Policy Research Working Paper 8742

Measuring Human Capital

Noam Angrist
Simeon Djankov
Pinelopi K. Goldberg
Harry A. Patrinos

WORLD BANK GROUP
Education Global Practice
&
Development Economics
Office of the Chief Economist
February 2019

Policy Research Working Paper 8742

Abstract

Students around the world are going to school but many of them are not learning—an emerging gap in human capital formation. To understand this gap, this paper introduces a new global data set measuring learning in 164 countries. The data cover 98 percent of the world's population from 2000 to 2017. The data set will be publicly available and updated at regular intervals by the World Bank and is designed to serve as a public good to accelerate global policy and research agendas focused on quality education and human capital formation. The paper presents several motivating facts in a first application of the data: (a) although enrollment has increased worldwide, learning progress is more limited; (b) girls outperform boys on learning—a positive gender gap—in contrast to a negative gender gap observed for schooling; (c) human capital when measured by both schooling and learning accounts for between a fifth to half of cross-country income differences—a middle ground in the recent development accounting literature and (d) average estimates mask important underlying heterogeneity by country income status and region. These stylized facts demonstrate the potential of this new global dataset to reveal insights into the relationship between human capital and economic development.

This paper is a product of the Education Global Practice and the Office of the Chief Economist, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at hpatrinos@worldbank.org.

Measuring Human Capital

NOAM ANGRIST, SIMEON DJANKOV, PINELOPI K. GOLDBERG,
AND HARRY A. PATRINOS*

Updated October 2020

JEL Codes: I20, O40, O15, H40, H52, J24, P50

*Angrist: University of Oxford, and World Bank, 1818 H Street NW, Washington, DC 20433 (e-mail: noam.angrist@bsg.ox.ac.uk); Djankov: London School of Economics and the Peterson Institute for International Economics (e-mail: s.djankov@lse.ac.uk); Goldberg: Yale University, and the Peterson Institute for International Economics (e-mail: penny.goldberg@yale.edu); Patrinos: World Bank 1818 H Street NW, Washington, DC 20433 (e-mail: hpatrinos@worldbank.org). We thank the referees for revisions which improved the paper. We are grateful to Syedah Aroob Iqbal and Husein Abdul-Hamid for research support. This work builds on co-authored work with Nadir Altinok. Particular thanks to Aart Kraay for detailed comments and contributions to the methodology. Valuable comments were provided by Eva L. Baker, Felipe Barrera-Osorio, Eduardo Cascallar, Paul Collier, Stefan Dercon, Deon Filmer, Roberta Gatti, Rachel Glennerster, Daniel Koretz, Julien Labonne, Silvia Montoya, George Psacharopoulos, Simon Quinn, Heiner Rindermann, Halsey Rogers, Jaime Saavedra, Shwetlena Sabarwal, and Eduardo Velez. This paper benefited from seminars held at the World Bank, World Congress of Cliometrics, American Economic Association, IDB, Oxford, USAID and FHI360. A great number of individuals and organizations supplied us with data. A special thanks to Rebecca Rhodes for access to the microdata for many EGRA learning assessments. The views expressed here are those of the authors and should not be attributed to the World Bank.

I. INTRODUCTION

The notion of human capital – resources imbedded in people – was alluded to as early as 1776 by Adam Smith and formalized two centuries later by Becker (1962). Ever since, the literature has explored the role of human capital in economic development. For decades, this literature proxied human capital and education with measures of schooling.[2] This applies even to the most prominent index of human capital to date, the United Nation's Human Development Index (HDI).

However, proxying human capital with schooling assumes that being in school translates into learning. Evidence suggests that this is often not the case (Pritchett 2013). A recent analysis reveals that six out of ten adolescents worldwide cannot meet basic proficiency levels in math and reading (UNESCO 2017). The gap between schooling and learning is acute in developing countries. In Kenya, Tanzania, and Uganda three-quarters of grade 3 students cannot read a basic sentence such as "the name of the dog is Puppy." In rural India, half of grade 3 students cannot solve a two-digit subtraction problem such as 46 minus 17 (World Bank, World Development Report 2018).

These stylized facts demonstrate a gap in human capital formation: students are in school but learning is limited.[3] Closing this gap presents a significant margin to drive economic development. Several papers have argued that when human capital is measured by schooling it fails to deliver the returns predicted by growth models. However, when measured by learning, human capital is more strongly associated with growth[4].

To date, much of the analysis of human capital when measured by learning has focused on advanced economies. This is due to the absence of comparable measures of learning in developing countries. This excludes a significant portion of the global distribution, in particular countries with the most potential to gain from human capital accumulation.

In this paper, we bridge this gap. We introduce a database of globally comparable learning outcomes for 164 countries covering 98 percent of the global population from 2000 to 2017.[5] This is the largest and most current global learning database, one of the first to disaggregate learning results by gender, and introduces methodological improvements, such as inclusion of standard errors to quantify uncertainty. The database will be updated at regular intervals and made available

[2] Examples include Mincer (1984), Mankiw, Romer, and Weil (1992), and Lutz and Samir (2011).
[3] We refer to 'schooling' as measured by enrollment or average years of school and 'learning' measured by the stock of cognitive skills on basic proficiencies including mathematics, reading and science.
[4] See related papers by Krueger and Lindahl (2001); Pritchett (2006); Hanushek and Woessmann (2012a).
[5] There are two administrative regions: Hong Kong and Macao, which we refer to as countries for simplicity.

for public use. We hope this database is a public good that will enable tracking of human capital formation and deepen understanding of the factors driving human capital formation and economic development. A large-scale effort using this database to track and understand human capital formation is the World Bank's new Human Capital Index.[6]

II. The New Database

The database includes 164 countries from 2000 to 2017 and was produced through a large-scale effort by the World Bank to identify, collect and collate student assessment data worldwide. We include seven assessment regimes total, including three international tests and three regional standardized achievement tests. We also include the Early Grade Reading Assessment (EGRA), which adds 48 countries to the database with at least one data point in the past 10 years, including large developing economies such as Bangladesh, Nigeria and Pakistan. Each test covers between 10 and 72 countries. By combining these assessments and making them comparable we include countries which represent 98 percent of the global population. The supplement includes a detailed description of the methodology we use to develop harmonized learning measures and all data included in the database.

The final database includes mean scores as well as standard errors for each measure to quantify uncertainty. Scores are disaggregated by schooling level (primary and secondary), subject (reading, math and science) and gender (male and female). We include year-by-year data. The database will be made publicly available by the World Bank and will be updated regularly. We do not extend the time series prior to 2000 since data quality is low and this does not significantly affect country coverage, with an addition of just two territories.

Several statistics demonstrate the coverage and detail of the database. Table I presents coverage for country-year observations by region. The database includes 2134 observations across all countries from 2000-2017. Disaggregation by gender is available for nearly all the data, with 2,105 country-year observations. Most data come from math scores, with 768 country-year observations, followed by science scores, with 690 and lastly by science scores, with 676. A third of scores are primary school scores, and two-thirds of observations are secondary school scores. Latin America

[6] The World Bank Human Capital Index includes additional measures of human capital, such as combining measures of school enrollment with learning, as well as including measures of survival and health (Kraay 2019).

and the Caribbean and Sub-Saharan Africa make up nearly a quarter of all available data. This provides the largest representation of developing countries to date in a learning database.

Our methodology leverages the growth of international assessments to construct globally comparable learning outcomes. These tests are derived from assessments conducted in the United States since the 1960s such as the SAT and the National Assessment of Educational Progress (NAEP). The tests are psychometrically designed, standardized assessments of cognitive skills. Since the 1990s, international assessments have been conducted by organizations such as the OECD. Two high profile examples are PISA and TIMSS which covered 71 and 65 countries in 2015. These assessments enable credible global comparison of learning across countries and over time. However, to date most analyses of these assessments focus mainly on OECD countries and cover few developing countries.[7] This limits the distribution of countries represented and has implications for our understanding of the link between human capital and economic development.

We include 164 countries, two-thirds of which are developing countries, by linking international assessments to their regional counterparts. Regional assessments cover much of Sub-Saharan Africa and Latin America, but have often been excluded from international comparisons. We employ methods to convert a regional test score to an international test score within subjects and schooling levels (primary and secondary) and within adjacent years. By including tests across the same testing round and at the disaggregated schooling and subject level, this minimizes the likelihood that test differences are a function of time, proficiency, schooling level, or data availability and are an accurate reflection of test difficulty. We then apply this conversion to a country that participates in regional test but not an international test to produce a comparable score (referred to as a Harmonized Learning Outcome (HLO) in the database). Means are also calculated for disaggregated groups such as by gender.

The success of this approach hinges on three key assumptions. First, linked tests must capture the same underlying population. This assumption is satisfied by using sample-based assessments representative at the national level where a country participated in both a regional and international assessment. This ensures that the underlying population tested is the same on average and we capture differences between tests. Second, tests should measure similar proficiencies. To this end, we link within subjects (math, reading and science) and schooling levels (primary and secondary)

[7] Earlier studies include Barro and Lee (2001); Hanushek and Kimko (2000); Hanushek and Woessmann (2012a); Altinok, Angrist and Patrinos (2018).

to ensure overlap. Third, the linking function should capture differences between tests rather than country-specific effects. This assumption is most likely to hold the larger the number of countries which participate in a given pair of tests being linked. To maximize the likelihood this assumption holds, we construct the linking function over the entire interval. This increases the sample size used to link tests, increasing the likelihood that we capture test-specific rather than country-specific differences. In fixing the linking function, we assume that the relationship between tests stays constant across rounds. This assumption is reasonable since the mid-1990s when assessments started to use a standardized approach and to link testing rounds with overlapping test items. A related advantage of a linking function over a fixed interval is that it guarantees that any changes in test scores over this interval are due to realized progress in learning rather than changing linking functions between tests. Of note, every update of the database increases the number of countries participating in a given pair of assessments. Thus, each update expands coverage and enhances the reliability of all estimates by enabling construction of a more robust linking procedure.

We use multiple methods to link regional to international assessments. Our primary approach uses regression when multiple countries participate in assessments being compared. When only one country participates, we use linear linking. Both methods adjust test scores by a constant as well as relative standard deviations across tests. These approaches build on a literature comparing scores across different tests (Kolen and Brennan 2014) as well as a more recent work linking aggregate level scores across states in the United States (Reardon, Kalogrides and Ho 2019). In the supplement we conduct a series of sensitivity tests, including conducting the conversion using country-fixed effects or random draws of countries and time periods. We further explore additional methods, such as mean linking and ratio conversions, highlighting the tradeoffs of each approach and examining robustness across them. We find a .99 and above correlation for scores and relative ranks across all robustness tests. We also compare our data to a smaller database using Item Response Theory (IRT) where tests share common test items and find a .98 correlation.

The tests used are conducted at school. To this end, learning data might be affected by enrollment patterns, and we advise users of the data to analyze learning outcomes alongside enrollment trends. For example, average test scores might be driven by lower-performing students entering the system rather than learning progress for those who were already in school. While this is a potential concern when analyzing average scores, it is mitigated for a few reasons. First, primary enrollment rates are relatively high, reaching 90 percent on average. Second, the direction of the bias is likely to

yield a conservative upper bound of learning in a given country. Since most countries at the bottom of the distribution of learning are also those with relatively lower enrollments, it is unlikely new school entrants will alter substantive conclusions – the lowest preforming countries will be revealed to be even lower performing. In addition, data at the primary level should be largely unaffected, since at this level students are being taught basic skills, such as reading "the name of the dog is Puppy." Even if new students enter the system, these students should still be expected to attain basic skills by the time they are tested in later primary school grades.

Through construction of a cross-test conversion between international and regional assessments we quantify the difference between them, adjust for this difference, and then place learning outcomes from regional assessments on a global scale. For a high-performance benchmark on the upper end of the distribution, we use the TIMSS benchmark of 625. For the low-performance benchmark on the lower end of the distribution, we use 300, which is the equivalent on the HLO scale of the minimum benchmarks on regional assessments such as LLECE and PASEC. This approach enables us to capture performance across the distribution of both international and regional benchmarks. The detailed methodology is described in the supplement.

III. Stylized Facts

Figure I presents learning outcomes for 164 countries from 2000-2017. The supplement provides a corresponding table for all countries and includes gender disaggregation. Figure I makes the global coverage of the database immediately apparent with typically excluded regions from international tests such as PISA and TIMSS included in our database. This includes the vast majority of Sub-Saharan Africa, Latin America and the Caribbean, and South Asia – economies with significant potential to close learning gaps for economic development.

A few trends emerge: advanced economies far outpace developing economies; Sub-Saharan African lags behind all regions; within Sub-Saharan Africa, a few countries such as Kenya and Tanzania lead, on par with many countries in Latin America; within Latin America, a few countries such as Chile are on par with European counterparts; the Middle East performs similarly or worse than Latin America; many Asian countries outperform North American and European counterparts, while a few South Asian countries such as India perform on par with Sub-Saharan African countries.

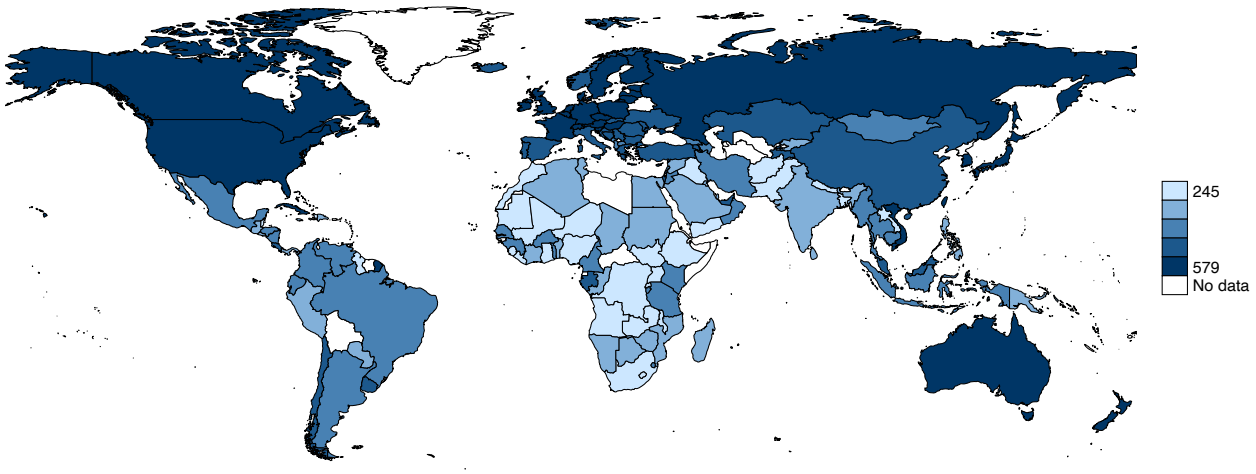


FIGURE I

Average Learning (2000-2017)

Notes: Learning levels are calculated as an average across subjects and levels over the given period of time.

