New Zealand	10	7	17	14	13	17	13	10
Italy	5	8	15	17	17	18	13.3	13
Spain	20	18	18	18	21	19	19	3
France	17	20	16	15	18	20	17.7	5
Greece	8	11	19	20	20	21	16.5	13

Table 8. Regression of GDP per capita growth 1970-2020 on education variables, mean values

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	All countries				OECD countries				Non-OECD countries			
Initial GDP pc	-0.528*** (0.072)	-0.491*** (0.070)	-0.600*** (0.057)	-0.587*** (0.055)	-0.589*** (0.116)	-0.569*** (0.115)	-0.604*** (0.117)	-0.595*** (0.114)	-0.540*** (0.088)	-0.436*** (0.115)	-0.563*** (0.085)	-0.583*** (0.072)
Years of Schooling	0.235*** (0.028)		0.122*** (0.035)		0.078*** (0.028)		0.047 (0.029)		0.268*** (0.034)		0.158*** (0.038)	
Quality of Schooling		0.833*** (0.104)	0.604*** (0.127)			0.437*** (0.124)	0.294** (0.130)			0.981*** (0.128)	0.725*** (0.151)	
LAYS				0.304*** (0.027)				0.098*** (0.031)				0.423*** (0.040)
Physical Capital	0.323 (0.287)	0.509** (0.245)	0.292 (0.228)	0.330 (0.247)	0.923* (0.461)	0.382 (0.445)	0.461 (0.437)	0.701 (0.431)	0.239 (0.308)	0.553** (0.249)	0.242 (0.225)	0.216 (0.242)
R-squared	0.476	0.566	0.629	0.559	0.643	0.648	0.663	0.651	0.506	0.599	0.694	0.663
Observations	101	101	101	101	32	32	32	32	69	69	69	69

*** p<.01, ** p<.05, * p<.1

Data sources: As described in the text. Note: Dependent variable: Growth of GDP per capita (1970-2020). Robust standard errors are in parentheses. Education variables are computed as the mean value across 1970-2020

Table 9. Comparison of main statistics between assessments for the restricted double countries samples

Assessment 1	Nb of countries	Mean	SD	Skewness	Kurtosis	Assessment 2	Mean	SD	Skewness	Kurtosis
ELCE math	4	512.88	94.14	0.37	3.87	TIMSS math	395.39	98.80	-0.07	2.89
ELCE science	4	508.14	92.52	0.22	3.30	TIMSS science	437.96	93.43	-0.24	3.02
ELCE reading	3	529.01	85.63	0.13	2.82	PIRLS reading	475.79	79.87	-0.10	2.76
SACMEQ II, math	2	496.97	96.77	0.84	5.09	TIMSS 2003, grade 8, math	304.23	102.70	0.22	2.86
SACMEQ III, reading	1	497.87	115.04	0.57	2.92	PIRLS 2006, reading	295.32	123.46	0.52	3.23
PISA 2000, math	21	485.76	110.92	-0.30	2.91	TIMSS 1999, math	505.07	100.21	-0.28	3.10
PISA 2000, science	21	488.90	104.20	-0.12	2.73	TIMSS 1999, science	510.65	97.09	-0.24	3.22
PISA math	45	471.03	105.48	0.00	2.75	TIMSS math	496.04	104.01	-0.10	2.94
PISA science	42	473.15	103.73	0.01	2.65	TIMSS science	504.40	98.50	-0.44	3.33
PASEC II math	1	540.02	125.09	-0.19	2.16	SACMEQ III, math	619.20	135.90	0.24	2.62
PASEC II, reading	1	566.67	131.40	-0.47	2.03	SACMEQ III, reading	570.87	120.17	0.05	2.28

Table 10. Effect of PISA results on TIMSS scores for double countries

	(1)	(2)	(3)	(4)
	Mean scores			
PISA results	0.919*** (0.052)	0.939*** (0.049)	2.060*** (0.136)	2.145*** (0.131)
Difference in grades				
2 years		-6.633 (5.612)		-3.729 (2.547)
3 years		-38.189*** (9.228)		-17.122*** (3.167)
Skills dummies	Yes	Yes	Yes	Yes
Years dummies	Yes	Yes	Yes	Yes
R-squared	0.800	0.815	0.714	0.739
Observations	304	304	304	304

*** p<.01, ** p<.05, * p<.1 Notes: Cluster-robust standard errors provided in brackets. Clusters are countries. Difference in grades is calculated by the rounded mean grade in PISA and the actual grade in TIMSS. Hence, a difference of two years means that the mean grade tested in PISA is grade 10 while the grade tested in TIMSS is always grade 8. Both mathematics and science scores are included. All years are taken into account (2000, 2003, 2006, 2011, 2015 and 2018). When years of implementation are close, we group assessments. For instance, PISA 2018 results are compared with TIMSS 2019 results

Appendix A. Student Achievement Tests

Below, we describe the student achievement tests which can be compared globally and over time. We divide the assessments into three main groups: The first consists of international assessments; the second contains regional assessments while the third is a hybrid assessment. A detailed summary of these assessments is provided in Table A.1

A.1. 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 achievements for international comparison. Tests began in the early 1960s. These tests were precursors of their more current counterparts: *Trends in International Mathematics and Science Study* (TIMSS) and *Progress in International Reading Literacy Study* (PIRLS). The precursors of 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 of 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 (Elley, 1994, Campbell and Mullis, 2001). Data sources are based on major reports and raw data (Postlethwaite, Foshay et al., 1962, Bloom, 1969, Comber and Keeves, 1973, Thorndike, 1973, Peaker, 1975, Walker et al., 1976, Livingstone, 1986, Garden, 1987, IEA, 1988, Robitaille and Garden, 1989, Westbury and Travers, 1990, Burstein, 1992, Keeves, 1992).

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 et al., 2000). 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 Chinapah et al., 2000). However, much of the data has not

been published. Since microdata is sparse or often unavailable for the MLA and IAEP data, we prefer not to include these series in our dataset.

The Modern ISATs (mid 1990s onward): In the mid-1990s, standardized, psychometrically robust and relatively consistent ISATs emerged. Below we describe the major ISATs which we used 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. Six 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.³⁰ 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 fourth round was performed in 2015 and covered 63 countries/areas. The latest round was closed in 2019 and covered 64 countries/areas.

PIRLS. The other dominant IEA survey is the Progress in International Reading Literacy Study (PIRLS). Four rounds of PIRLS have been held to date: in 2001, 2006, 2011 and 2016. The PIRLS tests pupils from primary schools in grade 4 in reading proficiency.³¹ 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 third round of PIRLS was carried out with TIMSS in 2011 and included 60 countries/areas. In 2016, PIRLS was also conducted and included 50 countries/areas.

In our database, we use all recent IEA studies across two subjects (mathematics and reading/literacy). We use results from official reports (Mullis et al., 2000, Mullis et al., 2003,

³⁰ IEA assessments define populations relative to specific grades, while PISA assessments focus on the age of pupils. In IEA studies, three different groups 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.

³¹ Similar to TIMSS, pupils from Grade 4 are chosen.

Mullis et al., 2004, Mullis et al., 2008, Martin et al., 2012, Mullis et al., 2012a, Mullis et al., 2012b, Mullis, 2016a, 2016b, Mullis et al., 2017, Mullis et al., 2020) and raw data provided by the IEA Repository website: <https://www.iea.nl/data-tools/repository>.

PISA. The Organization for Economic Co-operation and Development (OECD) launched the Program for International Student Assessment (PISA) in 1997 to provide comparable data on student performance. PISA emphasizes an extended concept of “literacy” and places the emphasis on lifelong learning. Literacy is considered more broadly because PISA studies are concerned with pupils’ capacity to extrapolate from what they have learnt and apply their knowledge to novel 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.³² 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. The number of countries taking part at PISA is growing over time. In 2000, 43 countries/areas participated while in 2018, 79 countries/areas participated. Data can be retrieved from OECD website. Main results can also be found in official reports (OECD, 2000, 2003, 2004, 2007, 2010a, 2010c, 2010b, 2013, 2016, 2019c, 2019a, 2019b).