Source: Our HLO learning database.

We present a series of stylized facts in a first application of the database. Figure II contrast years of schooling with learning for the most recent year in which data is available. A few trends emerge from this graph. First, the exponential shape of the line of best fit indicates that there might be a tipping point where countries realize more learning after about ten years of schooling. Second, there is high variance in learning conditional on years of schooling, with some countries reaching around 8 years of expected schooling yet less than 300 on learning, such as Nigeria, while others with similar years of schooling reach almost 400 on learning, such as Tanzania. Moreover, this graph reveals that many developing countries have successfully achieved high schooling, but have not yet realized high rates of learning. A few notable examples with high schooling but low learning include India, Brazil and Ghana. Brazil has 11.7 expected years of schooling, yet a learning score of just 411. India has 10.2 expected years of schooling, yet a learning score of just 366. Ghana has 11.6 years of expected schooling yet only a 229 score on learning. In contrast, some countries outperform the trend, with Vietnam achieving learning on the upper end of the distribution with a score of 514 on learning even without achieving the upper end of expected schooling. These trends reveal that schooling does not translate one-to-one into learning, with significant margins to improve learning both by increasing schooling as well conditional on schooling.

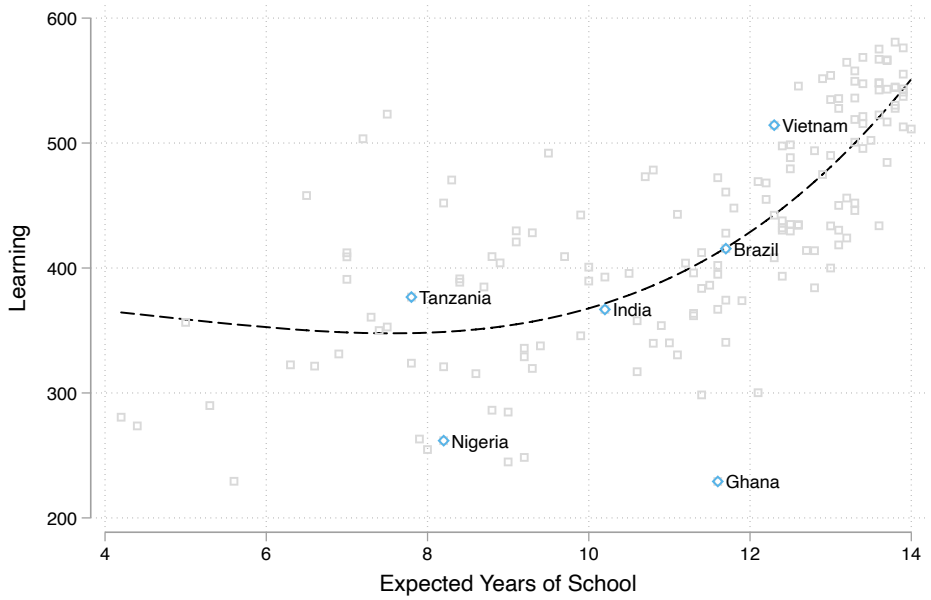


FIGURE II

Schooling versus Learning

Notes: We compare the average country learning in the latest available year from our database to expected years of schooling.

Source: Expected years of schooling are from the World Bank Human Capital Index based on data compiled by UNESCO; average learning outcomes are from our database. Both measures take the latest data available.

Figure III explores the contrast between changes in schooling and learning over time. We measure schooling using primary school enrollment rates. We compare this to our measure of learning in primary school for the years 2000-2015. We use data this period since it has the most overlap of schooling and learning measures. We restrict our comparison to countries with at least two data points over this time period for both enrollment and learning data in primary school to maximize comparability and minimize bias due to changing country coverage over the time period.

We see a clear trend towards increased schooling, while learning progress is inconsistent and in some cases stagnated. We observe limited learning progress even in regions where enrollments are relatively constant, such as Latin America and the Caribbean (LAC). We explicitly condition average learning on enrollment across countries and over time using multivariate regression and show results in Figure IV. We find learning has stagnated on average, even when conditioned on enrollment and when including country-fixed effects. Together, these data reveal a striking human capital gap: students are increasingly in school but learning progress is inconsistent and limited. This trend has been referred to as 'the learning crisis' (UNESCO 2017; World Bank 2018).

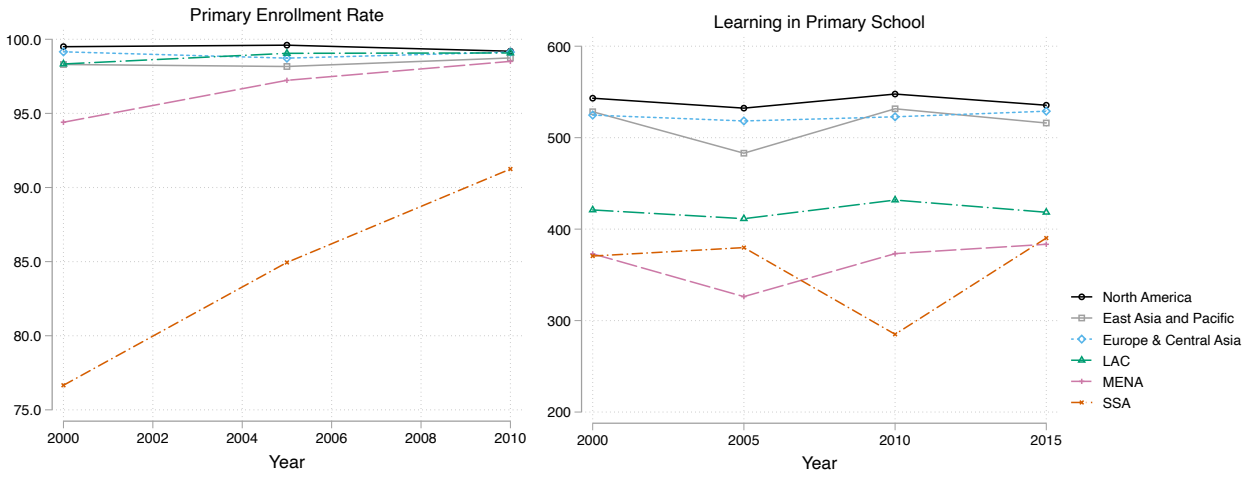


FIGURE III

Enrollment versus Learning (2000-2010), by Region

Notes: Primary enrollment and learning estimates are averaged within regions. LAC refers to Latin American and the Caribbean; MENA refers to the Middle East and North Africa; and SSA refers to Sub-Saharan Africa. We have a total of 73 countries which have both learning and enrollment data over this period and at least two data points for each.

Source: Primary enrollment rates are from Lee and Lee (2016) and are available until 2010. Learning estimates are taken from our database.

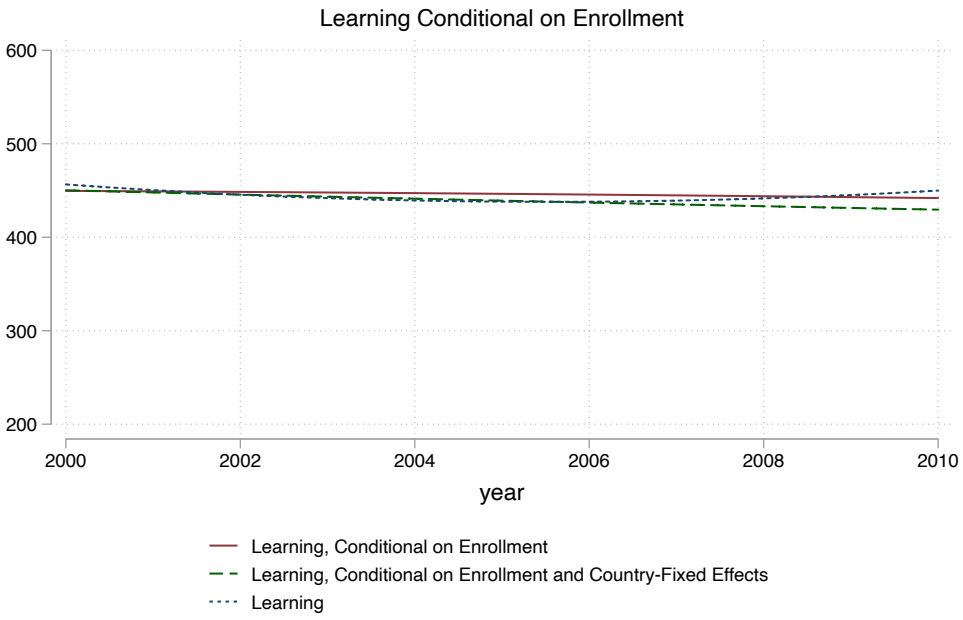


FIGURE IV

Learning Conditional on Enrollment (2000-2010)

Notes: Primary enrollment and learning estimates are averaged.

Source: Primary enrollment rates are from Lee and Lee (2016). Learning estimates are taken from our database.

Next, we explore gender gaps. We find gender gaps in learning are positive on average with girls outperforming boys across nearly all regions as shown in Figure V by region. This points in the opposite direction of the gender gap for years of schooling which is negative on average.

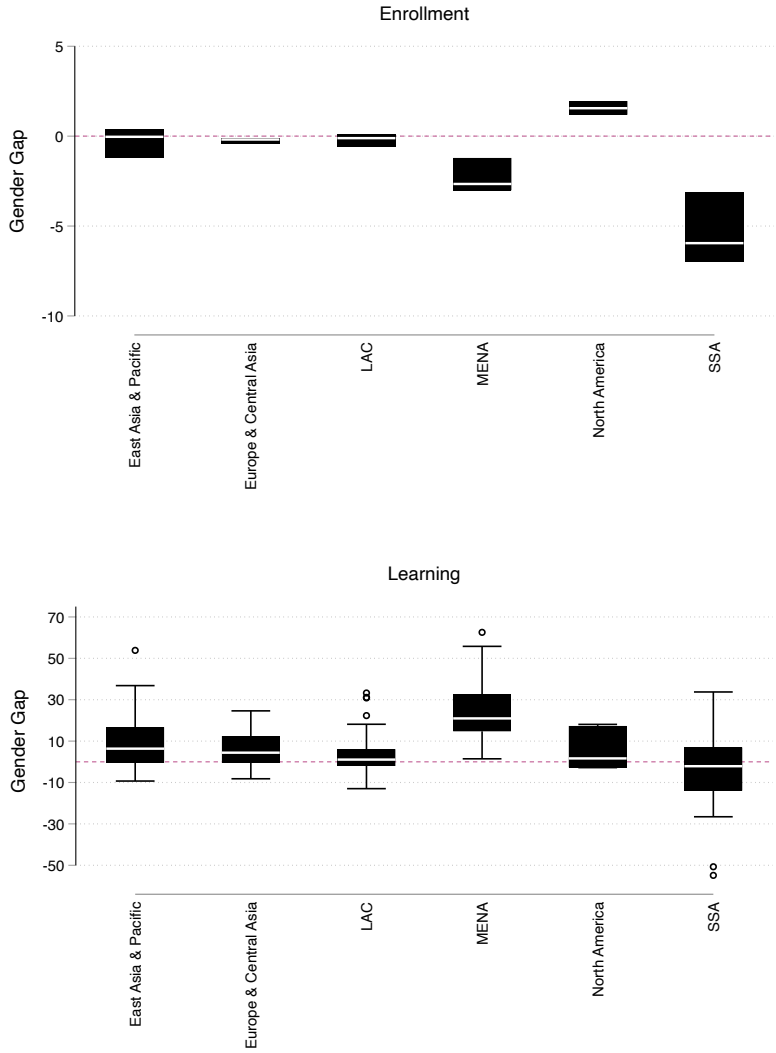


FIGURE V

Gender Gap – Enrollment versus Learning (2000-2010), By Region

Notes: LAC refers to Latin American and the Caribbean; MENA refers to the Middle East and North Africa; and SSA refers to Sub-Saharan Africa. The gender gap takes the difference of female and male enrollment or learning. A positive gender gap indicates females do better and vice-versa.

Source: Primary enrollment rates are from Lee and Lee (2016). Learning estimates are taken from our database.

This might suggest that as women increasingly join the labor market worldwide, girls who have attained schooling might realize large returns if they can obtain skilled work and might partially explain why in cross-country Mincerian returns estimates women have higher returns to schooling (Psacharopoulos and Patrinos 2018). Of note, the flip in the gender gap might be due to selection. In regions where enrollment is low, only high-achievers might be taking assessments. This explanation is consistent with trends observed for the Middle East and North Africa. However, in Sub-Saharan Africa, where enrollment is second lowest, the learning gender gap is negative, as is the enrollment gap, indicating selection is unlikely the only driver. We present the contrast in gender gaps in schooling versus learning not as definitive, but rather to motivate further in-depth exploration, which we hope this database can enable.

We next examine the relationship between human capital and economic development on a global scale. A development accounting literature studies the relative contribution of human capital in cross-country income differences. However, this question remains unsettled, in part due to difficulties in measuring human capital. While direct measures of years of schooling exist, the quality of schooling has often been inferred or measurement of quality has covered only a limited sample of countries.

Several approaches have emerged to estimate the quality of schooling, including cross-country differences in Mincerian wage returns (Hall and Jones 1999; Caselli 2005), immigrant returns (Schoellman 2011), and cross-country skill premia (Caselli and Coleman 2006). However, these approaches have encountered challenges such as the substitutability between skilled and unskilled workers (Jones 2019). The challenges in measuring quality have contributed to substantial variation in estimates of the role of human capital in accounting for cross-country income differences, ranging from nearly all to potentially none (Jones 2014; Caselli and Ciccone 2018).

In this paper, we provide a more direct and reliable measure of the quality of schooling on a global scale. We construct human capital stocks using our learning outcome data and produce development accounting results in a motivating application. We follow the literature (see Caselli (2005) for a review) and begin with a standard aggregate production function in its per-capita form following Klenow and Rodriguez-Clare (1997):

$$y = Ah\left(\frac{k}{y}\right)^{\frac{\alpha}{1-\alpha}}$$

where y represents output per worker, k denotes the capital-labor ratio, h denotes the level of human capital per capita, and A captures the residual, usually attributed to Total Factor Productivity (TFP). Taking the log on both sides decomposes cross-country income differences into three proximate sources: capital-output ratios, total factor productivity, and average human capital. Since we are only interested in the share of income differences that can be explained by variation in human capital, we provide decompositions of our baseline accounting results for the human capital share. In Table II, we provide decompositions as direct analogies to Schoellman (2011) who used inferred measures of quality as well as measures based on education quantity (Hall and Jones 1999; Hendricks 2002). In Table III we include an additional decomposition: $\frac{\ln(h_{90})-\ln(h_{10})}{\ln(y_{90})-\ln(y_{10})}$ which provides direct comparisons with a literature using various quality measures.

To measure human capital, we extend the standard Mincer specification that weights education by its micro labor market returns to consider learning as well as schooling:

$$h = e^{rS+wL} \tag{5}$$

where S is the quantity of schooling and L is a measure of learning, and r and w are their respective returns. For years of schooling, we use Barro-Lee (2010) data. For learning measures, we use the data presented in this paper. We assume rates of return based on the microeconomic literature. We take the value $r = .10$ for the rate of return per school year, and $w = .20$ per standard deviation increase in learning.[8] The .20 value is based on U.S. data. However, we might expect that returns to skills will be higher in developing countries, where the supply of skills is lower, as is the case in the returns to schooling literature. Significant work has been done to identify this parameter value. For the purpose of this paper, our intention is not to provide a final result, but rather to motivate the use of the data for future use in the development accounting literature. To this end, we take parameter values as given and conduct sensitivity analyses with values $w = .15$ and $w = .25$. We include 131 countries in this development accounting exercise.[9]

Table II shows our results in comparison to Schoellman (2011), Hall and C. Jones (1999) and Hendricks (2002). We find that when our human capital measure only captures quantity ($w = 0$), the share of human capital accounts for roughly 9-26 percent of output per worker differences.

[8] These values are based on Psacharopoulos and Patrinos (2004) and Hanushek and Zhang (2009) respectively.
[9] This includes all countries which have both Barro-Lee data as well as learning data.

However, when we include quality, we find that this share goes up to 20–44 percent. These results suggest that measuring human capital with quality substantially increases the role of human capital in explaining cross-country output per worker differences.

In Table III, we show results using the following decomposition: $\frac{\ln(h_{90})-\ln(h_{10})}{\ln(y_{90})-\ln(y_{10})}$. We compare results to the recent literature, which varies from nearly all (Jones 2014) to potentially none (Caselli and Ciccone 2018). We find that when including our measure of quality, the share of human capital varies between 46 to 58 percent. Together with Table II, our results suggest human capital accounts for between a fifth to around half of cross-country income differences – a middle ground in a literature which ranges from zero to nearly all. These results are consistent with models of human capital capturing the role of educated entrepreneurs and more comprehensive measures of human capital including schooling, learning, and health (Gennaioli, La Porta, Lopez-de-Silanes, and Shleifer 2013; Campbell and Üngör 2020). In this development accounting exercise, our central contribution is not to provide a conclusive result, but rather to motivate the use of a direct measure of schooling quality and thus a better measure of human capital.

In Table IV we further find the average relationship between learning and income masks significant heterogeneity across countries. First, we find human capital explains between a tenth and a fifth of cross-country income differences among low-income countries but up to two-thirds among high-income countries. This suggests human capital plays a more central role as economies develop. Second, we find the income gradient is often as steep or steeper as the quantity to quality gradient, more than tripling the contribution of human capital. We find even steeper gradients by regions. For example, when measured by schooling, human capital accounts for 31 percent of cross-country income differences in Europe and just 5 percent in sub-Saharan Africa. When we include learning, this gap widens to 67 percent in Europe but just 8 percent in sub-Saharan Africa. This substantial heterogeneity reveals the importance of the inclusion a global distribution of countries covering multiple stages of economic development to account for the role of human capital.

Finally, we compare our measure of human capital to alternatives in Table V. We find that our measure of human capital has a stronger and more statistically significant association with growth than human capital measures in prominent global databases such as the Penn World Tables (PWT) and the Human Development Index (HDI). This is the case in when comparing measures on their own in columns (1)-(4). Each variable is transformed to a log scale to compare percent changes or

elasticities in comparable units. We observe that a one percent change in learning is associated with a 6.5 percent change in annual growth. In contrast, a one percent change in the other human capital measures is associated with between a 1.6 to 3.3 percent change in annual growth. Moreover, the R-squared for the learning measure is highest at .275 relative to non-learning human capital measures which range from .240 to .261. We further observe when we include variables in the same multivariate regression that the relationship between learning and growth remains statistically significant between 4.7 to 5.5 percent, whereas other human capital variables have a reduced and statistically insignificant association with growth. However, we observe the overall model fit improves when all measures are included with an R-squared that increases, although only slightly, from .275 to between .281 to .298.

To this end, we observe that our measure of human capital individually and jointly appears to have a stronger relationship with economic growth. This is likely because alterative human capital measures rely largely on years of schooling and might underestimate the role of human capital in economic development by omitting learning.[10] However, their use remains standard practice in part since these data have the broadest coverage. By constructing learning data across 164 countries we fill a key gap: broad coverage over nearly two decades and a measure of human capital with strong links to economic development.

IV. Conclusion

To understand and track human capital formation, a critical ingredient for development, there is need for globally comparable measurement of learning. The growth of international standardized achievement tests is a significant step in this direction. However, many of the countries that participate in these tests are often already rich. This limits the ability to track, compare or understand education patterns in developing countries – the countries that might have the most potential to gain from human capital formation.

We bridge this gap, constructing a globally comparable database of 164 countries from 2000-2017, representing more than 98 percent of the global population and over two-thirds of countries included are developing countries. We document a series of motivating stylized facts in a first application of the data. First, we show that global learning progress has been relatively limited to

[10] This does not mean that schooling is not useful, but might lead to growth largely through the channel of learning.

date and that there is a female learning premium. We also contribute conduct a development accounting exercise, providing a direct measure of school quality. We estimate that the role of human capital in explaining cross-country income differences ranges from a fifth to half – a middle ground in a wide-ranging literature. Moreover, we find that average estimates mask significant heterogeneity by country income status and region. This reveals the importance of including countries at all stages of economic development for understanding the role of human capital. Finally, we show that our learning database provides a measure of human capital that is more closely associated with economic growth than current education and human capital measures included in the Penn World Tables 9.0 and the Human Development Index.

This database comes at a moment when a series of global efforts have been launched to measure and track learning on a global scale. While recent modelling suggests the world is on track to achieve universal primary enrollment by 2030 (Friedman et al. 2020), if learning continues to stagnate this achievement will be partial. In recognition of this, the Sustainable Development Goals (SDGs) include a focus on learning whereas the Millennium Development Goals focused largely on schooling. In addition to the SDGs, another notable effort to measure and track learning on a global scale is the World Bank's Human Capital Index which compares countries' levels of human capital around the world (Kim 2018; Kraay 2019; World Bank 2019). This effort aims to disseminate a measure of human capital that will encourage countries to invest in the education of their people. The Human Capital Index includes learning outcomes from this database as one of its core ingredients. The database in this paper will be updated regularly and made public to enable these large-scale efforts, among others, and to advance our understanding and ability to track human capital formation and potential links to economic development.

REFERENCES

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

Barro, Robert J., and Jong-Wha Lee. "International data on educational attainment: updates and implications." *Oxford Economic Papers* 53, no. 3 (2001): 541-563.

Barro, Robert J., and Jong Wha Lee. "A new data set of educational attainment in the world, 1950–2010." *Journal of Development Economics* 104 (2013): 184-198.

Becker, Gary S. "Investment in human capital: A theoretical analysis." *Journal of Political Economy* 70, no. 5, Part 2 (1962): 9-49.

Campbell, Susanna G., and Murat Üngör. 2020. "Revisiting human capital and aggregate income differences." *Economic Modelling*.

Caselli, Francesco and Antonio Ciccone. "The Human Capital Stock: A Generalized Approach. Comment." *American Economic Review*, 109, no. 3 (2019): 1155-74

Gennaioli, Nicola, Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. "Human capital and regional development." *The Quarterly Journal of Economics* 128, no.1 (2013): 105-164.

Friedman, Joseph, Hunter York, Nicholas Graetz, Lauren Woyczynski, Joanna Whisnant, Simon I. Hay, and Emmanuela Gakidou. "Measuring and forecasting progress towards the education-related SDG targets." *Nature* (2020): 1-4.

Hanushek, Eric A. and Dennis D. Kimko. "Schooling, Labor-force Quality, and the Growth of Nations." *American Economic Review* 90, no.5 (2000): 1184-1208.

Hanushek, Eric A., and Ludger Woessmann. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17, no.4 (2012a): 267-321.

Jones, Benjamin F. "The Human Capital Stock: A Generalized Approach." *American Economic Review* 104, no.11 (2014): 3752-77.

Jones, Benjamin F. "The Human Capital Stock: A Generalized Approach: Reply." *American Economic Review* 109, no. 3 (2019): 1175-95.

Kim, Jim Yong. "The Human Capital Gap: Getting Governments to Invest in People." *Foreign Affairs* 97 (2018): 92

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

Kraay, Aart. "The World Bank Human Capital Index: A Guide." *The World Bank Research Observer* 34, no. 1 (2019): 1-33

Krueger, Alan B. and Mikael Lindahl. "Education for Growth: Why and For Whom?" *Journal of Economic Literature* 39, no.4 (2001): 1101-1136.

Lange, Glenn-Marie, Quentin Wodon and Kevin Carey. *The Changing Wealth of Nations 2018: Building a Sustainable Future.* The World Bank, 2018.

Lee, Jong-Wha, and Hanol Lee. "Human Capital in the Long Run." *Journal of Development Economics* 122 (2016): 147-169.

Lutz, Wolfgang, and K. C. Samir. "Global human capital: Integrating education and population." *Science* 333, no. 6042 (2011): 587-592.

Mankiw, N. Gregory, David Romer and David N. Weil. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics* 107, no. 2 (1992): 407-437.

Mincer, Jacob. "Human Capital and Economic Growth." *Economics of Education Review* 3, no.3 (1984): 195-205.

OECD. *PISA 2015 Technical Report.* OECD Publishing, 2015.

Psacharopoulos. George and Harry Anthony Patrinos. "Returns to investment in education: a decennial review of the global literature." *Education Economics* 26, no. 5 (2018): 445-458.

Pritchett, Lant. "Does Learning to Add Up Add Up? The Returns to Schooling in Aggregate Data." *Handbook of the Economics of Education* 1 (2006): 635-695.

Pritchett, Lant. *The rebirth of education: Schooling ain't learning*. CGD Books, 2013.

Reardon, Sean F., Demetra Kalogrides, and Andrew D. Ho. "Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale." *Journal of Educational and Behavioral Statistics* (2019).

Smith, Adam. *An Inquiry into the Nature and Causes of the Wealth of Nations (Volume One)*. London: printed for W. Strahan; and T. Cadell, 1776.

UNESCO. *More Than One-half of Children and Adolescents are not Learning Worldwide*. UIS Fact Sheet No. 46, 2017.

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

World Bank. *World Development Report 2019: The Changing Nature of Work.* Washington, DC, 2019.

TABLE I
COUNTRY-YEAR OBSERVATIONS BY DISAGGREGATION AND REGION

Region	Total	Female	Male	Math	Reading	Science	Primary	Secondary
East Asia & Pacific	360	358	358	130	100	130	99	261
Europe & Central Asia	984	984	984	343	298	343	255	729
Latin America & Caribbean	249	249	249	84	88	77	111	138
Middle East & North Africa	271	271	271	106	60	105	90	181
North America	62	62	62	22	18	22	17	45
South Asia	10	9	9	1	8	1	7	3
Sub-Saharan Africa	198	172	172	82	104	12	175	23
Total	2134	2105	2105	768	676	690	754	1380

Notes: This table presents coverage for country-year observations by region. The database includes 2134 observations across all countries from 2000-2017.

TABLE II

BASELINE ACCOUNTING RESULTS AND COMPARISON TO QUANTITY LITERATURE

	Our Estimates				Literature		
	$w = 0$	$w = .15$	$w = .20$	$w = .25$	Hall and C. Jones (1999)	Hendricks (2002)	Schoellman (2011)
h_{90}/h_{10}	1.80	2.45	2.70	3.06	2.00	2.10	4.70
$\frac{h_{90}/h_{10}}{y_{90}/y_{10}}$	0.26	0.35	0.39	0.44	0.09	0.22	0.21
$\frac{var[log(h)]}{var[log(y)]}$	0.09	0.20	0.25	0.30	0.06	0.07	0.26

Notes: y: real output per worker (2000-2010); h: human capital based on school and learning estimates (2000-2010). We convert HLO units into standard deviations based on a cross-country standard deviation of 68 to correspond to the returns to schooling parameter value which is given per standard deviation.

Source: Schooling data are from Barro-Lee (2013). GDP data are from PWT 9.0. Learning estimates are from our database. Literature estimates are derived from the referenced papers

TABLE III

BASELINE ACCOUNTING RESULTS AND COMPARISON TO QUALITY LITERATURE

	Our Estimates				Recent Literature			
	$w = 0$	$w = .15$	$w = .20$	$w = .25$	B. Jones (2014)	Hanushek and Woessmann (2012)	Hendricks and Shoellman (2017)	Caselli and Ciccone (2018)
$\frac{\ln(h_{90})-\ln(h_{10})}{\ln(y_{90})-\ln(y_{10})}$	0.31	0.46	0.51	0.58	*Nearly All*	0.51	0.62	*Potentially None*

Notes: y: real output per worker (2000-2010); h: human capital based on school and learning estimates (2000-2010). We convert HLO units into standard deviations based on a cross-country standard deviation of 68 to correspond to the returns to schooling parameter value which is given per standard deviation. We assume rates of return based on the microeconomic literature. We take the value $r = .10$ for the rate of return per school year, and $w = .20$ per standard deviation increase in learning. The .20 value is based on U.S. data. However, we might expect that returns to skills will be higher in developing countries, where the supply of skills is lower, as is the case in the returns to schooling literature. Significant work has been done to identify this parameter value. For the purpose of this paper, our intention is not to provide a final result, but rather to motivate the use of the data for future use in the development accounting literature. To this end, we take parameter values as given and conduct sensitivity analyses with values $w = .15$ and $w = .25$. When $w = 0$, our accounting de facto only includes schooling; for any value $w > 0$, we include learning as well as schooling. We include 131 countries in this development accounting exercise.

Source: Schooling data are from Barro-Lee (2013). GDP data are from PWT 9.0. Learning estimates are from our database. Literature estimates are derived from the referenced papers.