A.2. 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

³²As 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.

countries met in Paris in July 2004 and established SACMEQ. The current 16 SACMEQ education members are: Angola,³³ 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 performances 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 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 pupils' 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) spanned 2006-2011 and covered the same countries as SACMEQ II plus Zimbabwe. SACMEQ III also assessed the achievement of grade 6 pupils. The latest round of SACMEQ (SACMEQ IV) began in 2013 in 15 countries, but results have not been published. We therefore included only the first three rounds of SACMEQ in our database.

PASEC. The "Programme d'Analyse des Systèmes Éducatifs" (PASEC, or "Program 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

³³ Angola is a recent member of SACMEQ, but has not implemented any survey projects yet.

round includes ten countries. PASEC I ran 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 1994.³⁴ Moreover, data from the first four assessments are not available.³⁵ PASEC was modified significantly in 2014, rendering results hard to compare with previous PASEC items. While 10 countries took part at PASEC 2014, this number increased up to 14 in 2019,³⁶ in the second round.

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. The main aim of this survey is to garner information on pupil performance and performance-related factors likely to guide politicians in educational policymaking.

Assessments conducted by the LLECE focus on achievements 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 (Treviño, 2014). 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 (Flotts et al., 2015, Treviño et al., 2015). The fourth and latest round to date was conducted in 2019 and included 16 countries, with

³⁴ 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).

³⁵ The first four assessments were mainly pilot studies and the purpose was not to disseminate results.

³⁶ In 2014, the following countries participated to the PASEC study: Benin, Burkina Faso, Burundi, Cameroun, Congo, Côte d'Ivoire, Niger, Senegal, Chad, and Togo. In addition to 2014 participating countries, Guinea, Gabon, Mali and Madagascar joined the PASEC initiative in 2019.

Cuba the new country compared to the third round. Our analysis will include all LLECE results, since these assessments are mostly similar and cover comparable grades (UNESCO/Unicef, 2021). However, raw data for the 2019 study are not yet available (in May 2022). Therefore, we extracted results from official reports and the "Laboratory Portal" available at the following link: <https://lleceunesco.org/>.

A.3. Hybrid Standardized Achievement Tests (RSATs)

In addition to traditional international and regional student achievement tests, we also include some "hybrid tests" following the terminology of Wagner (2017).

Hybrid assessments can be considered as a mix of international and regional achievement tests. Wagner (2011) argues that EGRA represents a hybrid type of assessment. The Early Grade Reading Assessment (EGRA) is an individually administered oral assessment of the most basic foundation skills for literacy acquisition in early grades. The assessment requires about 15 minutes per child. It was designed as an inexpensive and simple diagnostic of individual student progress in reading. EGRA includes up to thirteen subtasks, such as "oral reading fluency", "vocabulary", "diction", and "reading comprehension". We compile and include data from the proportion of pupils with a 0-score in the "Oral Reading Fluency" test provided in almost all EGRA tests. We believe this is the best comparable measure since it is not related to the complexity of languages, which may differ across countries. Although EGRA has not been conducted in rounds as PISA or TIMSS have, we consider three different rounds of EGRA in years 2010, 2015 and 2019. The number of countries with comparable data is respectively 29, 40 and 5. The list of countries with data for EGRA is presented in Table A.1. Similarly to EGRA, the Annual Status of Education Report (ASER) is a quick oral assessment of early grades. It has been conducted in several countries, including India and Pakistan. ASER is the largest annual household survey carried out among citizens of India to understand whether children are enrolled in school and whether they are learning. The main advantage of ASER is that it reaches a representative sample of children from every rural district in India. Since 2006, the ASER sample size has included 30 villages per district and more than 700,000 children surveyed. Table A.2 presents the coverage of ASER over time for India. One potential advantage of ASER is the possibility to obtain results which are provided in each state within Pakistan and India. Moreover, since EGRA and ASER tests are quite similar, we were able to match scores for both tests and include results in our analysis. Data

for EGRA have been mainly extracted from raw data.³⁷ When data were not available, we used results from the EGRA Barometer, available at <https://earlygradereadingbarometer.org/>. Results for ASER India and ASER Pakistan were downloaded from their respective websites.³⁸

Table A.3 summarizes the availability and details of the various international and regional assessments listed above.

³⁷ We are very grateful to Luis Crouch and Jennifer Ryan for their support with the EGRA datasets.

³⁸ ASER India: <http://www.asercentre.org/>. ASER Pakistan: <http://www.aserpakistan.org/>. We are very grateful to Preeti Mnchanda, Wilima Wadhwa and Sahar Saeed for their support during the analysis of raw data for ASER India and ASER Pakistan surveys.

Table A.1. Review of Student Achievement Tests

<i>No</i>	<i>Year</i>	<i>Organization</i>	<i>Abbr.</i>	<i>Subject</i>	<i>Countries or Areas</i>	<i>Grade</i>	<i>Included</i>
1	1959-60	IEA	Pilot Study	M,S,R	12	7/8	
2	1964	IEA	FIMS	M	12	7/FS	■
3	1970-71	IEA	SRC	R	15	4/8/FS.	
4	1970-72	IEA	FISS	S	19	4/8/FS.	■
5	1980-82	IEA	SIMS	M	19	8/FS	■
6	1983-84	IEA	SISS	S	23	4/8/ FS	■
7	1988/1990-91	NCES	IAEP	M,S	6-19	4/7-8	
8	1990-91	IEA	RLS	R	32	3-4/7-8	■
9	1995/1999/2003/2007/2011/2015/2019	IEA	TIMSS	M,S	45-38-26-48-66-65-64	3-4/7-8/ FS	■
10	1992-97	UNESCO	MLA	M,S,R	72	6-8	
11	1997/2006/2013/2019	UNESCO	LLECE	M,S,R	13-16-15-16	3-6	■
12	1999/2002/2007	UNESCO	SACMEQ	M,R	7-15-16	6	■
13	2000/2010/2014/2019	CONFEMEN	PASEC	M,R	22-22-10-14	2/5 then 3-6	■
14	2001/2006/2011/2016	IEA	PIRLS	R	35-41-55-50	4	■
15	2000/2003/2006/2009/2012/2015/2018	OECD	PISA	M,S,R	43-41-57-74-65-71-79	Age 15	■
16	2010/2015/2019	USAID/RTI	EGRA	R	29-40-5	1 to 6	■
17	2008-2019	ASER	ASER	R	2	1 to 6	■

Note: For the meaning of abbreviations, please consult section 2. Only assessments for which there is information in the "Survey Series" column are included in our dataset. Subjects: M=math; S=science; R=reading.