TABLE IV

HUMAN CAPITAL SHARE BY INCOME STATUS AND REGION

	Schooling	Schooling and Learning		
	$w = 0$	$w = .15$	$w = .20$	$w = .25$
Human Capital Share				
Baseline	0.09	0.20	0.25	0.30
High Income	0.18	0.41	0.52	0.65
Upper Middle Income	0.29	0.54	0.65	0.79
Lower Middle Income	0.11	0.16	0.19	0.23
Low Income	0.07	0.11	0.13	0.16
East Asia & Pacific	0.19	0.32	0.38	0.45
Europe & Central Asia	0.31	0.49	0.57	0.67
Latin America & Caribbean	0.10	0.13	0.15	0.18
Middle East & North Africa	0.10	0.19	0.23	0.28
North America	0.51	0.76	0.86	0.97
South Asia	0.35	0.47	0.52	0.56
Sub-Saharan Africa	0.05	0.06	0.07	0.08

Notes: y: real output per worker (2000-2010); h: human capital based on school and learning estimates (2000-2010). We convert HLO units into standard deviations based on a cross-country standard deviation of 68 to correspond to the returns to schooling parameter value which is given per standard deviation. We assume rates of return based on the microeconomic literature. We take the value $r = .10$ for the rate of return per school year, and $w = .20$ per standard deviation increase in learning. The .20 value is based on U.S. data. However, we might expect that returns to skills will be higher in developing countries, where the supply of skills is lower, as is the case in the returns to schooling literature. Significant work has been done to identify this parameter value. For the purpose of this paper, our intention is not to provide a final result, but rather to motivate the use of the data for future use in the development accounting literature. To this end, we take parameter values as given and conduct sensitivity analyses with values $w = .15$ and $w = .25$. When $w = 0$, our accounting de facto only includes schooling; for any value $w > 0$, we include learning as well as schooling. We include 131 countries in this development accounting exercise. We define the human capital share as $\frac{\text{var}[\log(h)]}{\text{var}[\log(y)]}$.

Source: Schooling data are from Barro-Lee (2013). GDP data are from PWT 9.0. Learning estimates are from our database. Literature estimates are derived from the referenced papers.

TABLE V

HUMAN CAPITAL AND ECONOMIC DEVELOPMENT – COMPARING MEASURES

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Human Capital - Learning	0.065***				0.047**	0.049**	0.055**	0.050**
	(0.019)				(0.023)	(0.022)	(0.023)	(0.023)
Human Capital - Penn World Tables		0.033***			0.019			0.021
		(0.011)			(0.012)			(0.036)
Human Capital - Schooling			0.016***			0.009		0.012
			(0.006)			(0.006)		(0.020)
Human Capital - HDI Education Index				0.020**			0.008	-0.022
				(0.008)			(0.009)	(0.022)
Observations	107	107	107	107	107	107	107	107
R-squared	0.275	0.261	0.255	0.240	0.291	0.290	0.281	0.298

Notes: Dependent variable: annual growth rates averaged across 2000-2010. Human Capital - Schooling refers to estimates in 2000, the beginning of the time period. Human Capital – Penn World Tables refers to the measure of human capital in the Penn World Tables. Human Capital – Learning refers to the measure of human capital in this database from 2000 onwards. Human Capital - HDI Education Index refers to the measure of education included in the HDI in the year 2000. Results exclude countries in civil war, inflation crises and with rents from natural resources above 15 percent. All independent variables are transformed to log units to derive comparable elasticities. We control for initial GDP per capita refers to levels at the beginning of the period in the year 2000 in all specifications following standard practice in the growth literature.

Source: Schooling data are from Barro-Lee (2013). HDI data are from the Human Development Index Education Index. GDP and human capital data are from PWT 9.0. Learning estimates are from our database.

SUPPLEMENT

I. METHODOLOGY AND DATA

We leverage the growth of international assessments to construct globally comparable learning outcomes. These tests are derived from assessments conducted in the United States since the 1960s such as the SAT and the National Assessment of Educational Progress (NAEP). The tests are psychometrically designed, standardized assessments of cognitive skills. Since the 1990s, international assessments have been conducted by organizations such as the OECD. Two high profile examples are PISA and TIMSS which covered 71 and 65 countries in 2015. These assessments enable credible global comparison of learning across countries and over time. Earlier analyses focus mainly on OECD countries and cover few developing countries. This limits the distribution of countries represented and has implications for our understanding of the link between human capital and economic development.

We include 164 countries, two-thirds of which are developing countries, by linking international assessments to their regional counterparts. Regional assessments cover much of Sub-Saharan Africa and Latin America. Thus, through construction of a linking procedure between international and regional assessments we quantify the difference between them, adjust for this difference, and then place learning outcomes from regional assessments on a global scale.

A. The Linking Procedure

The central intuition behind the construction of globally comparable learning outcomes is the production of a linking function between international and regional assessments. This function can be produced for countries that participate in a given pair of assessments and captures the difference in difficulty between the two assessments. This linking function can then be used to place scores for countries that only participate in regional assessments on the international scale. This enables construction of globally comparable learning outcomes.

We use multiple methods to produce globally comparable scores. Our primary approach uses regression when multiple countries participate in assessments being compared. When only one country participates, we use linear linking. Both methods adjust test scores by a constant as well as relative standard deviations across tests. These approaches build on a literature comparing scores across different tests (Kolen and Brennan 2014) as well as a more recent work linking aggregate level scores across states in the United States (Reardon, Kalogrides and Ho 2019).

The conversion can be implemented by regressing mean scores from countries that partake in a regional and international assessment to derive α and β and produce a linking function between assessments:

$$\mu_{Yi} = \alpha + \beta\mu_{Xi} + \varepsilon_i$$

where μ denotes the mean scores, *X* is a regional assessment, *Y* is an international assessment and *i* denotes countries that have scores on both assessments. We can then convert scores from

countries that only participate in regional assessment X onto an international scale Y using α and β.

The success of this approach hinges on three key assumptions. First, linked tests must capture the same underlying population. This assumption is satisfied by using sample-based assessments representative at the national level where a country participated in both a regional and international assessment. This ensures that the underlying population tested is the same on average and we capture differences between tests.

Second, tests should measure similar proficiencies. To this end, we link within subjects (math, reading and science) and schooling levels (primary and secondary) to ensure overlap.

Third, the linking function should capture differences between tests rather than country-specific effects. This assumption is most likely to hold the larger the number of countries which participate in a given pair of tests being linked. To ensure this last assumption holds, we use the same linking parameters over the entire interval. This increases the sample size used to link tests, increasing the likelihood that we capture test-specific rather than country-specific differences. In fixing the linking function over time, we assume that the relationship between tests stays constant across rounds. This assumption is reasonable since the mid-1990s when assessments started to use a standardized approach and to link testing rounds with overlapping test items. A related advantage of fixing the linking function is that it guarantees that any changes in test scores over this interval are due to realized progress in learning rather than changing relationship between tests. Of note, every update of the database increases the number of countries participating in a given pair of assessments. Thus, each update both expands coverage as well as enhances the reliability of all estimates by enabling construction of a more robust linking procedure.

Below we capture a level of precision needed to satisfy the above assumptions. We produce a linking function within subjects and schooling levels (primary and secondary) from test X to test Y:

$$\mu_{Yisl} = \alpha + \beta\mu_{Xisl} + \varepsilon_{isl}$$

where *i* is a country in the set countries that participate in both tests X and Y in a given subject s, and schooling level *l.* Scores from test X and Y are further matched by testing round. We consider tests to be in the same round if they are five years apart and optimize to have the rounds as tight as possible. Most often the time window is one to two years. In some cases, this extends to three to five years apart. In a few exceptions, we average adjacent years across one another. This minimizes the likelihood that test differences are a function of time, proficiency, schooling level, or data availability and are an accurate reflection of test difficulty.

We present a simplified and illustrative example. In 2006 Colombia and El Salvador participated in the regional test in Latin America and the Caribbean called LLECE as well as an international test, TIMSS. Thus, they have primary science scores on both assessments representative at the national level. In 2013, Chile and Honduras participated in both assessments and have primary science scores on both assessments representative at the national level. A regression for this set of countries of LLECE on TIMSS at primary level and on math scores yields an estimate β of .816

and a constant adjustment α of 15.824. We can then use this estimated relationship to convert scores from countries which only took part in regional assessments to an international scale. For example, Argentina has a score of 501.32 in primary science in 2013 on LLECE and would thus have an equivalent international score of around 425.

We can also use an alternative approach called linear linking when only one country participates in pairwise assessments. This approach uses information on within-country standard deviations and mean scores to estimate α and β as follows:

$$Y = \alpha + \beta X$$

where $\alpha = \mu_Y - \beta\mu_X$, $\beta = \frac{\sigma_Y}{\sigma_X}$, and σ denotes within-country standard deviations on test X and Y. Both methods adjust test scores by a constant as well as relative standard deviations across tests.

By producing a linking function and placing regional scores on an international scale, we are able to compare learning outcomes on a global scale. On this scale, 625 represents advanced attainment and 300 represents minimum attainment. This interpretation is derived by taking establish benchmarks already used on international and regional assessments. For the high-performance benchmark on the upper end of the distribution, we use the TIMSS benchmark of 625. For the low-performance benchmark on the lower end of the distribution, we use 300, which is the equivalent on the HLO scale of the minimum benchmarks on regional assessments such as LLECE and PASEC. This approach enables us to capture performance across the distribution and accounts for floor and ceiling effects that would be introduced by taking either international or regional benchmarks on both ends of the distribution. Supplement Table VI includes descriptions of each assessment to enable derivation of linking functions.

B. Sensitivity Tests

We conduct a series of sensitivity tests. First, we examine the degree to which linking functions are stable across countries using two approaches. For tests where we have multiple participating countries and for which we use the regression method we can also produce linking functions using country-fixed effects.

SUPPLEMENT TABLE I
SCORES USING REGRESSION WITH AND WITHOUT COUNTRY-FIXED EFFECTS

	(1)	(2)	(3)	(5)	(6)	(7)
	EGRA reading	LLECE math	LLECE reading	LLECE science	PISA math	PISA science
HLO	348.4	383.3	454.1	423.2	490.0	496.2
HLO - Country Fixed Effects	329.4	368.4	462.7	414.5	474.0	489.0
Correlation	1.000	1.000	1.000	1.000	1.000	1.000

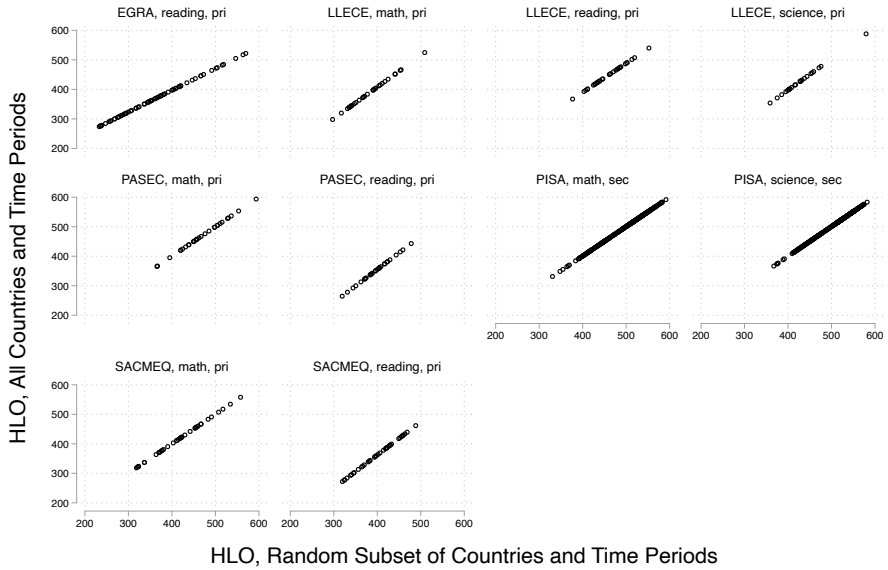
Notes: HLO references Harmonized Learning Outcomes produced with a linking function without country-fixed effects. HLO- Country Fixed Effects refers to HLO scores produced from with a linking function derived using a

regression which includes country-fixed effects. We only compute scores using the regression method for LLECE, EGRA and PISA since SACMEQ and PASEC only have a single country used to make score comparisons.

Supplement Table I compares scores with and without country-fixed effects linking methods. We observe differences in scores ranging from 10 to 20 points. These differences are smaller for tests with more overlapping countries, with the smallest differences for PISA, followed by LLECE and then EGRA. This supports the approach of fixing the linking function across countries and over time, since more overlapping countries and time periods maximize the likelihood that linking parameters are a function of relative difficulty rather than country factors. Overall, while we observe differences in scores, they are relatively small, and we find a perfect correlation among scores within test and subject.

We further test the robustness of linking by conducting a random draw of half of all available countries and time periods per test-subject-level to produce the linking function using linear linking for consistency of method. Supplement Figure I shows a scatter plot of scores with all countries and time periods relative to linking functions using a random sample. Supplement Table II quantifies these differences.

We find average point differences of less than 1 point for PISA, followed by 7 points for EGRA and LLECE, and 20-25 points for SACMEQ and PASEC. This variation is consistent with country-fixed effects results suggesting smaller differences where there is more country overlap and data availability. EGRA and LLECE converge similarly to PISA with the difference in scores falling within standard error margins of 2 to 7. PASEC and SACMEQ score differences vary more widely, necessitating caution when interpreting precise scores. Overall, we find consistently high correlations above .95 indicating while scores are not identical, they change in consistent directions. This indicates relative rankings and country groupings are preserved.



SUPPLEMENT FIGURE I:
Learning Scores with All Countries and Time Periods vs. Random Subset

Notes: We compare Harmonized learning Outcomes (HLO) using all county and time periods over the fixed linking period with HLO scores computed using a random subset of half of available countries for each test-subject-level.

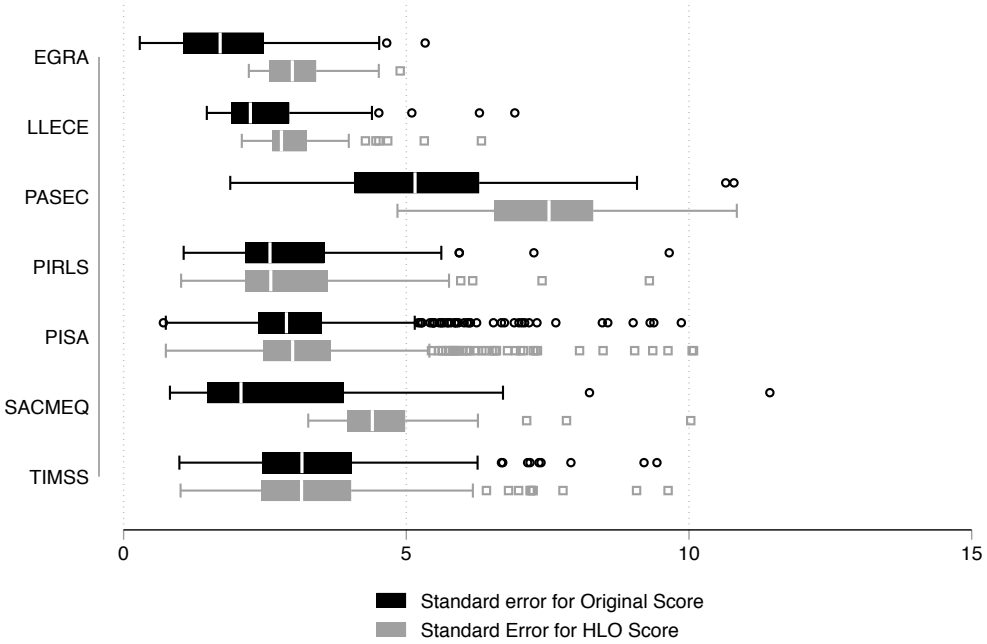
SUPPLEMENT TABLE II:
LEARNING SCORES WITH ALL COUNTRY AND TIME PERIODS VS. RANDOM SUBSET

	(1) EGRA	(2) LLECE	(3) PASEC	(5) PISA	(6) SACMEQ
HLO	368.6	420.4	410.5	497.9	390.6
HLO - Random Set of Countries and Time Intervals	361.6	422.2	433.1	497.3	409.5
Correlation	1.000	0.990	0.975	1.000	0.955

Notes: We compare Harmonized learning Outcomes (HLO) using all county and time periods over the fixed linking function period with HLO scores computed using a random subset of half of available countries for each test-subject-level.

Next we explicitly account for linking errors by including measures of uncertainty to quantify the degree of confidence around our estimates by test. We capture two sources of uncertainty: scores on the original test and uncertainty in the estimation of linking parameters across tests. We calculate the variance by bootstrapping. We consider each average country score on a given subject, test, and schooling level as a random variable with a mean – the score itself – and a standard deviation which captures the sampling variation across students. This distribution of scores is asymptotically normal by virtue of the central limit theorem. We take 1,000 draws from the distribution of subject-level average test scores for each testing regime. We do this as a computational shortcut, rather than bootstrapping subsamples of students from each test. We derive the linking function and scores from each bootstrapped sample. We take the 2.5^{th} and 97.5^{th} percentiles of the distribution and use this to construct lower and upper bounds of uncertainty.

We find small uncertainty intervals overall, as shown in Supplement Figure II with an average of 3.5 points and ranging from 1 to 11 points. This is consistent with original standard errors from each respective testing regime. Consistent with sensitivity tests, we find larger uncertainty for our estimates relative to original scores when testing regimes have fewer countries participating in a given pair of tests. Supplement Figure II decomposes standard errors due to within-test sampling variation as well as variance in the linking function. This figure shows that for tests where there is no need to produce a linking function, or many pair-wise countries which we can use to produce this linking, the final standard errors remain similar to standard errors from the original test. For tests with fewer pair-wise countries, the linking has more uncertainty, such as PASEC, where the average standard error increases from 5.3 on the original test to 7.5 for the HLO. By quantifying the degree of uncertainty, we can more reliably bound our estimates.

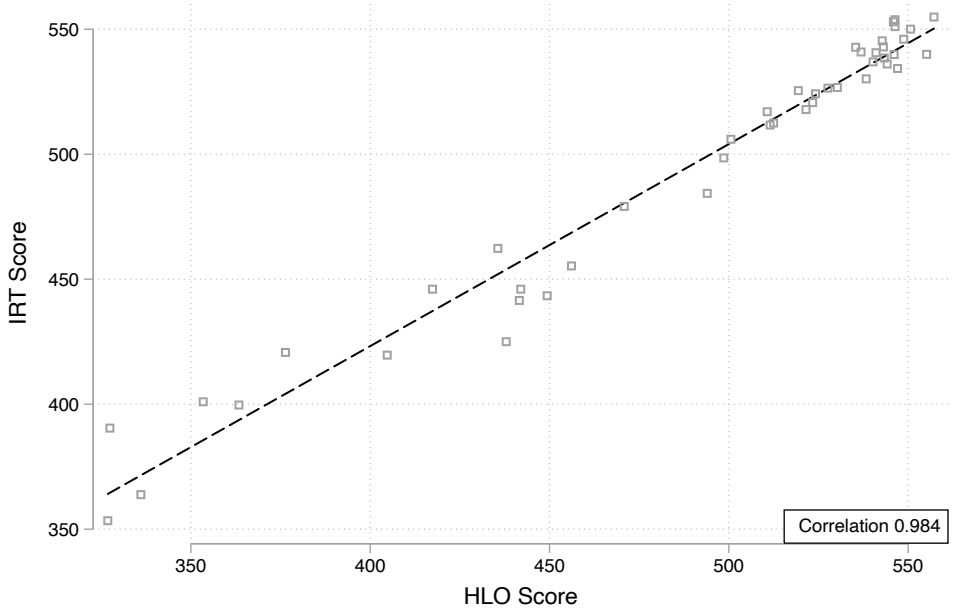


SUPPLEMENT FIGURE II
Standard Errors by Test

Notes: We decompose standard errors on the overall HLO score versus the original test. This reveals the source of uncertainty derived from sampling variation from the original test relative to variation introduced from the production of the linking function.

Finally, we supplement the primary approaches to link regional to international assessments with alternative linking methods. We compare results with learning scores using Item-Response Theory, often considered one of the most reliable methods in the psychometric literature (Kolen and Brennan 2014). IRT models the probability a given pupil answers a given test item correctly as a function of pupil and item-specific characteristics (Mislevy, Beaton, Kaplan, and Sheehan 1992; Holland and Dorans 2006). This methodology is used to construct the underlying tests we use. However, to use it to compare learning *across* assessments would require enough overlap in the test items across each assessment. This is not true for a large enough set of tests and time periods to create a globally comparable panel data set. For example, TIMSS 1995 and SACMEQ 2000 included overlapping math items, but only had three items to make this comparison. 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.

While IRT might not be a reliable approach when there is limited item-by-item overlap, we conduct a few comparisons where overlap is larger. We compare our results to the *Linking International Comparative Student Assessment* (LINCS) project which uses IRT methods and has significant overlap in items for a subset of international studies focused on reading at primary school (Steinmann, Strietholt**,** and Bos 2014). We find that our database can produce similar results to IRT methods for average scores where there is overlap with a correlation coefficient above 0.98 for primary reading scores.

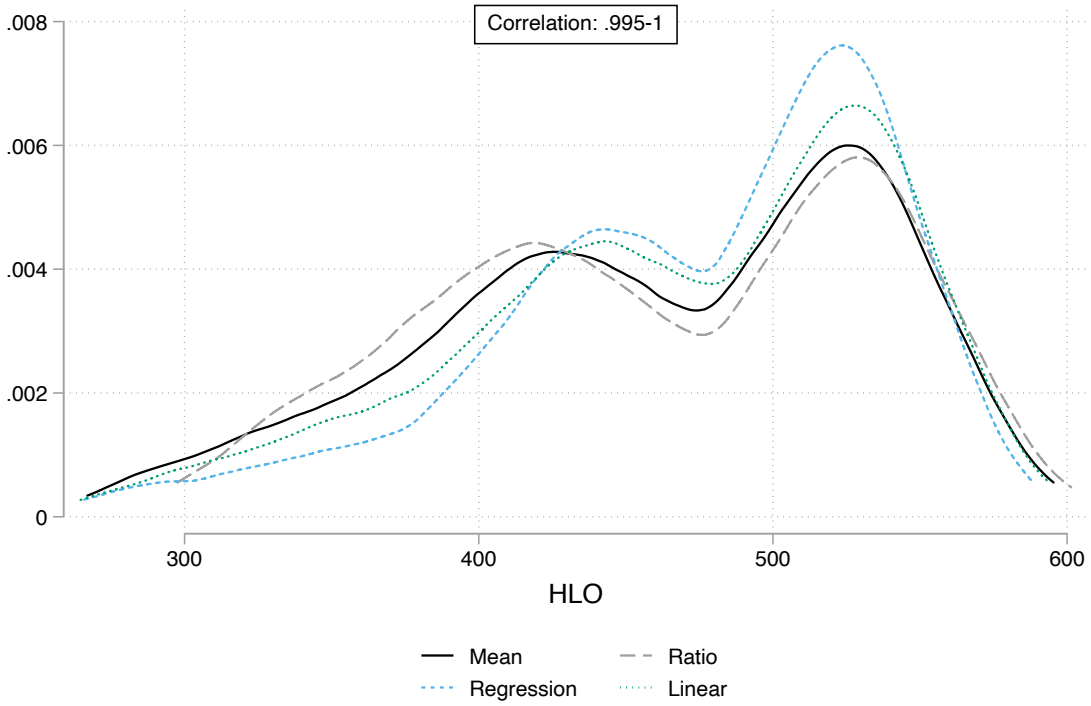


SUPPLEMENT FIGURE III
Comparison to Item Response Theory (IRT) Linking (2000-2010)

Notes: We compare our data to the *Linking International Comparative Student Assessment* (LINCS) project which uses Item-Response Theory (IRT) methods to link test score data. IRT methods are used to create scores for the underlying international and regional assessments used. However, to compare across assessments, IRT would require enough overlap in the test items. This is not true for a significant enough set of tests and time intervals to create a globally comparable panel data set. However, for a subset of tests, this is the case, such as a series of international studies focused on reading at primary school. The LINCS project produces scores on a globally comparable scale using this subset of data (Steinmann, Strietholt**,** and Bos 2014). We compare results on the HLO as well as the LINCS data for primary school reading on average between 2000-2010.

Supplement Figure III compares scores for the same subject (reading), schooling level (primary) and time period (2000-2010), with a correlation of .984. This comparison indicates that as we expand coverage to 164 countries, we maintain high consistency with alternative measures where there is overlap.

Finally, we compare our primary linking approach using regression and linear linking with two alternative approaches and compare robustness across them. First, we use simple mean linking which introduces a constant adjustment between tests matched with rounds, and which we average across testing rounds. This approach assumes constant standard deviations across tests. Second, we use a ratio between test means and also take an average across rounds. This approach assumes a constant scalar adjustment λ between means and standard deviations across tests. The ratio approach is salient and intuitive for policymakers. However, a potential challenge in applying ratios is that they are in principle sensitive to the scale of the test. For example, given score scales have no absolute zero, in theory we can add 300 points to the mean of each test and preserve the interval properties of the scale, but will alter the conversion ratios (i.e. exchange rates). We address this potential issue by having strict inclusion criteria for the underlying tests: they have a uniform scale with a mean of 500 and standard deviation of 100. “Exchange rates” are derived using the same scale and applied on the same scale. Thus, while in theory changing score scales might bias results, by design this is not the case. This increases the likelihood we capture differences in test difficulty rather than arbitrary scaling variation.

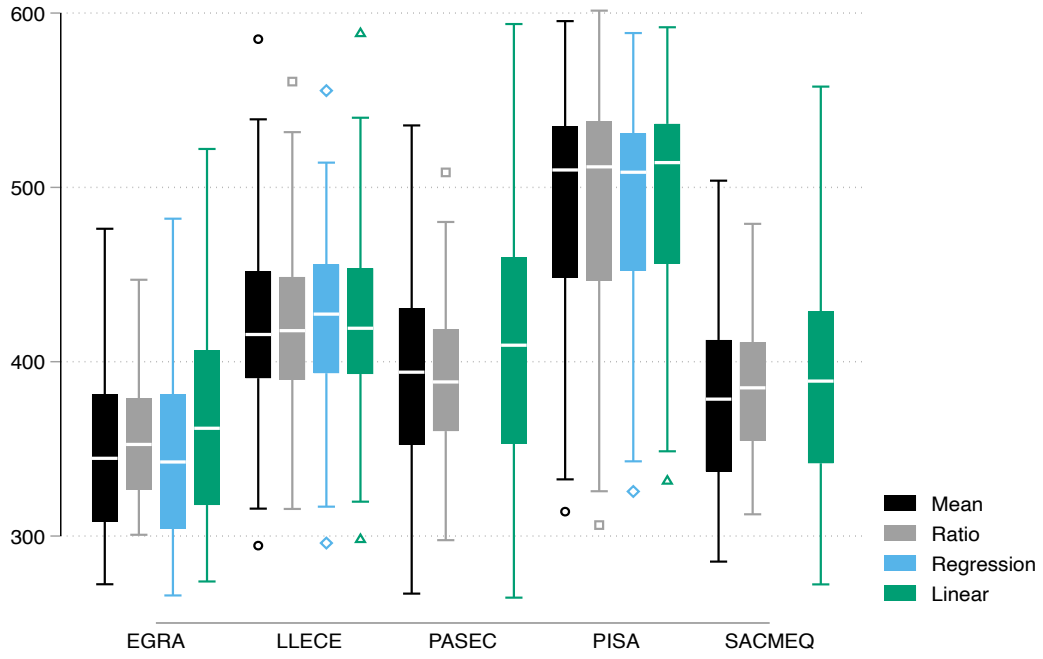


SUPPLEMENT FIGURE IV
Comparison of Scores Across Linking Methods

Notes: We compute Harmonized Learning Outcomes scores using multiple methods including regression, linear, mean and ratio linking. This figure compares scores using a density plot.

Supplement Figure IV shows how scores compare across methods. Overall, we find a correlation of .995 and above, indicating high levels of robustness. Supplement Figure V below breaks down scores distributions by test and method. This reveals similar patterns, with testing regimes with more overlapping countries showing more consistent scores across method. Taken together, these results reveal overall robustness. A caveat is that scores from regional assessments from PASEC and SACMEQ in particular should be interpreted carefully and focus less on precise scores and more on relative ranks and country groupings.

Over time, as more countries participate in more assessments, we anticipate the linking functions used to produce harmonized scores will become increasingly robust. The approach outlined here produces a first set of global comparisons, demonstrates aggregate reliability, quantifies uncertainty to bound estimates, and provides a foundation for continuingly more robust data and comparisons as more countries partake in regional and international assessments.



SUPPLEMENT FIGURE V
Comparison of Scores Across Linking Methods by Test

Notes: We compute Harmonized Learning Outcomes scores using multiple methods including regression, linear, mean and ratio linking functions. This figure compares scores using a density plot. We only compute scores using the regression method for LLECE, EGRA and PISA since SACMEQ and PASEC only have a single country used to make score comparisons.

C. Potential Limitations

A potential limitation is the representativeness of the data of the total stock of cognitive skills in a given country. While the tests used are nationally representative, they are conducted at the school. To this end, learning data might be affected by enrollment patterns, and we advise users of the data to analyze learning outcomes alongside enrollment trends. For example, as marginal students enter the schooling system, average test scores might be driven by selection rather than true learning progress. While this is a potential concern, it is mitigated for a few reasons. First, primary enrollment rates are relatively high, reaching 90 percent on average, and above 75 percent even in the furthest behind regions, such as Sub-Saharan Africa. Second, the direction of the bias is likely to yield a conservative upper bound of learning in a given country. If all students enrolled, the average test score would be even lower, since the marginal students would pull the average down. Since most countries at the bottom of the distribution of learning are also those with relatively lower enrollments, it is unlikely this will alter substantive conclusions – the lowest preforming countries will be revealed to be even lower performing. In addition, data at the primary level should be largely unaffected, since at this level students are being taught basic skills, such as reading "the name of the dog is Puppy." Thus, even if marginal students enter the system, these students should still be expected to attain basic skills by the time they are tested in later primary school grades. Of note, in future work, we aim to include household-based learning data to sign and quantify the

degree of selection present in school-based testing. However, current household-based data is limited and not yet comparable across a significant number of countries.