Table A.2. Countries with comparable data for the EGRA/ASER tests

Country	Years available	Grades	Source
Afghanistan	2016	2;3;4;5	Raw data
Bangladesh	2014;2015;2016	2;3	Raw Data
Cambodia	2018;2019	1;2	EGRA Barometer
DRC	2010;2012;2013;2015	2;3;4;5;6	Raw Data
Egypt	2009;2013;2014	2;3;4	Raw Data
El Salvador	2017	2;3	EGRA Barometer
Ethiopia	2010;2013;2014	2;3;4	Raw Data
Gambia	2007	1;2;3	Raw Data
Ghana	2013;2015	2	Raw Data
Guyana	2008	2;3;4	Raw Data
Haiti	2009;2014	1;2;3	Raw Data
Honduras	2008	2;3;4	Raw Data
India	2008-2014; 2016;2018	1;2;3;4;5;6	ASER India
Indonesia	2012;2013;2014	2;3	Raw Data
Iraq	2012	2;3	Raw Data
Jordan	2012;2014	2;3	Raw Data
Kenya	2009;2012;2013;2016	1;2;3	EGRA Barometer (for 2016)
Kiribati	2016	1;2;3	Raw Data
Kyrgyzstan	2016;2017	2;4	Raw Data
Lao PDR	2012	3;4;5	Raw Data
Liberia	2011;2015	1;2;3	Raw Data
Macedonia	2015;2016	2;3	Raw Data
Malawi	2011;2012	2;4	Raw Data
Mali	2009;2011;2015	2;3;4	Raw Data
Mexico	2014;2015;2016	1;2;3;4;5;6	Raw Data
Morocco	2011	2;3	Raw Data
Mozambique	2013; 2016	1;2;3;4;5;6	Raw Data ; EGRA Barometer (for 2013)
Myanmar	2014;2015	1;2;3	Raw Data
Nepal	2014;2016;2018	1;2;3;4;5;6	Raw Data
Nicaragua	2008;2009	1;2;3;4	Raw Data
State of Palestine	2014	2	Raw Data
Pakistan	2012-2016;2018;2019	1;2;3;4;5;6	ASER Pakistan
Papua New Guinea	2011;2012;2013	1;2;3;4	Raw Data
Philippines	2013;2014;2015;2019	1;2;3	Raw Data ; EGRA Barometer (for 2019)
Rwanda	2011;2015;2016	4;6	Raw Data
Samoa	2017	1;2;3	Official Report
Senegal	2009	3	Raw Data
Sierra Leone	2014	2;4	Raw Data
Solomon Islands	2017	1;2;3	Official Report
Somalia	2013	2;3;4	EGRA Barometer
South Africa	2013	5	EGRA Barometer
South Sudan	2017;2018	3	Official Report
Sudan	2016	3	Raw Data
Tajikistan	2017	2;4	Raw Data
Tanzania	2013;2017	2	Raw Data
Timor-Leste	2009;2011	1;2;3	Raw Data
Tonga	2009;2014	1;2;3	Raw Data
Tuvalu	2016	1;2;3	Raw Data
Uganda	2009;2013;2014;2017	1;2;3	Raw Data
Vanuatu	2010	1;2;3	Raw Data
Yemen	2011	2;3	Raw Data
Zambia	2011;2014;2015	2;3	Raw Data ; EGRA Barometer (for 2015)

Table A.3. ASER India coverage over time

	2008	2009	2010	2011	2012	2013	2014	2016	2018
No. of districts covered	489	557	569	577	580	567	564	568	550
No. of villages covered	9,593	15,841	16,131	16,303	16,484	16,149	16,159	16,243	15,941
No. of households surveyed	192,517	322,425	320,719	337,315	331,791	333,042	328,141	331,49	325,827
No. of children surveyed (total)	330,101	762,252	723,969	703,047	679,271	655,81	631,137	595,139	566,661
Children aged 3-5		148,054	125,416	119,322	119,238	124,349	116,618	112,773	107,392
Children aged 6-14	330,101	521,116	529,889	509,188	478,901	455,769	439,168	411,519	390,838
Children aged 15-16		93,082	68,664	74,537	81,132	75,692	75,351	70,847	68,431
No. of schools observed	8,306		14,066		14,748	14,24	14,373	14,662	14,724

Source: ASER India website. Link: <http://img.asercentre.org/docs/domains2005-2018.pdf>

Annex B. Presentation of 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 the 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, the 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, 1999). This is also known as *linking* in the terminology of Holland and Dorans (2006), Linn (1993) and Mislevy (1992). As explained by Kolen and Brennan (2014), similar statistical procedures are used in linking and equating, although their purposes are different. In this paper, we use the term *linking* instead of *equating* since the tests we link 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 of a given pupil answering 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 an overlap in test items.³⁹ This is not true for a significant enough number of tests and time intervals to create a globally comparable panel dataset. Moreover, even when there is an 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.

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 target population T , takes two different assessments X and Y . Here, we suppose that any differences in the score distributions on X

³⁹ 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.

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:

$$X - \mu(X) = Y - \mu(Y) \quad (\text{B.1})$$

We then solve for y and obtain:

$$\text{linking}_Y^m(x) = y = x - \mu(X) + \mu(Y) \quad (\text{B.2})$$

In this equation, $\text{linking}_Y^m(x)$ 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:

$$\frac{x-\mu(X)}{\sigma(X)} = \frac{y-\mu(Y)}{\sigma(Y)} \quad (\text{B.3})$$

Solving for y in Eq. (3),

$$\text{linking}_Y^l(X) = y = \sigma(Y) \left[\frac{x-\mu(X)}{\sigma(X)} \right] + \mu(Y) \quad (\text{B.4})$$

where $\text{linking}_Y^l(X)$ 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 $\text{linking}_Y^l(X)$ is:

$$\text{linking}_Y^l(X) = y = \frac{\sigma(Y)}{\sigma(X)}x + \left[\mu(Y) - \frac{\sigma(Y)}{\sigma(X)}\mu(X) \right] \quad (\text{B.5})$$

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

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

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, this assumes the distribution is the same, and Eq. (3) could 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. (2018) 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, which can be considered as an exchange rate between achievement tests:

$$\text{linking}_Y^{pl}(X) = y = \frac{\mu(Y)}{\mu(X)}x \quad (\text{B.7})$$

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* and $\frac{\mu(Y)}{\mu(X)}$ can be considered as an exchange rate between the two tests. We prefer to use this hybrid approach instead of linear linking to preserve the 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 a similar approach, but adjust the coefficient with both means and standard deviations:

$$linking_Y^{pl2}(X) = y = \left[\frac{\mu(Y)}{\mu(X)} \times \frac{\sigma(Y)}{\sigma(X)} \right] x \quad (B.8)$$

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 in 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.

Equipercntile linking. Equipercntile linking was developed by Braun (1982). Equipercntile 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 scores but less difficult for low scores. The equipercntile linking function is developed by identifying scores on *Anchored Test X* that have the same percntile 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: $F(x) = P\{X \leq x | T\}$ where $P\{. | T\}$ 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: $G(y) = P\{Y \leq y | T\}$ where $P\{. | T\}$ is the probability or population proportion in each population T.

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

$$F(x) = G(y) \quad (\text{B.9})$$

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

$$\text{linking}_Y^e(X) = G^{-1}[F(x)] \quad (\text{B.10})$$

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

In summary, equipercentile linking is broken down 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 the 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, we 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 the 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

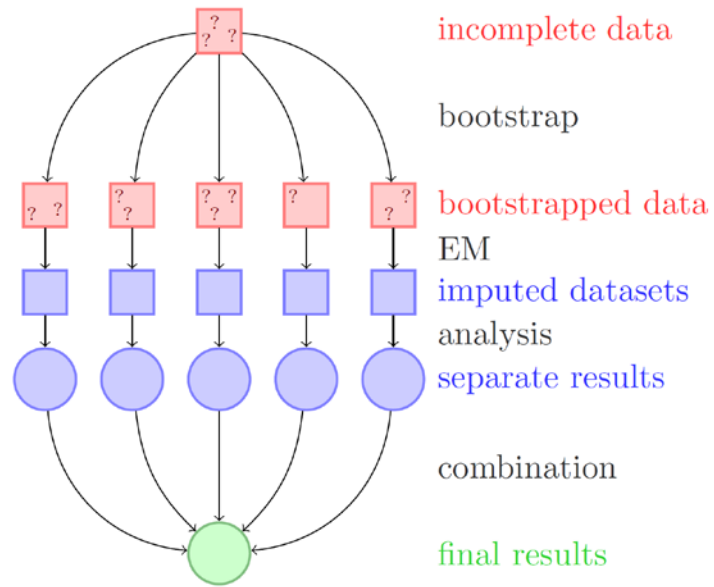
method used by the ETS, and is based on Von Davier and Yamamoto (2004) and Holland and Thayer (1987, 2000).⁴⁰

Three assumptions must hold for the linking methods above to be valid. First, they must test the same underlying population. Given that 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.

In order to obtain comparable means scores across achievement tests, we use pseudo-linear linking, also called the 'exchange rate' approach. 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.

⁴⁰ We used R Statistics software for the equipercentile linking. In particular, we use the "equate" package. See Albano (2016) for more information.

Figure B.1. A schematic approach to multiple imputation with the EMB algorithm.



Source: Honaker et al. (2011)

Table B.1. Source and definition of variables

Variable	Source	Definition
Group 1 – Proxies for learning outcomes with large missing values		
% of teachers who are female	EdStats / UIS	Number of female teachers at the primary level expressed as a percentage of the total number of teachers (male and female) at the primary level in a given school year.
% of qualified teachers	EdStats / UIS	Percentage of teachers by level of education taught (pre-primary, primary, lower secondary and upper secondary education) who have at least the minimum academic qualifications required for teaching their subjects at the relevant level in a given country, in a given academic year. A high value indicates that students are being taught by teachers who are academically well qualified in the subjects they teach.
% of trained teachers	EdStats	Trained teachers in primary/secondary/tertiary education are the percentage of primary/secondary/tertiary school teachers who have received the minimum organized teacher training (pre-service or in-service) required for teaching in a given country.
Group 2 - Proxies for learning outcomes with few missing values		
Repetition rate	EdStats	Number of repeaters in a given grade in a given school year, expressed as a percentage of enrolment in that grade the previous school year. Divide the number of repeaters in a given grade in school year t+1 by the number of pupils from the same cohort enrolled in the same grade in the previous school year t.
Pupil-teacher ratio	EdStats/ UIS	Primary school pupil-teacher ratio is the average number of pupils per teacher in primary school.
Government expenditure per student (% of GDP per capita)	WDI/UIS	Government expenditure per student is the average general government expenditure (current, capital, and transfers) per student in the given level of education, expressed as a percentage of GDP per capita.
Government expenditure on education, total (% of government expenditure)	EdStats	Total general (local, regional and central) government expenditure on education (current, capital, and transfers), expressed as a percentage of total general government expenditure on all sectors (including health, education, social services, etc.). It includes expenditure funded by transfers from international sources to government. Public education expenditure includes spending by local/municipal, regional and national governments (excluding household contributions) on educational institutions (both public and private), education administration, and subsidies for private entities (students/households and other private entities). In some instances, data on total public expenditure on education refers only to the ministry of education and can exclude other ministries that spend a part of their budget on educational activities. The indicator is calculated by dividing total public expenditure on education incurred by all government agencies/departments by total government expenditure and multiplying by 100.
Government expenditure on education, total (% of GDP)	EdStats	Total general (local, regional and central) government expenditure on education (current, capital, and transfers), expressed as a percentage of GDP. It includes expenditure funded by transfers from international sources to government. Divide total government expenditure for a given level of education (e.g. primary, secondary, or all levels combined) by GDP, and multiply by 100. A higher percentage of GDP spent on education shows a higher government priority for education, but also a higher capacity of the government to raise revenues for public spending, in relation to the size of the country's economy. When interpreting this indicator, however, one should keep in mind that in some countries, the private sector and/or households may fund a higher proportion of total funding for education, thus making government expenditure appear lower than in other countries.
Group 3 – Variables related to school enrollment		
Completion rate (%)	EdStats	The number of persons in the relevant age group who have completed the last grade of the given level of education is expressed as a percentage of the total population (in the survey sample) of the same age group. The primary completion rate is the percentage of a cohort of children or young people aged 3-5 years above the intended age for the last grade of primary education who have completed that grade. The intended age for the last grade of primary education is the age at which pupils would enter the grade if they had started school at the official primary entrance age, had studied full-time and had progressed without repeating or skipping a grade. For example, if the official age of entry into primary education is 6 years, and if primary education has 6 grades, the intended age for the last grade of primary education is 11 years. In this case, 14-16 years (11 + 3 = 14 and 11 + 5 = 16) would be the reference age group for calculation of the primary completion rate.
Gross enrolment ratio (%)	EdStats	Total enrollment in primary education, regardless of age, expressed as a percentage of the population of official primary education age. GER can exceed 100% due to the inclusion of over-aged and under-aged students because of early or late school entrance and grade repetition.
Survival rate to the last grade of primary education (%)	EdStats	Percentage of a cohort of students enrolled in the first grade of primary education in a given school year who are expected to reach the last grade of primary education, regardless of repetition. Divide the total number of students belonging to a school-cohort who reached each successive grade of primary education by the number of students in the school-cohort, i.e., those originally enrolled in the first grade of primary education,

		and multiply the result by 100. The survival rate is calculated on the basis of the reconstructed cohort method, which uses data on enrolment and repeaters for two consecutive years.
Total net enrolment rate (%)	EdStats	Total number of students of the official age group for primary education who are enrolled in any level of education, expressed as a percentage of the corresponding population. Divide the total number of students in the official school age range for primary education who are enrolled in any level of education by the population of the same age group and multiply the result by 100. The difference between the total NER and the adjusted NER provides a measure of the proportion of children in the official relevant school age group who are enrolled in levels of education below the one intended for their age. The difference between the total NER and the adjusted NER for primary education is due to enrolment in pre-primary education. The total NER should be based on total enrolment of the official relevant school age group in any level of education for all types of schools and education institutions, including public, private and all other institutions that provide organized educational programs.
Out-of-school rate for each age group (%)	EdStats	Children in the official primary school age range who are not enrolled in either primary or secondary schools. Total number of lower secondary school age adolescents who are not enrolled in lower secondary education.
Adjusted net enrollment rate (%)	WDI	Adjusted net enrollment is the number of pupils of the school-age group for primary education, enrolled in either primary or secondary education, expressed as a percentage of the total population in that age group.
Enrolment rate (%)	Lee and Lee	Primary/Secondary/Tertiary adjusted enrolment ratio (original names: "pri", "sec", "ter")
Average years of schooling, 15-64 age group	Barro and Lee (2017)	Average years of total schooling, 15+, total is the average years of education completed among people over age 15.
GDP per capita	WDI	GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2015 U.S. dollars.
Group 4 – Alternative measure for learning outcomes		
Literacy data for developing countries	Le Nestour et al. (2022)	Literacy rate for an adult with a given level of schooling (here, 5 years of schooling). Estimation based on repeated cross-sections from the Demographic Health Surveys (DHS) run by USAID and Multiple Indicator Cluster Surveys (MICS) run by UNICEF. Note: Results are disaggregated for women and men. We imputed results for both genders and then computed the mean literacy rate by supposing similar weights across genders.

Table B.2. Correlation matrix

Primary level – mathematics

	1	2	3	4	5	6	7	8	9	10	
Learning outcomes - pri - math	1	-									
Years of schooling	2	0.7969*	-								
GDP per capita	3	0.6491*	0.3855*	-							
Pupil-teacher ratio, primary	4	-0.6384*	-0.5996*	-0.4679*	-						
Survival rate, primary education	5	0.6339*	0.6576*	0.3266*	-0.6496*	-					
Adult literacy score	6	0.5353*	0.7543*	0.5627*	-0.5037*	0.3997*	-				
Adult numeracy score	7	0.5929*	0.6972*	0.5289*	-0.5776*	0.4214*	0.9609*	-			
Lim et al. data, secondary	8	0.9519*	0.7161*	0.5134*	-0.6204*	0.6093*	0.8001*	0.7714*	-		
Primary enrolment rate	9	0.2005*	0.7046*	0.2192*	-0.4093*	0.4949*	0.3887*	0.3539*	0.4010*	-	
Gov. expenditure on pri. ed. as % of GDP	10	-0.4026*	-0.0226	-0.1412*	0.0411*	-0.0009	0.1246*	0.0951*	-0.2085*	0.0845*	-