A second limitation regards data availability. While this is the largest learning outcomes database to date, data are still sparse for some countries. This introduces bias if data availability is correlated with education quality or progress. For example, if countries that perform worse have data only in later years (because they were later to introduce assessments), their average score will be likely biased upwards, as the test scores will reflect more recent testing, not stronger performance. Since we provide year-by-year scores this can be accounted for.

Relatedly, when averaging data across subjects, levels and over time, there is a possibility that averages reflect the availability of data rather than learning gains. For example, let's examine a case where a country has a score of 500 in 2000 in math and jumps to 550 in 2005. If this country added reading in 2005 and scored 450, the average score across subjects in 2005 would be 500, reflecting no learning progress since average scores would be 500 in both years. However, an apples-to-apples comparison in math shows learning grains from 500 to 550. To address this issue, we construct disaggregated measures by subject and schooling levels as well as aggregated ones. This enables analyses at each level considering the trade-offs.

A point of emphasis is that while learning measures human capital better than prior proxies, such as enrollment, learning does not capture the concept of human capital in its totality. Moreover, assessments do not capture only cognitive skills. For example, recent evidence suggests test scores pick up as differential effort as well as cognitive ability (Gneezy, List, Livingston, Sadoff, Qin, and Xu 2017). We use learning outcomes in this paper with these caveats in mind.

II. Descriptive Statistics Supplement

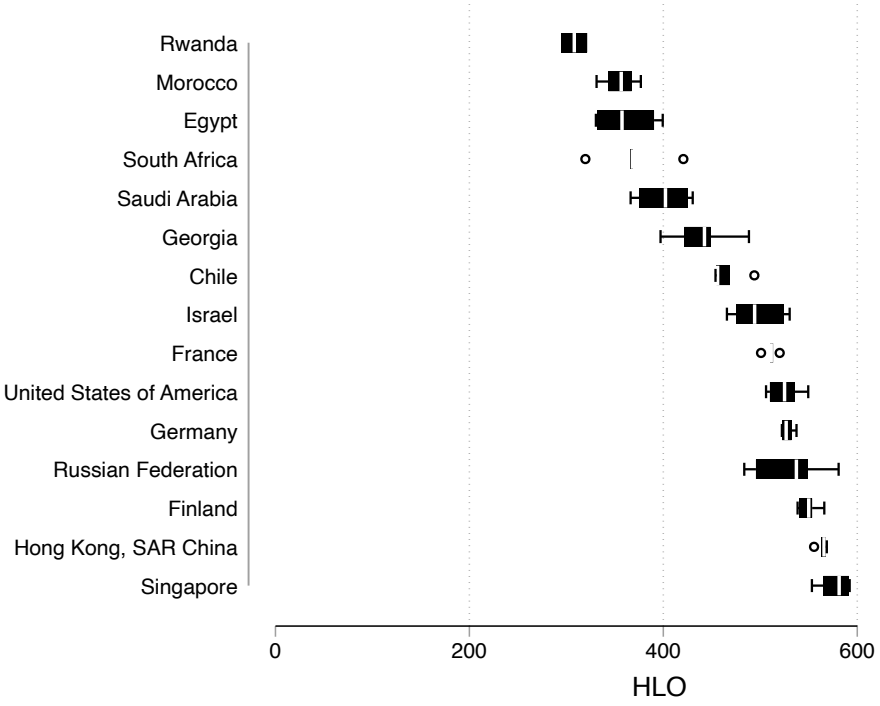
Several statistics demonstrate the coverage and detail of the database. Supplement Table III presents country-subject-level observations by year. The data are spread over time, slightly weighted towards recent years since countries are increasingly participating in assessments. A related feature of the data is a large influx of data in particular testing years. This is more prevalent for developing regions which participate in sporadic assessment.

SUPPLEMENT TABLE III
COUNTRY-SUBJECT-LEVEL OBSERVATIONS BY YEAR

Total	Total	Female	Male	Reading	Math	Science	Primary	Secondary
2000	155	155	155	56	56	43	26	129
2001	34	34	34	1	33	0	34	0
2002	4	4	4	2	2	0	4	0
2003	250	250	250	105	40	105	44	206
2004	2	2	2	1	1	0	2	0
2006	277	277	277	88	123	66	107	170
2007	193	193	193	96	15	82	97	96
2008	3	3	3	0	3	0	3	0
2009	225	225	225	73	79	73	6	219
2010	6	5	5	0	6	0	6	0
2011	240	240	240	92	56	92	152	88
2012	202	201	201	64	74	64	10	192
2013	78	54	54	27	36	15	78	0
2014	28	27	27	10	18	0	28	0
2015	378	378	378	153	75	150	98	280
2016	55	54	54	0	55	0	55	0
2017	4	3	3	0	4	0	4	0
Total	2134	2105	2105	768	676	690	754	1380

Notes: This table presents country-subject-level observations by year.

Supplement Figure VI summarizes learning for selected countries for the last decade. We observe a few interesting case studies. Russia outperforms the United States. Chile outperforms Eastern European countries such as Georgia. Saudi Arabia places near the bottom outperforming only African countries. The gap between Morocco and Singapore is substantial. Singapore and Finland have low variation due to a potential plateau on the upper end of performance. Rwanda has low variation due to limited data. Russia has high variation due to improving learning, whereas South Africa has high variation due to declining learning.

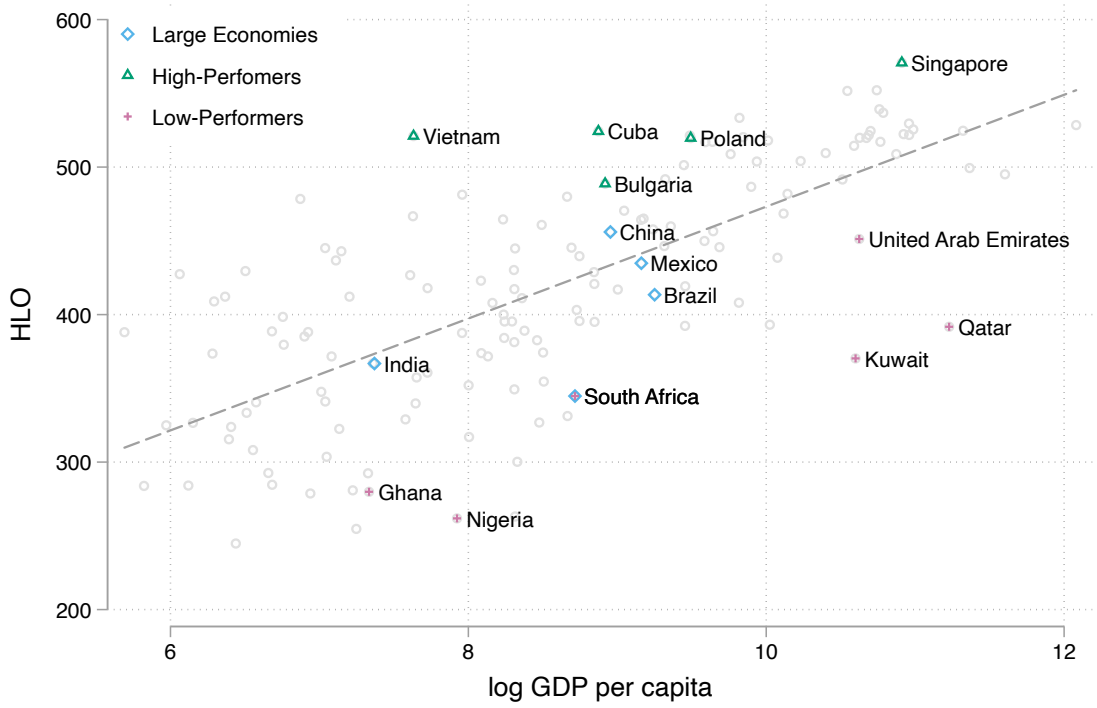


SUPPLEMENT FIGURE VI

Average Learning (2007-2017) – Selected Countries

Notes: Average learning is calculated across subjects and schooling levels over the given period of time.
Source: Our learning outcomes database.

Supplement Figure VII plots average learning for each country by the log of their GDP per capita. This graph illuminates cases where countries have managed to improve learning despite a lack of resources, as well as cases where countries have resources to invest in to date unrealized learning potential. Former or current centrally planned economies display better learning outcomes than their income would suggest, such as Singapore, Poland, Bulgaria, Cuba and Vietnam. Countries in the Middle East and Africa reach lower learning levels than predicted by income, such as Qatar, Kuwait, United Arab Emirates, South Africa, Nigeria and Ghana. We also highlight large developing countries: India, China, Mexico, and Brazil. China outperforms its counterparts, Mexico, India and Brazil perform slightly below where their income would predict, and South Africa trails far behind.



SUPPLEMENT FIGURE VII

Average Learning (2000-2017) versus 2015 GDP per capita

Notes: Average learning is calculated across subjects and schooling levels over the given time period.
Source: GDP per capita estimates are from World Bank national accounts data; learning outcomes are from our database.

III. DATA DESCRIPTION SUPPLEMENT

A. International Standardized Achievement Tests (ISATs)

In the mid-1990s, there was an emergence of standardized, psychometrically robust and relatively consistent ISATs. Below we describe the major ISATs we use in this database.

*TIMSS***.** The Trends in International Mathematics and Science Study (TIMSS) is conducted by the IEA. Five TIMSS rounds have been held to date in Math and Science subjects covering grades 4 and 8. The first, conducted in 1995, covered 45 national educational systems and three groups of students.[11] 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

[11] 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). 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.

on both primary and secondary education (grades 4 and 8). In 2007, the fourth survey covered grades 4 and 8 and more than 66 educational systems. In 2011, the survey covered 77 educational systems across grades 4 and 8. The last round was performed in 2015 and covered 63 countries/areas. The precise content of the questionnaires varies but remains systematic across countries.

*PIRLS***.** The Progress in International Reading Literacy Study (PIRLS) survey is also conducted by the IEA. The PIRLS tests pupils in primary schools in grade 4 in reading proficiency. Four rounds of PIRLS have been held to date in 2001, 2006, 2011 and 2016.

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 2011 round of PIRLS was carried out alongside TIMSS and included 60 countries/areas. The newest round of PIRLS in 2016 includes 50 countries.

*PISA***.** The Organization for Economic Co-operation and Development (OECD) launched the Programme for International Student Assessment (PISA) in 1997 to provide comparable data on student performance. Since 2000, PISA has assessed the skills of 15-year-old pupils every three years. PISA concentrates on three subjects: mathematics, science and literacy. The framework for evaluation remains the same across time to ensure comparability. In 2009, 75 countries/areas participated; in 2012, 65 countries/areas participated and in 2015, 72 countries/areas participated. An important 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.

B. Regional Standardized Achievement Tests (RSATs)

In addition to the above international assessments, a series of 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). SACMEQ is a psychometrically designed, standardized test which generally assesses math, reading and English in grade 6 pupils. The first SACMEQ round took place between 1995 and 1999. SACMEQ I covered seven different countries and assessed performance only in reading. The participating countries were Kenya, Malawi, Mauritius, Namibia, United Republic of Tanzania (Zanzibar), Zambia and Zimbabwe. The studies shared common features (instruments, target populations, sampling and analytical procedures). SACMEQ II surveyed 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) spans 2006-2011 and covers the same countries as SACMEQ II plus

Zimbabwe. SACMEQ collected its latest round of data in 14 countries in East and Southern Africa from 2012-2014. These include Botswana, Kenya, Lesotho, Mauritius, Malawi, Mozambique, Namibia, Seychelles, South Africa, Tanzania, Uganda, Zambia, Zanzibar and Zimbabwe. SACMEQ was designed and scaled to be comparable to past rounds. We include microdata from prior rounds, and estimates from reports for the latest round of SACMEQ since the microdata are pending.

*PASEC***.** The "Programme d'Analyse des Systèmes Éducatifs" (PASEC, or "Programme of Analysis of Education Systems") was launched by the Conference of Ministers of Education of French-Speaking Countries (CONFEMEN). These surveys are conducted in French-speaking countries in Sub-Saharan Africa in primary school (grades 2 and 5) in math and French. Each round includes 10 countries. PASEC I occurred from 1996 to 2003; PASEC II from 2004 to 2010 and PASEC III was conducted in 2014. Of note, PASEC has not always been conducted simultaneously across countries and participation has varied considerably since 1994.[12] The most recent PASEC in 2014 uses Item Response Theory (IRT). Ten countries participated, including Benin, Burkina Faso, Burundi, Cameroon, Chad, Republic of Congo, Côte d'Ivoire, Niger, Senegal and Togo. We include these countries using available microdata. Madagascar also participated in 2015 and was scaled to the PASEC 2014 round. We include Madagascar in our database using estimates from reports. To this end, inclusion of the recent PASEC data adds countries as well as enhances the quality of data from Sub-Saharan Africa. This marks significant improvement over past data sets. To provide a link to past PASEC rounds, which used classical test theory, we create an inter-temporal comparison using a linking function derived based on Togo, which participated in all rounds of PASEC. However, given that PASEC did not conduct intertemporal scaling calibration directly, intertemporal comparisons for PASEC should be analyzed with this caveat in mind.

*LLECE***.** The Latin American Laboratory for Assessment of the Quality of Education (LLECE) was formed in 1994 and is coordinated by the UNESCO Regional Bureau for Education in Latin America and the Caribbean. Assessments conducted by the LLECE focus on achievement in reading and mathematics in primary school. The first round was conducted in 1998 across grades 3 and 4 in 13 countries. These countries include: Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Honduras, Mexico, Paraguay, Peru and Venezuela. The second round of the LLECE survey was initiated in 2006 in the same countries as LLECE I. In round two, called the Second Regional Comparative and Explanatory Study (SERCE), pupils were tested in grade 3 and grade 6. The Third Regional Comparative and Explanatory Study (TERCE), was done in 2013 across grades 3 and 6 and included 15 Latin American and Caribbean countries. We only include SERCE and TERCE data in this database, since these assessments are most similar and cover comparable grades.

C. The Early Grade Reading Assessment (EGRA)

[12] 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), Republic of Congo (2007), Senegal (2007), Burkina Faso (2007), Burundi (2009), Côte d'Ivoire (2009), Comoros (2009), Lebanon (2009), Togo (2010), Democratic Republic of Congo (2010), and Chad (2010). Additional countries took a slightly different test between 2010 and 2011 (Lao PDR, Mali, Cambodia and Vietnam).