Primary level – reading

	1	2	3	4	5	6	7	8	9	10	
Learning outcomes - pri - read	1	-									
Years of schooling	2	0.8137*	-								
GDP per capita	3	0.6496*	0.3855*	-							
Pupil-teacher ratio, primary	4	-0.7728*	-0.5996*	-0.4679*	-						
Survival rate, primary education	5	0.6960*	0.6576*	0.3266*	-0.6496*	-					
Adult literacy score	6	0.7168*	0.7543*	0.5627*	-0.5037*	0.3997*	-				
Adult numeracy score	7	0.6379*	0.6972*	0.5289*	-0.5776*	0.4214*	0.9609*	-			
Lim et al. data, secondary	8	0.8949*	0.7161*	0.5134*	-0.6204*	0.6093*	0.8001*	0.7714*	-		
Primary enrolment rate	9	0.2981*	0.7046*	0.2192*	-0.4093*	0.4949*	0.3887*	0.3539*	0.4010*	-	
Gov. expenditure on pri. ed. as % of GDP	10	-0.2836*	-0.0226	-0.1412*	0.0411*	-0.0009	0.1246*	0.0951*	-0.2085*	0.0845*	-

Primary level – science

	1	2	3	4	5	6	7	8	9	10	
Learning outcomes - primary - science	1	-									
Years of schooling	2	0.7008*	-								
GDP per capita	3	0.5186*	0.3855*	-							
Pupil-teacher ratio, primary	4	-0.4751*	-0.5996*	-0.4679*	-						
Survival rate, primary education	5	0.5010*	0.6576*	0.3266*	-0.6496*	-					
Adult literacy score	6	0.4543*	0.7543*	0.5627*	-0.5037*	0.3997*	-				
Adult numeracy score	7	0.3890*	0.6972*	0.5289*	-0.5776*	0.4214*	0.9609*	-			
Lim et al. data, secondary	8	0.9049*	0.7161*	0.5134*	-0.6204*	0.6093*	0.8001*	0.7714*	-		
Primary enrolment rate	9	0.1282*	0.7046*	0.2192*	-0.4093*	0.4949*	0.3887*	0.3539*	0.4010*	-	
Gov. expenditure on pri. ed. as % of GDP	10	-0.2188*	-0.0226	-0.1412*	0.0411*	-0.0009	0.1246*	0.0951*	-0.2085*	0.0845*	-

Secondary level – mathematics

	1	2	3	4	5	6	7	8	9	10
Learning outcomes - sec - math	1	-								
Years of schooling	2	0.5890*	-							
GDP per capita	3	0.5662*	0.3855*	-						
Secondary enrolment rate	4	0.4433*	0.8292*	0.3776*	-					
Total net enrolment rate, lower secondary	5	0.3629*	0.7441*	0.3450*	0.7555*	-				
Lower secondary completion rate	6	0.1924*	0.7359*	0.3696*	0.7624*	0.8392*	-			
Adult literacy score	7	0.7573*	0.7543*	0.5627*	0.7411*	0.6734*	0.4164*	-		
Adult numeracy score	8	0.7599*	0.6972*	0.5289*	0.6810*	0.6589*	0.4041*	0.9609*	-	
Lim et al. data, secondary	9	0.9363*	0.7161*	0.5134*	0.6715*	0.5868*	0.6242*	0.8001*	0.7714*	-
Gov. expenditure on sec. ed. as % of GDP	10	0.1394*	0.3443*	0.0644*	0.2907*	0.3724*	0.3275*	0.5283*	0.5621*	0.2561*

Secondary level – reading

	1	2	3	4	5	6	7	8	9	10
Learning outcomes - sec - reading	1	-								
Years of schooling	2	0.5927*	-							
GDP per capita	3	0.5706*	0.3855*	-						
Secondary enrolment rate	4	0.5255*	0.8292*	0.3776*	-					
Total net enrolment rate, lower secondary	5	0.4042*	0.7441*	0.3450*	0.7555*	-				
Lower secondary completion rate	6	0.1609*	0.7359*	0.3696*	0.7624*	0.8392*	-			
Adult literacy score	7	0.8686*	0.7543*	0.5627*	0.7411*	0.6734*	0.4164*	-		
Adult numeracy score	8	0.7790*	0.6972*	0.5289*	0.6810*	0.6589*	0.4041*	0.9609*	-	
Lim et al. data, secondary	9	0.8610*	0.7161*	0.5134*	0.6715*	0.5868*	0.6242*	0.8001*	0.7714*	-
Gov. expenditure on sec. ed. as % of GDP	10	0.2827*	0.3443*	0.0644*	0.2907*	0.3724*	0.3275*	0.5283*	0.5621*	0.2561*

Secondary level – science

	1	2	3	4	5	6	7	8	9	10
Learning outcomes - sec - science	1	-								
Years of schooling	2	0.5562*	-							
GDP per capita	3	0.5441*	0.3855*	-						
Secondary enrolment rate	4	0.4413*	0.8292*	0.3776*	-					
Total net enrolment rate, lower secondary	5	0.3461*	0.7441*	0.3450*	0.7555*	-				
Lower secondary completion rate	6	0.2032*	0.7359*	0.3696*	0.7624*	0.8392*	-			
Adult literacy score	7	0.7496*	0.7543*	0.5627*	0.7411*	0.6734*	0.4164*	-		
Adult numeracy score	8	0.6918*	0.6972*	0.5289*	0.6810*	0.6589*	0.4041*	0.9609*	-	
Lim et al. data, secondary	9	0.9275*	0.7161*	0.5134*	0.6715*	0.5868*	0.6242*	0.8001*	0.7714*	-
Gov. expenditure on sec. ed. as % of GDP	10	0.1484*	0.3443*	0.0644*	0.2907*	0.3724*	0.3275*	0.5283*	0.5621*	0.2561*

Table B.3. Test linking architecture

Linking number	Anchored assessment	Level	Subject	Reference assessment	List of countries used for linking
1	FIMS, SIMS, FISS, SISS, IAEP, first wave of TIMSS & PIRLS, RLS	P + S	M+S+R	NAEP	USA
2	LLECE	P	M + S	TIMSS	Colombia, Chile, Honduras, El Salvador
3	LLECE	P	R	PIRLS	Colombia, Chile, Honduras
4	SACMEQ	P	M + S	TIMSS	Botswana, South Africa
5	SACMEQ	P	R	PIRLS	South Africa
6	PISA	S	M + S	TIMSS	Australia, Bulgaria, Canada, Chile, Chinese Taipei, Colombia, Czech Republic, Finland, Georgia, Hong-Kong China, Hungary, Indonesia, Israel, Italy, Japan, Jordan, Kazakhstan, Republic of Korea, Latvia, Lebanon, Lithuania, North Macedonia, Malaysia, Malta, Netherlands, New Zealand, Norway, Qatar, Romania, Russian Federation, Serbia, Singapore, Slovakia, Slovenia, Sweden, Thailand, Tunisia, Turkey, USA, UAE
7	EGRA	P	R	PIRLS	Egypt, Honduras, Indonesia
8	PASEC Round I	P	M+R	SACMEQ	Mauritius
9	PASEC Round II	P	M+R	EGRA	Senegal

Notes: For representation, we include countries used at any point in time for each test linking procedure. Since tests should be administered in adjacent years to be linked for a given round, some countries are not included in some rounds. This is especially the case with linking number 6. A more detailed architecture by year is available on request.

Table B.4. Proportion of missing values, learning outcomes data

	01 MEAN	02 PRI	03 SEC	04 MATH	05 READ	06 SCIE	07 PRI+M	08 PRI+R	09 PRI+S	10 SEC+M	11 SEC+R	12 SEC+S	Total
1970	0.46	0.52	0.65	0.53	0.49	0.62	0.63	0.55	0.73	0.66	0.74	0.65	0.60
1975	0.44	0.50	0.65	0.53	0.46	0.62	0.62	0.53	0.73	0.66	0.74	0.65	0.59
1980	0.43	0.49	0.64	0.52	0.45	0.62	0.62	0.52	0.73	0.65	0.74	0.65	0.59
1985	0.43	0.49	0.64	0.53	0.45	0.61	0.62	0.52	0.73	0.66	0.74	0.64	0.59
1990	0.33	0.41	0.57	0.44	0.38	0.53	0.57	0.47	0.68	0.58	0.67	0.57	0.52
1995	0.33	0.41	0.57	0.43	0.37	0.53	0.55	0.47	0.67	0.58	0.67	0.57	0.51
2000	0.34	0.42	0.57	0.44	0.38	0.53	0.57	0.48	0.68	0.58	0.67	0.57	0.52
2005	0.35	0.43	0.57	0.44	0.40	0.53	0.57	0.50	0.68	0.58	0.67	0.57	0.52
2010	0.34	0.42	0.57	0.44	0.38	0.53	0.57	0.48	0.68	0.58	0.67	0.57	0.52
2015	0.32	0.41	0.57	0.43	0.35	0.53	0.57	0.46	0.68	0.58	0.66	0.57	0.51
2020	0.50	0.62	0.62	0.43	0.43	0.53	0.64	0.71	0.70	0.62	0.69	0.66	0.60
Total	0.39	0.47	0.60	0.47	0.41	0.56	0.59	0.52	0.70	0.61	0.69	0.61	0.55

Note: "Mean" = mean of all existing levels and skills. "PRI" = primary level, "SEC" = secondary level, "MATH" or "M" = mathematics, "READ" or "R" = reading, "SCIE" or "S" = science.

Table B.5. Proportion of imputed values, learning outcomes data

	01 MEAN	02 PRI	03 SEC	04 MATH	05 READ	06 SCIE	07 PRI+M	08 PRI+R	09 PRI+S	10 SEC+M	11 SEC+R	12 SEC+S	Total
1970	0.85	0.88	0.77	0.83	0.84	0.79	0.84	0.87	0.78	0.76	0.71	0.77	0.82
1975	0.85	0.89	0.77	0.83	0.85	0.79	0.86	0.89	0.80	0.76	0.71	0.77	0.82
1980	0.82	0.89	0.71	0.78	0.81	0.73	0.86	0.89	0.80	0.71	0.64	0.71	0.79
1985	0.75	0.81	0.59	0.70	0.74	0.63	0.75	0.81	0.65	0.59	0.52	0.59	0.69
1990	0.72	0.73	0.68	0.67	0.70	0.60	0.66	0.71	0.55	0.68	0.63	0.68	0.67
1995	0.52	0.56	0.52	0.45	0.49	0.39	0.45	0.53	0.34	0.52	0.44	0.52	0.48
2000	0.41	0.52	0.34	0.33	0.38	0.28	0.38	0.48	0.33	0.34	0.23	0.34	0.37
2005	0.31	0.37	0.23	0.21	0.28	0.19	0.23	0.34	0.20	0.22	0.15	0.23	0.25
2010	0.11	0.22	0.08	0.07	0.11	0.04	0.15	0.18	0.11	0.06	0.03	0.08	0.11
2015	0.08	0.13	0.08	0.07	0.07	0.03	0.09	0.11	0.03	0.06	0.01	0.08	0.07
2020	0.20	0.25	0.06	0.32	0.33	0.18	0.22	0.20	0.07	0.06	0.01	0.05	0.18
Total	0.50	0.56	0.42	0.45	0.49	0.37	0.47	0.54	0.36	0.42	0.34	0.42	0.45

Note: "Mean" = mean of all existing levels and skills. "PRI" = primary level, "SEC" = secondary level, "MATH" or "M" = mathematics, "READ" or "R" = reading, "SCIE" or "S" = science.

Table B.6. Variables related to the quantity of schooling and proxies for learning outcomes

Variable	Levels of schooling	Years available	% of missing values	% of imputation
Group 1 – Proxies for learning outcomes with large missing values				
% of teachers who are female	All levels	2000-2020	20	41
% of qualified teachers	All levels	2000-2020	62	34
% of trained teachers	All levels	2000-2020	62	25
Group 2 - Proxies for learning outcomes with few missing values				
Repetition rate	P – LS	1970-2020	21	45
Pupil-teacher ratio	All levels	1970-2020	21	35
Government expenditure per student (% of GDP per capita)	P – S – T	1970-2020	18	69
Government expenditure on education, total (% of government expenditure)	-	1970-2020	23	50
Government expenditure on education, total (% of GDP)	-	1970-2020	21	45
Group 3 – Variables related to school enrollment				
Completion rate (%)	P – LS	1970-2020	11	50
Gross enrolment ratio (%)	P – LS	1970-2020	10	29
Survival rate to the last grade of primary education (%)	P	1970-2020	6	64
Total net enrolment rate (%)	P - LS	1970-2020	11	52
Out-of-school rate for each age group (%)	P – LS	1970-2020	11	53
Adjusted net enrollment rate (%)	P	1970-2020	11	55
Enrolment rate (%)	P – S – T	1970-2020	6	52
Average years of schooling, 15-64 age group	P – S – T	1970-2020	6	27
GDP per capita	-	1970-2020	13	10
Group 4 – Alternative measure for learning outcomes				
Literacy data for developing countries	P	1970-2005	77	8
Adult literacy data	A	1950-2015	0	0

Note: Levels of schooling are pre-primary (PP), primary (P), lower-secondary (LS), secondary (S), upper-secondary (US), tertiary (T), Adult (A). For average years of schooling, we also add "total schooling". Degree of imputation is lower for literacy variable, since data are not available for most developed countries and after 2005. Source of data and definitions are available in Appendix Table B.1.

Table C.1. Regression of GDP per capita growth 1970-2020 on education variables, initial values

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	All countries				OECD countries				Non-OECD countries			
Initial GDP pc	-0.449*** (0.133)	-0.407*** (0.083)	-0.551*** (0.100)	-0.438*** (0.122)	-0.657** (0.247)	-0.559** (0.233)	-0.654** (0.250)	-0.639** (0.250)	-0.520*** (0.169)	-0.246** (0.092)	-0.525*** (0.119)	-0.557*** (0.129)
Years of Schooling	0.166*** (0.051)		0.089** (0.040)		0.067 (0.041)		0.056 (0.040)		0.271*** (0.080)		0.169** (0.068)	
Quality of Schooling		0.625*** (0.122)	0.513*** (0.117)			0.228 (0.174)	0.131 (0.157)			0.825*** (0.132)	0.646*** (0.137)	
LAYS				0.217*** (0.061)				0.080 (0.050)				0.471*** (0.098)
Physical Capital	0.224 (0.159)	0.243 (0.158)	0.235 (0.149)	0.203 (0.163)	0.168 (0.294)	0.164 (0.315)	0.078 (0.295)	0.143 (0.287)	0.285* (0.142)	0.314** (0.139)	0.339*** (0.117)	0.287** (0.131)
R-squared	0.211	0.341	0.385	0.203	0.505	0.476	0.515	0.496	0.341	0.463	0.567	0.435

*** p<.01, ** p<.05, * p<.1

Data sources: As described in the text. Note: Dependent variable: Growth of GDP per capita (1970-2020). Robust standard errors are in parentheses. Education variables are computed as the initial value in 1970 or the closest year