The Early Grade Reading Assessment (EGRA) is a basic literacy assessment conducted in early grades. The assessment is conducted most often in grades 2-4. Since 2006, EGRA has been conducted in over 65 countries. EGRA was developed by RTI and is typically implemented by USAID, RTI and local partners (Gove 2009).

The assessment is a short oral assessment conducted with a child one-on-one. EGRA is designed to be flexible and adapted across countries and contexts, while maintaining core modules and similarities. EGRA is a timed test, enabling uniformity in how it is conducted. The tests often represent the most common features of the local language and align with the expectations of the grade level. EGRA includes up to thirteen subtasks, such as 'oral reading fluency', 'vocabulary', 'diction', and 'reading comprehension.' Multiple questions are included in each subtask to test proficiency. Of the thirteen subtasks, there are a few subtasks encouraged to be delivered across all countries and contexts (Dubeck and Gove 2015).

We compile and include data from the 'reading comprehension' indicator in EGRA from 48 countries. This indicator is available in nearly all EGRA data sets and is less sensitive to differences in context, implementation and language. It also and has a strong conceptual link to RSATs and ISATs (Abadzi 2008; Dubeck and Gove 2015) which also measure reading comprehension. To ensure robustness to language effects, we only include data when students took the test in their language of instruction. We use data for grades 2-4, which EGRA is designed for, although certain countries will participate out of this range. We restrict data used for our database to grades 2-4 to be consistent with the design of EGRA. We scale the EGRA microdata to a mean of 500 and standard deviation of 100. This scale corresponds to the scale used by RSATs and ISATs. We include all EGRA data from 2007-2017 as one round. This ensures our scaling is not biased by changing distributions of countries. In the future, we will consider new EGRA data as part of a future round and will conduct intertemporal comparisons using a similar approach to PISA (OECD 2015). Patrinos and Angrist (2018) provide additional detailed analysis and robustness checks on the inclusion of EGRA data.

The inclusion of EGRA adds 48 countries to the database with at least one data point in the past 10 years, nearly all of which are developing economies. Of the 48 countries, nearly two-thirds (31 countries) have data that is nationally representative. Linking functions for EGRA are derived using countries with nationally representative data only, to ensure the assumptions underlying the construction of the linking function hold. We include countries with non-representative data only when the alternative is no data. We include a dummy variable indicating when the data is not nationally representative to enable users of the database to analyze the data accordingly.

D. Summary of Assessments Included in the Database

We include eight learning assessments in our database. Supplement Table IV summarizes the assessments included. Supplement Table V further describes the distribution of source assessments included in our database by country-level-year observations. Most regional assessments are done at the primary level. Moreover, regional assessments comprise nearly 40 percent of primary country-level-year observations, marking substantial representation of developing countries.

SUPPLEMENT TABLE IV
REVIEW OF STUDENT ACHIEVEMENT TESTS

Organization	Abbr.	Year	Subject	Countries/ Areas	Grade/Age
IEA	TIMSS	Every four years since 2003 (latest round is 2015)	M,S	38, 26, 48, 66, 65	4,8
UNESCO	LLECE	2006, 2013	M,S,R	13, 16 (only 6 for science)	3,6
UNESCO	SACMEQ	2000, 2003, 2007, 2013	M,R	7, 15, 16	6 6
CONFEMEN	PASEC	2006, 2014	M,R	22 (before 2014), 10	Until 2014: 2,5 After 2014: 3, 6
IEA	PIRLS	Every five years since 2001 (latest round is 2016)	R	35, 41, 55	4
OECD	PISA	Every three years since 2000 (latest round is 2015)	M,S,R	43, 41, 57, 74, 65, 71	Age 15
RTI/USAID	EGRA	2007-2017	R	65	2,3,4

Notes: When denoting subjects, M=math; S=science; and R=reading.

SUPPLEMENT TABLE V
DISTRIBUTION OF SOURCE TEST FOR HLO

	Total		Primary		Secondary	
Test	Frequency	Percent	Frequency	Percent	Frequency	Percent
EGRA	72	0.03	72	0.10	0	0.00
LLECE	86	0.04	86	0.11	0	0.00
PASEC	60	0.03	60	0.08	0	0.00
PIRLS	160	0.07	160	0.21	0	0.00
PISA	1034	0.48	0	0.00	1034	0.75
SACMEQ	78	0.04	78	0.10	0	0.00
TIMSS	644	0.30	298	0.40	346	0.25

Notes: We include country-year-level observations by source test based on the metadata.

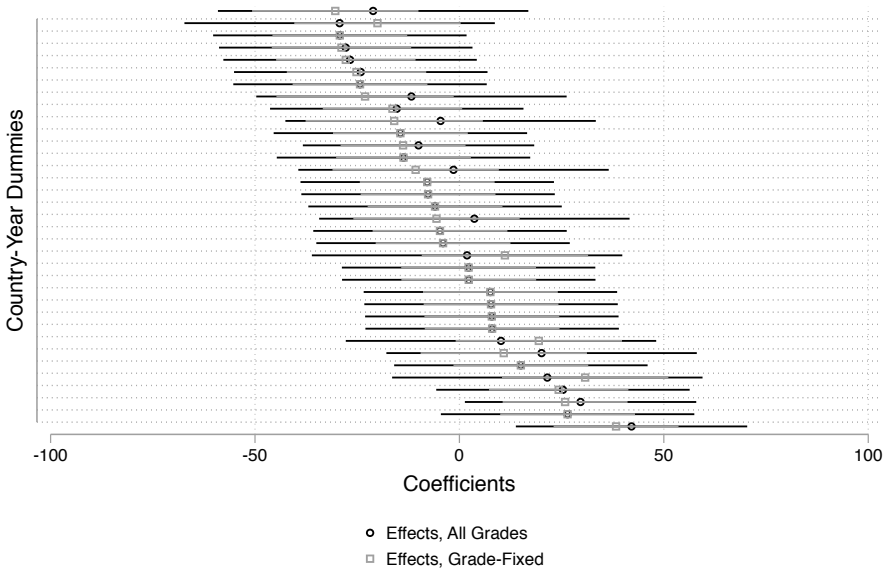
III. ADDITIONAL METHODOLOGICAL PARAMETERS

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

Time Intervals.— While this is one of the largest and most comprehensive comparable learning outcomes databases produced to date, it is still sparse given limited test frequency. In other databases, the data is often disaggregated over 5-year periods. This produces continuously spaced intervals, is designed to reduce noise by averaging results within these intervals, and is comparable to the Barro-Lee approach for years of schooling. In this database, we have moved away from this approach. We now provide the year of test as documented in official reports. This enables greater granularity and precision of the data and enables the users of the database to make trade-offs at their discretion.

Schooling Levels.— We construct a score for each grade and pool across grades within each schooling level to produce primary and secondary school scores. We distinguish primary from secondary schooling since enrollment rates drop off between levels in many developing countries. This introduces a potential selection term in secondary school scores, with the highest performing students progressing in the system, biasing scores up due to selection rather than actual learning.

Conceptually, the broader categories of 'primary' and 'secondary' scores enable us to categorize learning at schooling levels across assessments which span multiple grades and age groups. If the test is designed for an age group (for example, PISA) we code it at the relevant schooling level (for example, secondary for PISA). We specify an approach to including specific grade levels to ensure we have a tight grade interval within one to two years to minimize scope for grade-fixed effects. While the interval is relatively small, it still leaves room for grade-fixed effects rather than test-fixed effects when linking tests. For example, linking PIRLS 2001 grade 4 with SACMEQ 2000 grade 6 might capture a grade difference in PIRLS in addition to difficulty. However, to enable greater country coverage, we put up with the need to expand beyond single grade level intervals. Moreover, these differences are often small and since linking functions are applied to all tests being linked, original ranks will be preserved. An analysis of EGRA in Supplement Figure VIII demonstrates sensitivity to grade. We run a regression with and without grade-fixed effects comparing mean scores relative to a country which participates across all three grade levels 2-4. We find small differences, with near complete overlap in the confidence intervals on the grade and non-grade-fixed estimates. This sensitivity analysis increases our confidence that the EGRA data, and other regional assessment data, is robust to data availability by grade. We also include a variable with grade information in the database to make this transparent.



SUPPLEMENT FIGURE VIII
EGRA, Grade-Fixed Effects

Notes: We run a regression with and without grade-fixed effects comparing mean scores relative to a given country which participates across all three grade levels 2-4 using EGRA data. We only include nationally representative data. The darker "All Grades" bar represents the confidence interval without grade-fixed effects. The lighter "grade-fixed effects" represents the confidence interval with grade-fixed effects.

Subjects.— We construct linking functions specific to reading math, and science. While the proficiency is not granular at the test item level, this ensures that there is significant proficiency overlap when tests are being put on a global scale.

Subsamples.—When calculating the HLO by gender we apply the average linking function to each subsample, rather than constructing subsample specific linking functions. While performance is likely to vary across subsamples in a given test, the relationship between pair-wise tests being linked is unlikely to vary across subsamples nor relative to the full sample.

Metadata versus Aggregate Data.— Our database is disaggregated by subject, schooling level, grade, year and source test. We call this version the 'metadata.' The final data series used in the Human Capital Index (HCI) aggregates the metadata presented in this paper. The aggregation used in the HCI is described in depth in Kraay (2019). There are multiple ways to aggregate the data. For example, the HCI averages data across schooling levels and subjects and uses the most recent year available. The HCI further combines data differently depending on the testing source, for example, including EGRA data in the final time series only when no other data is available. This implicitly weights the importance of testing source over schooling level or subject. Alternative aggregations of the metadata are possible. We present the metadata in this database to enable users to make judgements based on the purpose of their analysis and for maximum transparency.

Exceptions.— In unusual cases, the procedures practiced for a given international or regional test are adapted for the country context. Sri Lanka took a national assessment with items linked to the PISA test to provide comparable scores. Sixth grade students in Botswana took TIMSS instead of fourth grade in 2011. Another example includes India and China, where only certain states and cities participated in PISA. These variations are acknowledged by the underlying tests and the data is caveated with an asterisk in published reports. We preserve this information in our data, and include notes in the metadata for each case. We describe each case in detail in Patrinos and Angrist (2018). In the case of India, we verify that the state data is likely to be nationally representative using national assessment data.

We make an adjustment beyond the underlying tests in the case of China given the likelihood that China's current PISA data is biased. The China HLO based on 2015 PISA data is from four cities (Beijing, Shanghai, Jiangsu, and Guangdong) and is 532. However, this data is likely biased upwards since the cities participating are urban and rich relative to the national average. We adjust the score based on socioeconomic information by city and across the nation and produce an average HLO of 462 at the secondary level, which is plausibly representative at the national level. The detailed procedure is described in Patrinos and Angrist (2018).

SUPPLEMENT REFERENCES

Abadzi, Helen. "Efficient learning for the poor: New insights into literacy acquisition for children." *International Review of Education* 54, no. 5-6 (2008): 581-604.

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

Barro, Robert J., and Jong Wha Lee. "A new data set of educational attainment in the world, 1950–2010." *Journal of Development Economics* 104 (2013): 184-198.

Caselli, Francesco. "Accounting for Cross-country Income Differences." *Handbook of Economic Growth*, 1 (2005), pp.679-741.

Caselli, Francesco and Antonio Ciccone. "The Human Capital Stock: A Generalized Approach. Comment." *American Economic Review*, 109, no. 3 (2019): 1155-74

Caselli, Francesco and John Coleman. "The World Technology Frontier." *American Economic Review* 96.3 (2006): 499-522.

Das, Jishnu and Tristan Zajonc. "India Shining and Bharat Drowning: Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement." *Journal of Development Economics* 92, no.2 (2010): 175–187

Dubeck, Margaret M. and Amber Gove. "The Early Grade Reading Assessment (EGRA): Its Theoretical Foundation, Purpose, and Limitations." *International Journal of Educational Development* 40 (2015): 315-322.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Sally Sadoff, Xiangdong Qin, and Yang Xu. *Measuring success in education: the role of effort on the test itself*. No. w24004. National Bureau of Economic Research, 2017.

Gove, Amber. *Early Grade Reading Assessment Toolkit.* RTI International, USAID and the World Bank, 2009.

Hall, Robert E. and Charles I. Jones. "Why Do Some Countries Produce So Much More Output Per Worker Than Others?" *Quarterly Journal of Economics*, *114*, no.1 (1999): 83-116.

Hanushek, Eric A., and Ludger Woessmann. "Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation." *Journal of Economic Growth* 17, no.4 (2012a): 267-321.

Hanushek, Eric A., and Lei Zhang. "Quality-consistent estimates of international schooling and skill gradients." *Journal of Human Capital* 3, no. 2 (2009): 107-143.

Hendricks, Lutz. "How Important is Human Capital for Development? Evidence from Immigrant Earnings." *American Economic Review 92*, no.1 (2002), 198-219.

Hendricks, Lutz. and Todd Schoellman. "Human Capital and Development Accounting: New Evidence from Wage Gains at Migration." *Quarterly Journal of Economics*, *133* no. 2 (2017): 665-700.

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

Jones, Benjamin F. "The Human Capital Stock: A Generalized Approach." *American Economic Review* 104, no.11 (2014): 3752-77.

Jones, Benjamin F. "The Human Capital Stock: A Generalized Approach: Reply." *American Economic Review* 109, no. 3 (2019): 1175-95.

Klenow, Peter J., and Andres Rodriquez-Clare. *The neoclassical revival in growth economics: has it gone too far?*" NBER Macroeconomics Annual, MIT Press, Cambridge, MA (1997): 83–103.

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

Kraay, Aart. "The World Bank Human Capital Index: A Guide." *The World Bank Research Observer* 34, no. 1 (2019): 1-33

Mankiw, N. Gregory, David Romer and David N. Weil. "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics* 107, no. 2 (1992): 407-437.

Mislevy, Robert J., Albert E. Beaton, Bruce Kaplan and Kathleen M. Sheehan. "Estimating Population Characteristics from Sparse Matrix Samples of Item Responses." *Journal of Educational Measurement* 29, no.2 (1992): 133-161.

OECD. *PISA 2015 Technical Report.* OECD Publishing, 2015.

Patrinos, Harry Anthony, and Noam Angrist. *Global Dataset on Education Quality: A Review and Update (2000–2017)*. The World Bank, 2018.

Psacharopoulos, George, and Harry Anthony Patrinos. "Returns to investment in education: a further update." *Education Economics* 12, no. 2 (2004): 111-134.

Reardon, Sean F., Demetra Kalogrides, and Andrew D. Ho. 2019. "Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale." *Journal of Educational and Behavioral Statistics* (2019).

Schoellman, Todd. "Education Quality and Development Accounting." *The Review of Economic Studies* 79, no. 1 (2011): 388-417.