References

- AERA, A., NCME (1999). "The Standards for Educational and Psychological Testing." *American Psychological Association*.
- Aghion, P., P. Howitt, M. Brant-Collett and C. García-Peñalosa (1998). Endogenous growth theory, MIT press.
- Albano, A. D. (2016). "equate: An R package for observed-score linking and equating." *Journal of Statistical Software*, 74(8), 1-36.
- Altinok, N. (2017). "Mind the Gap: Proposal for the standardized measure for SDG4-Education 2030 Agenda." *UIS Information Paper*, 46.
- Altinok, N., N. Angrist and H. A. Patrinos (2018). "Global data set on education quality (1965-2015)." *World Bank Policy Research Working Paper*, 8314.
- Altinok, N. and A. Aydemir (2017). "Does one size fit all? The impact of cognitive skills on economic growth." *Journal of Macroeconomics*, 53, 176-190.
- Altinok, N., C. Diebolt and J.-L. Demeulemeester (2014). "A new international database on education quality: 1965–2010." *Applied Economics*, 46(11), 1212-1247.
- Altinok, N. and H. Murseli (2007). "International database on human capital quality." *Economics Letters*, 96(2), 237-244.
- Angrist, N., S. Djankov, P. K. Goldberg and H. A. Patrinos (2021). "Measuring human capital using global learning data." *Nature*, 592(7854), 403-408.
- Angrist, N., H. A. Patrinos and M. Schlotter (2013a). An expansion of a global data set on educational quality: a focus on achievement in developing countries, The World Bank.
- Angrist, N., H. A. Patrinos and M. Schlotter (2013b). "An Expansion of a Global Data Set on Educational Quality: A Focus on Achievement in Developing Countries." *The World Bank: Policy Research Working Papers*.
- Arellano, M. and O. Bover (1995). "Another look at the instrumental variable estimation of error-components models." *Journal of econometrics*, 68(1), 29-51.
- ASER (2021). Annual Status of Education Report 2021. New Dehli, ASER Centre.
- Avvisati, F. (2021). How much do 15-year-olds learn over one year of schooling?, OECD Publishing.
- Barro, R. J. and J. W. Lee (2013). "A new data set of educational attainment in the world, 1950–2010." *Journal of development economics*, 104, 184-198.
- Bloom, B. S. (1969). Cross-National Study of Educational Attainment: Stage I of the IEA Investigation in Six Subject Areas. Volume II. Washington DC, The IEA.
- Blundell, R. and S. Bond (1998). "Initial conditions and moment restrictions in dynamic panel data models." *Journal of econometrics*, 87(1), 115-143.

Bold, T., D. Filmer, G. Martin, E. Molina, B. Stacy, C. Rockmore, J. Svensson and W. Wane (2017). "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa." *Journal of Economic Perspectives*, 31(4), 185-204.

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

Burstein, L. (1992). The IEA study of mathematics III: Student growth and classroom processes, Elsevier.

Campbell, K. and M. Mullis (2001). Framework and specifications for PIRLS assesment. Sainsbury.

Casassus, J., J. Froemel, J. Palafox and S. Cusato (1998). "First International Comparative Study of Language, Mathematics, and Associated Factors in Third and Fourth Grade." *Santiago, Chile: Latin American Laboratory for Evaluation of the Quality of Education*.

Casassus, J., J. Froemel, J. Palafox and S. Cusato (2002). First International Comparative Study of Language, Mathematics, and Associated Factors in Third and Fourth Grade. Second Report. C. L. A. L. f. E. o. t. Q. o. E. Santiago.

Chinapah, V., E. M. H'ddigui, A. Kanjee, W. Falayajo, C. O. Fomba, O. Hamissou, A. Rafalimanana and A. Byomugisha (2000). With Africa for Africa. Towards Quality Education for All. 1999 MLA Project, ERIC.

Comber, L. C. and J. P. Keeves (1973). Science education in nineteen countries. New York, IEA.

Das, J. and T. Zajonc (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. and R. Doménech (2021). "Cross-country data on skills and the quality of schooling: a selective survey." *Available at SSRN 3975955*.

Demirgüç-Kunt, A. and I. Torre (2022). "Measuring human capital in middle income countries." *Journal of Comparative Economics*.

Diebolt, C. (2016). Cliometrica after 10 years: definition and principles of cliometric research, Springer. 10, 1-4.

Diebolt, C. and M. Hauptert (2019a). Handbook of Cliometrics. Springer References. Berlin, Springer Nature, 1768.

Diebolt, C. and M. Hauptert (2019b). "Measuring success: Clio and the value of database creation." *Annali della Fondazione Luigi Einaudi*, 53(2), 59-79.

Diebolt, C. and M. Hauptert (2020). "How Cliometrics has Infiltrated Economics—and Helped to Improve the Discipline." *Annali della Fondazione Luigi Einaudi*, 54(1), 219-230.

Diebolt, C. and M. Hauptert (2021). *Cliometrics: Past, Present, and Future*. Oxford Research Encyclopedia of Economics and Finance. Oxford, Oxford University Press.

Diebolt, C. and M. Hauptert (2022a). "Cliometrics and the Future of Economic History." *Essays in Economic & Business History*, 40, 1-20.

Diebolt, C. and M. Hauptert (2022b). *The Role of Cliometrics in History and Economics*. Bloomsbury History: Theory and Method. London, Bloomsbury Publishing.

Dunn, L. (1959). *Series of plates for the Peabody picture vocabulary test*. Minneapolis, American Guidance Service, Inc.

Elley, W. B. (1994). The IEA study of reading literacy: Achievement and instruction in thirty-two school systems, Pergamon Press.

Filmer, D., H. Rogers, N. Angrist and S. Sabarwal (2020). "Learning-adjusted years of schooling (LAYS): Defining a new macro measure of education." *Economics of Education Review*, 77, 101971.

Flotts, M. P., J. Manzi, D. Jiménez, A. Abarzúa, C. Cayuman and M. J. García (2015). *Informe de resultados TERCE: logros de aprendizaje*. Santiago, OREALC.

Foshay, A. W., R. L. Thorndike, F. Hotyat, D. A. Pidgeon and D. A. Walker (1962). Educational Achievements of Thirteen-year Olds in Twelve Countries: Results of an International Research Project, 1959-61, Unesco.

Gakidou, E., K. Cowling, R. Lozano and C. J. L. Murray (2010). "Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis." *The lancet*, 376(9745), 959-974.

Garden, R. A. (1987). "The second IEA mathematics study." *Comparative Education Review*, 31(1), 47-68.

Glawe, L. and H. Wagner (2022). "Is schooling the same as learning?—The impact of the learning-adjusted years of schooling on growth in a dynamic panel data framework." *World Development*, 151, 105773.

Gustafsson, M. A. (2014). Education and country growth models. Doctoral dissertation, Stellenbosch University.

Hanushek, E. A. (2002). "Publicly provided education." *Handbook of public economics*, 4, 2045-2141.

Hanushek, E. A. and D. D. Kimko (2000). "Schooling, labor-force quality, and the growth of nations." *The American Economic Review*, 90(5), 1184-1184.

Hanushek, E. A. and L. Woessmann (2007). "The Role of School Improvement in Economic Development." *NBER Working Papers*(12832).

Hanushek, E. A. and L. Woessmann (2008). "The role of cognitive skills in economic development." *Journal of economic literature*, 46(3), 607-668.

Hanushek, E. A. and L. Woessmann (2012). "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation." *Journal of Economic Growth*, 17(4), 267-321.

Hanushek, E. A. and L. Woessmann (2015). The knowledge capital of nations: Education and the economics of growth, MIT Press.

Holland, P. W. and N. J. Dorans (2006). "Linking and equating." *Educational measurement*, 4, 187-220.

Holland, P. W. and M. Hoskens (2003). "Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test." *Psychometrika*, 68(1), 123-149.

Holland, P. W. and D. T. Thayer (1987). "Notes on the use of log-linear models for fitting discrete probability distributions." *ETS Research Report Series*, 1987(2), i-40.

Holland, P. W. and D. T. Thayer (2000). "Univariate and bivariate loglinear models for discrete test score distributions." *Journal of Educational and Behavioral Statistics*, 25(2), 133-183.

Honaker, J., G. King and M. Blackwell (2011). "Amelia II: A program for missing data." *Journal of statistical software*, 45(1), 1-47.

IEA (1988). *Science Achievement in Seventeen Countries. A Preliminary Report*. Oxford, IEA.

Keeves, J. P. (1992). Learning Science in a Changing World. Cross-National Studies of Science Achievement: 1970 to 1984. The Hague, IEA.

Kolen, M. J. (1984). "Effectiveness of analytic smoothing in equipercentile equating." *Journal of Educational Statistics*, 9(1), 25-44.

Kolen, M. J. and R. L. Brennan (2014). Test equating, scaling, and linking: Methods and practices, Springer Science & Business Media.

Le Nestour, A., L. Moscoviz and J. Sandefur (2022). *The Long-Run Decline of Education Quality in the Developing World. CGD Working Paper 608*. C. f. G. Development. Washington, DC.

Lee, J. W. and R. J. Barro (2001). "Schooling quality in a cross-section of countries." *Economica*, 68(272), 465-488.

Lim, S. S., R. L. Updike, A. S. Kaldjian, R. M. Barber, K. Cowling, H. York, J. Friedman, R. Xu, J. L. Whisnant and H. J. Taylor (2018). "Measuring human capital: a systematic analysis of 195 countries and territories, 1990–2016." *The Lancet*, 392(10154), 1217-1234.

Linn, R. L. (1993). "Educational assessment: Expanded expectations and challenges." *Educational evaluation and policy analysis*, 15(1), 1-16.

Livingstone, I. (1986). Second international mathematics study: Perceptions of the intended and implemented mathematics curriculum, Office of Educational Research and Improvement, US Department of Education

Lucas, R. (1988). "On the mechanics of economic development." *Journal of Monetary Economics*, 22(1), 3-42.

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

Martin, M. O., I. V. Mullis, P. Foy and G. M. Stanco (2012). TIMSS 2011 International Results in Science. Boston College, TIMSS International Study Center.

McEwan, P. J. (2015). "Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments." *Review of Educational Research*, 85(3), 353-394.

Mislevy, R. J. (1992). "Linking Educational Assessments: Concepts, Issues, Methods, and Prospects."

Mourshed, M., C. Chijioke and M. Barber (2010). "How the world's best performing school systems keep getting better." *London: McKinsey*.

Mullis, I. V., M. O. Martin and P. Foy (2008). "TIMSS 2007 international mathematics report. Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades." *International Mathematics Report. TIMSS&PIRLS International Study Center, Boston College*.

Mullis, I. V., M. O. Martin, P. Foy and A. Arora (2012a). TIMSS 2011 international results in mathematics, ERIC.

Mullis, I. V., M. O. Martin, P. Foy and K. T. Drucker (2012b). PIRLS 2011 International Results in Reading, ERIC.

Mullis, I. V., M. O. Martin, P. Foy and M. Hooper (2017). PIRLS 2016 International Results in Reading. Amsterdam, The Netherlands, International Association for the Evaluation of Educational Achievement.

Mullis, I. V., M. O. Martin, E. J. Gonzalez and S. J. Chrostowski (2004). TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades, ERIC.

Mullis, I. V., M. O. Martin, E. J. Gonzalez, K. D. Gregory, R. A. Garden, K. M. O'Connor, S. J. Chrostowski and T. A. Smith (2000). TIMSS 1999 international mathematics report, Boston: International Study Center, Lynch School of Education, Boston College.

Mullis, I. V., M. O. Martin, E. J. Gonzalez and A. M. Kennedy (2003). PIRLS 2001 international report, International Association for the Evaluation of Educational Achievement.

Mullis, I. V. M., M.O., Foy, P. & Hooper, M. (2016a). TIMSS 2015 International Results in Mathematics. Boston.

Mullis, I. V. M., M.O., Foy, P. & Hooper, M. (2016b). TIMSS 2015 International Results in Science. Boston.

Mullis, I. V. S., M. O. Martin, P. Foy, D. L. Kelly and B. Fishbein (2020). "TIMSS 2019 international results in mathematics and science." Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <https://timssandpirls.bc.edu/timss2019/international-results>.

National Academies of Sciences, E. and Medicine (2017). Evaluation of the achievement levels for mathematics and reading on the National Assessment of Educational Progress, National Academies Press.

Nelson, R. R. and E. S. Phelps (1966). "Investment in humans, technological diffusion, and economic growth." *The American economic review*, 56(1/2), 69-75.

OCDE (2021). Education at a Glance 2021 : OECD Indicators. Paris, Editions OCDE.

OECD (2000). Measuring student knowledge and skills: The PISA 2000 assessment of reading, mathematical and scientific literacy. Paris, OECD Publishing.

OECD (2003). Literacy skills for the world of tomorrow: further results from PISA 2000. Paris, OECD Publishing.

OECD (2004). Learning for Tomorrow's World: First Results from PISA 2003. Paris, OECD Publishing.

OECD (2007). PISA 2006: Science Competencies for Tomorrow's World. Paris, OECD Publishing.

OECD (2010a). PISA 2009 results: learning trends: changes in student performance since 2000 (vol. v), OECD, Paris, France.

OECD (2010b). PISA 2009 results: Overcoming social background: Equity in learning opportunities and outcomes (Volume II). Paris, OECD Publishing.

OECD (2010c). PISA 2009 Results: What Students Know and Can Do: Student Performance in Reading, Mathematics and Science (Volume I). Paris, OECD Publishing.

OECD (2013). PISA 2012 Results: What Students Know and Can Do (Volume I, Revised edition, February 2014), OECD Publishing.

OECD (2016). PISA 2015 Results (Volume I), OECD Publishing.

OECD (2019a). PISA 2018 Results (Volume I). Paris, OECD Publishing.

OECD (2019b). PISA 2018 Results (Volume II). Paris, OECD Publishing.

OECD (2019c). PISA 2018 Results (Volume III). Paris, OECD Publishing.

Peaker, G. F. (1975). An Empirical Study of Education in Twenty-One Countries: A Technical Report. International Studies in Evaluation VIII. Stockholm, Almqvist and Wiksell

Postlethwaite, N. School organization and student achievement.

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

Raven, J. C. (1936). "Mental tests used in genetic, The performance of related individuals on tests mainly educative and mainly reproductive." *MSC thesis Univ London*.

Robitaille, D. F. and R. A. Garden (1989). The IEA study of mathematics II: Contexts and outcomes of school mathematics, Pergamon.

Ross, K. and T. Postlethwaite (1991). Indicators of the quality of education: a study of Zimbabwean primary schools, Harare: Ministère de l'Éducation et de la Culture.

Sandefur, J. (2016). "Internationally Comparable Mathematics Scores for Fourteen African Countries."

Sreekanth, Y. (2015). What Students of Class V Know and Can Do: A Summary of India's National Achievement Survey, Class V (Cycle 4), 2015. NCERT. New Delhi, NCERT New Delhi.

Thorndike, R. L. (1973). Reading Comprehension Education in Fifteen Countries: An Empirical Study. I. International Studies in Evaluation. Stockholm, Almqvist and Wiksell.

Treviño, E. (2014). Factors associated with student achievement. Results of the Second Comparative and Explanatory Regional Study (SERCE). Santiago.

Treviño, E., P. Fraser, A. Meyer, L. Morawietz, P. Inostroza and E. Naranjo (2015). Informe de resultados TERCE: factores asociados. Santiago, OREALC.

UNESCO (2019). How Fast Can Levels of Proficiency Improve? Information paper. UNESCO. Montreal, 28.

UNESCO/Unicef (2021). Los aprendizajes fundamentales en América Latina y el Caribe. Evaluación de logros de los estudiantes. Estudio Regional Comparativo y Explicativo (ERCE 2019). Santiago, Chile.

Uwezo (2014). Are Our Children Learning? Literacy and Numeracy Across East Africa. Nairobi, Uwezo East Africa.

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

Vos, T., A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, R. S. Abdulkader, A. M. Abdulle, T. A. Abebo and S. F. Abera (2017). "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016." *The Lancet*, 390(10100), 1211-1259.

Wagner, D. A. (2011). Smaller, quicker, cheaper: Improving learning assessments for developing countries, UNESCO, International Institute for Educational Planning.

Wagner, D. A. (2017). "Children's Reading in Low-Income Countries." *The Reading Teacher*, 71(2), 127-133.

Walker, D. A., A. C. Arnold and R. M. Wolf (1976). "The IEA six subject survey: an empirical study of education in twenty-one countries." *International Studies in Evaluation; IX*.

Wechsler, D. (1949). Wechsler intelligence scale for children : manual. New York, Psychological Corp.

Westbury, I. and K. Travers (1990). Second International Mathematics Study. Urbana, University of Illinois.

World Bank, T. (2018). World development report 2018: Learning to realize education's promise, The World Bank.

Wu, M. (2010). "Comparing the Similarities and Differences of PISA 2003 and TIMSS." *OECD Education Working Papers*(32), 0_1.