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

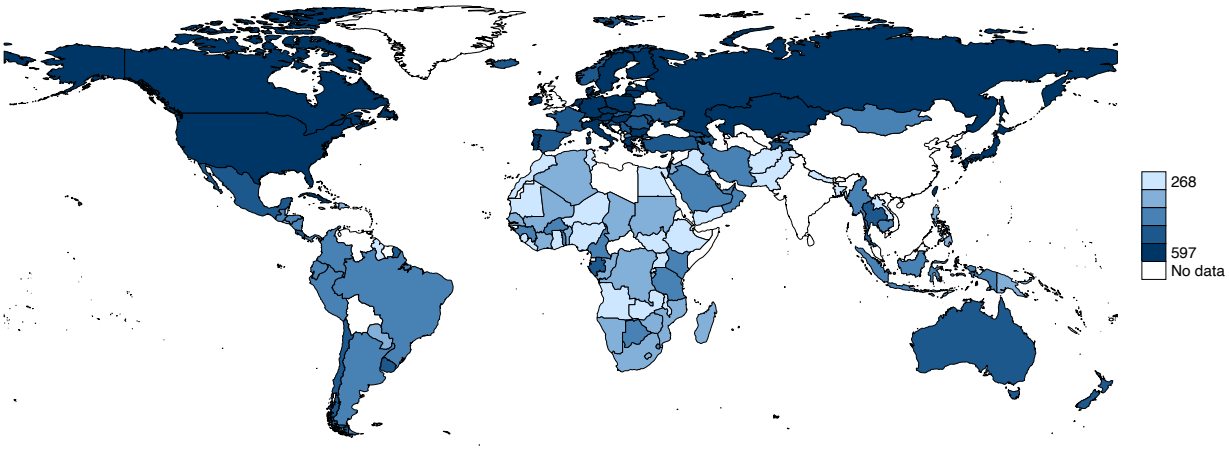
SUPPLEMENTAL TABLES

SUPPLEMENT TABLE VI
TEST LINKING ARCHITECTURE

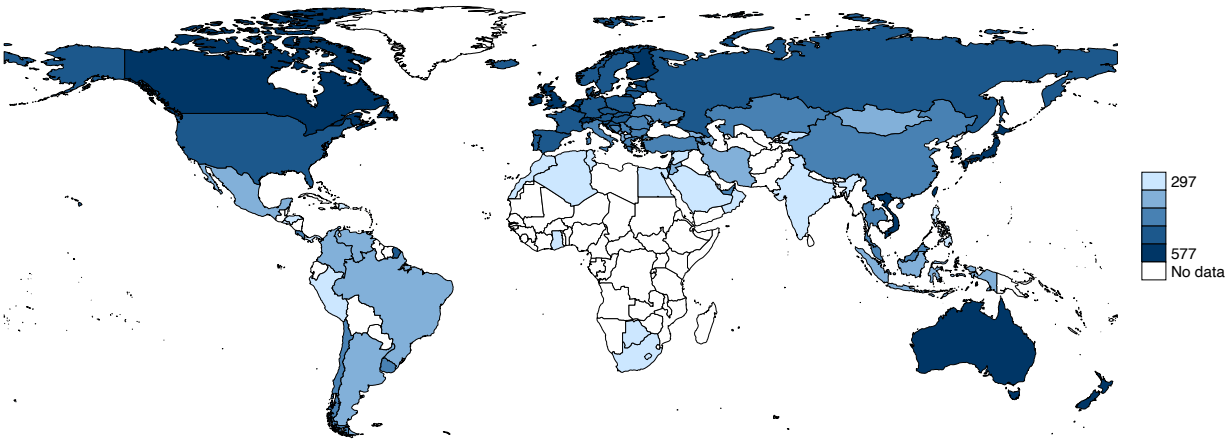
Test X	Test Y	Subject	Level	Overlapping Countries
PISA	TIMSS	Math, Science	Secondary	Australia, Bulgaria, Canada, Chile, Chinese Taipei, Colombia, Czech Republic, Finland, Georgia, Hong Kong – China, Hungary, Indonesia, Israel, Italy, Japan, Jordan, Kazakhstan, Korea, Republic of, Latvia, Lebanon, Lithuania, Macedonia F.Y.R., Malaysia, Malta, Netherlands, New Zealand, Norway, Qatar, Romania, Russian Federation, Serbia, Singapore, Slovakia, Slovenia, Sweden, Thailand, Tunisia, Turkey, USA, UAE.
SACMEQ	PIRLS	Reading	Primary	Botswana
SACMEQ	TIMSS	Math	Primary	Botswana
LLECE	PIRLS	Reading	Primary	Colombia, Chile, Honduras
LLECE	TIMSS	Math, Science	Primary	Colombia, Chile, Honduras, El Salvador
PASEC Round 1	SACMEQ	Reading, Math	Primary	Mauritius
PASEC Round 2	PASEC Round 1	Reading, Math	Primary	Togo
EGRA	PIRLS	Reading	Primary	Egypt, Honduras, Indonesia

Notes: For ease of representation, we include countries used at any point in time for each test linking procedure. In some rounds, some countries are not included, since we specify that for a given round to be linked, tests should be administered in adjacent years. A more detailed architecture by year is available on request.

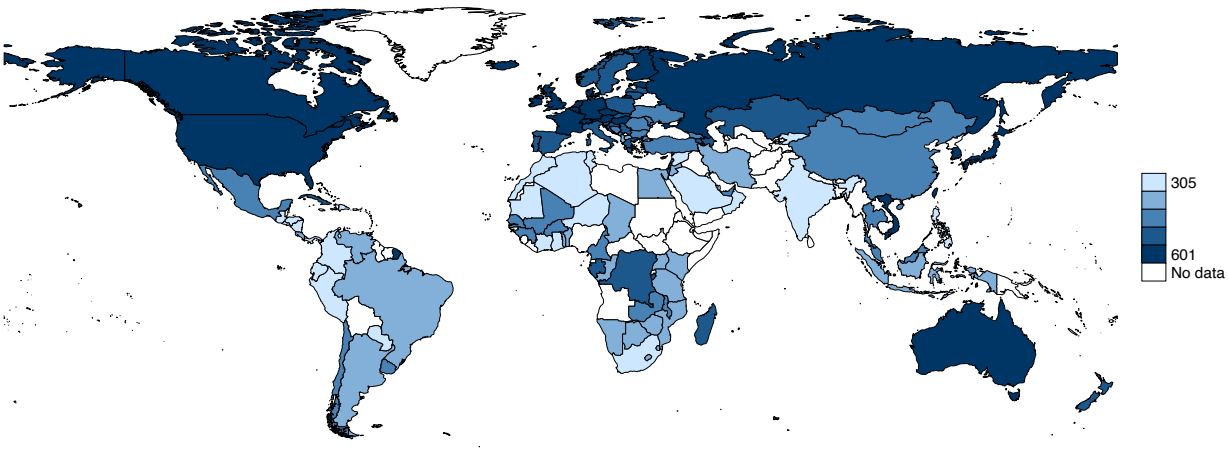
SUPPLEMENTAL FIGURES



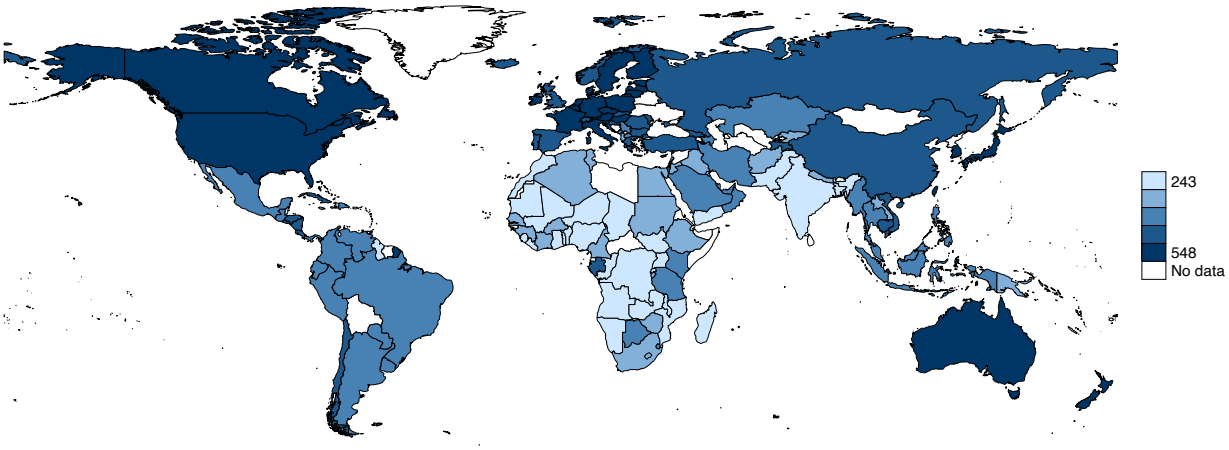
SUPPLEMENT FIGURE IX
Primary Learning Score



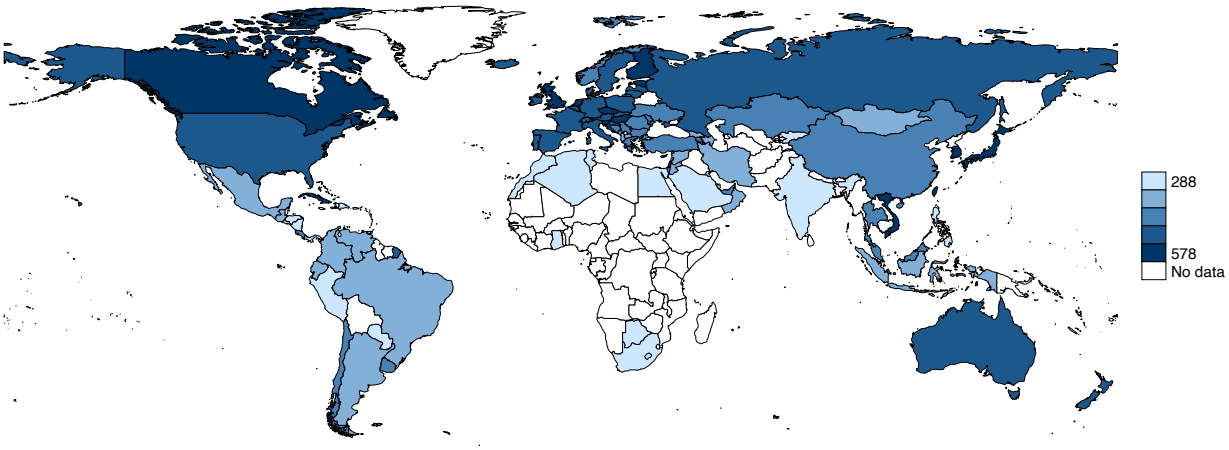
SUPPLEMENT FIGURE X
Secondary Learning Score



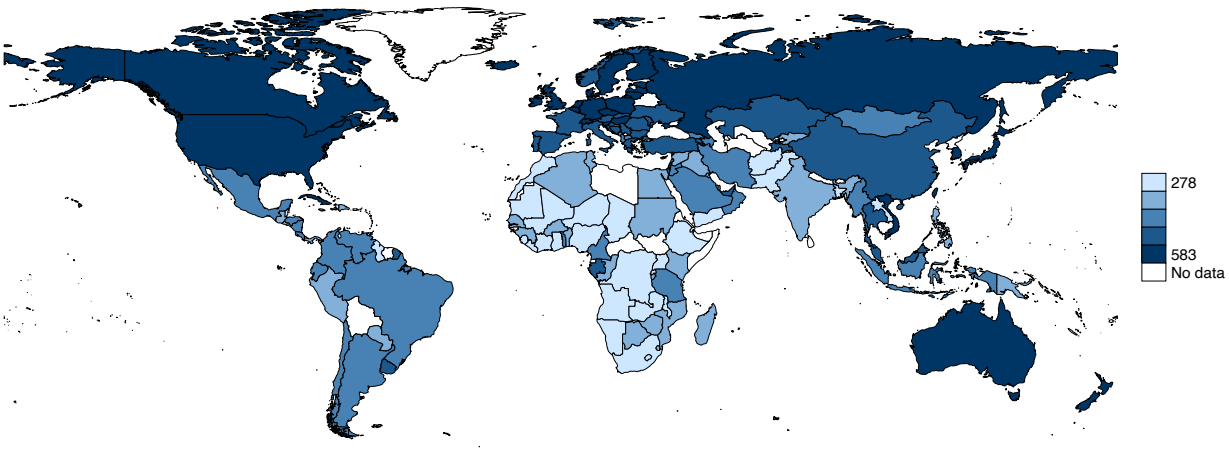
SUPPLEMENT FIGURE XI
Math Learning Score



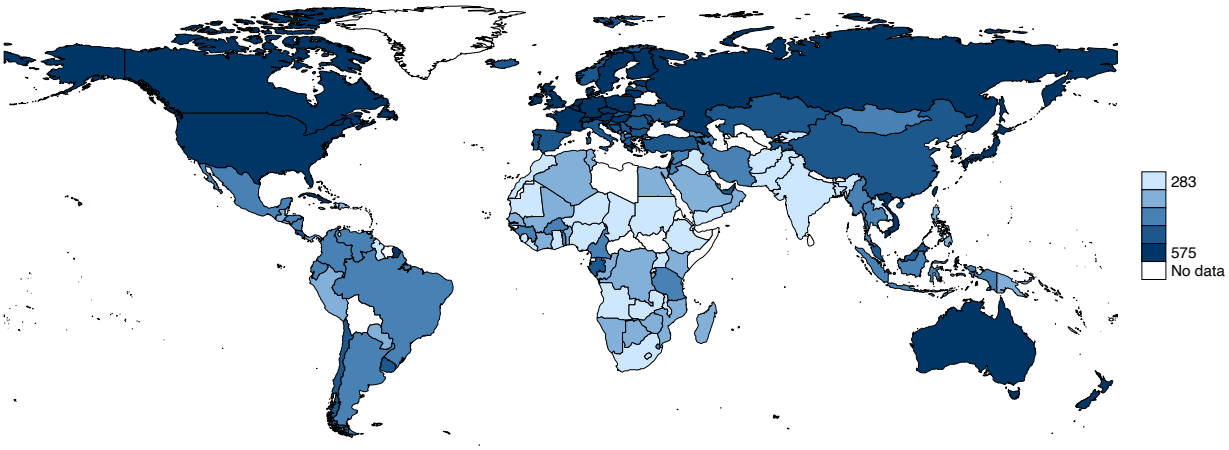
SUPPLEMENT FIGURE XII
Reading Learning Score



SUPPLEMENT FIGURE XIII
Science Learning Score



SUPPLEMENT FIGURE XIV
Female Learning Score



SUPPLEMENT FIGURE XV
Male Learning Score