RISE
RESEARCH ON IMPROVING SYSTEMS OF EDUCATION

RISE Working Paper 19/032
December 2019

The Political Economy of Testing in Latin America and Sub-Saharan Africa

Barbara Bruns, Maryam Akmal, and Nancy Birdsall

Abstract

Most countries in sub-Saharan Africa have not implemented testing of children's learning that can be benchmarked regionally or globally. In contrast, in the last two decades, almost all countries in Latin America have participated in regionally and globally benchmarked testing initiatives. Our analysis of the political economy of cross-national learning measurement in Latin America suggests that policymakers perceive the risks of exposing their education system's performance by joining cross-national assessments, but they also value the quality of the data generated, the strengthening of domestic technical capacity, and the political benefits in using comparative results to argue for reforms or to advertise progress. We document that in Ecuador and Peru cross-national tests played an important role in both stimulating and justifying reforms that have produced major improvements in learning. In sub-Saharan Africa, no cross-national test has been implemented as consistently or widely as the Latin American regional test. The context in Africa makes regional cooperation on cross-national testing more daunting—countries are poorer and more linguistically diverse than in Latin America, raising the relative costs of developing and administering cross-national tests. The experience of Latin America suggests that a coordinated and efficient application of resources in implementing cross-national tests in Africa, on the part of countries and with support of the international community, could help build countries' national capacity, strengthen focus on learning, support better research, and help diffuse reforms that raise learning.

UKaid from the British people
Australian Aid
Bill & Melinda Gates foundation
Oxford Policy Management
Blavatnik School of Government
University of Oxford

The Political Economy of Testing in Latin America and Sub-Saharan Africa

Barbara Bruns
Center for Global Development

Maryam Akmal
Center for Global Development

Nancy Birdsall
Center for Global Development

Acknowledgements:

This research was funded by the Research on Improving Systems of Education (RISE) Program. The analysis benefited enormously from comments by Luis Crouch and David Evans. Aisha Ali provided excellent research assistance. The views expressed here do not necessarily reflect those of the Center for Global Development, its board, or its funders. All errors are ours.

This is one of a series of working papers from "RISE"—the large-scale education systems research programme supported by funding from the United Kingdom's Department for International Development (DFID), the Australian Government's Department of Foreign Affairs and Trade (DFAT), and the Bill and Melinda Gates Foundation. The Programme is managed and implemented through a partnership between Oxford Policy Management and the Blavatnik School of Government at the University of Oxford.

Please cite this paper as:
Bruns, B., Akmal, A., Birsdall, N. (2019). The Political Economy of Testing in Latin America and Sub-Saharan Africa. RISE Working Paper Series.19/032.

Contents

1 Introduction

In many respects, the Millennium Development Goal (MDG) to achieve universal primary education by 2015 was a success. Over a fifteen year period, the number of primary school aged children who were out-of-school fell by nearly half, and the primary completion rate in developing countries rose to 89 percent worldwide, driven by remarkable gains in low-income countries of sub-Saharan Africa and South Asia, especially for girls: girls' primary completion increased from 49 to 68 percent in sub-Saharan Africa and from 62 to 96 percent in South Asia.[1]

But enrollment, and even completion of primary school, has not assured even minimal levels of learning. By heroically piecing together scattered regional and country data, UNESCO estimates that over one-half of all children globally fail to master basic literacy and numeracy skills, despite spending five or more years in school (UNESCO, 2018). In the last decade, increasing attention has gone to the reality that increased enrollment in schools is not producing learning (Pritchett, 2013), and that increased spending on school inputs, including in some of the world's richest countries, is not making any real difference (Beatty et al., 2018). School systems that fail to deliver learning are a deep waste of children's potential and countries' education resources. Growing awareness of a "global learning crisis" over the past decade has turned the focus of governments, civil society, academics, and international organizations from schooling access to education quality (WDR, 2018) (Education Commission, 2017).

In 2015, the Sustainable Development Goals for education for the first time set explicit global targets for higher student learning by 2030. The 2030 learning goals are aspirational in many ways, not least because there is no baseline measure of what the world's children know now at the three levels where the SDG goals envision testing reading and mathematics learning: (i) in Grades 2/3, (ii) at the end of primary, and (iii) at the end of lower secondary. Approximately half of all countries today have no measures of learning at these points in the

[1]Data accessed from World Bank website on October 6, 2018.

education system that are comparable over time or comparable with other countries—two fundamental conditions for aggregating and tracking global progress.

The problem of low learning cannot galvanize action without a measure of learning—"what gets measured gets done." But more than half of developing countries have no national assessments, and even fewer have participated in regional or international assessments. In sub-Saharan Africa in particular, more than half of the countries in the region have not participated in any regionally or internationally benchmarked assessments. Ideally, all countries would engage in standardized learning assessment at the national level—i.e., comparable over time. This is critical as input to management of education systems—to establish meaningful targets or accountability for progress. In addition, countries would also participate in regional and international tests that generate learning results that can be benchmarked against other countries (per SDG 4.1.1)[2] (Birdsall et al., 2016), which can be key inputs to informing education policies within countries.

What explains the paucity of core data in some but not all developing countries? What has made some countries decide to measure student learning in transparent, globally comparable ways and not others? And perhaps more importantly, is there evidence that cross-national testing that measures how well (or poorly) an education system is doing relative to other countries have impact on the adoption of policies to improve learning? Our focus is the politics and impact of participation in cross-nationally comparable tests,[3] either regional or international assessments, for two main reasons.[4] First, these tests—which trans-

[2]SDG 4.1.1: Proportion of children and young people: (a) in Grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.

[3]We use the term cross-national assessments to include both regional assessments and international assessments (which involve countries in multiple regions).

[4]Clearly, cross-national assessments are not the only source of learning data that can stimulate countries to adopt education reforms. Three other types of assessment are important. First, national assessments, which typically reflect a country's own curriculum and may be applied to much larger samples of students, provide more complete metrics on education system performance and in many ways a stronger platform for the design of national reforms and tracking of progress. No one would argue that cross-national tests are a substitute for sound national assessment systems. Our question is how important they are as a complement. Second, in many countries that do not have national assessments, citizen-led assessments, such as ASER in India and Pakistan, and Uwezo tests in East Africa, have documented education system dysfunction (Akmal and Pritchett, 2019), but had little effect on education programs or policy. Third, oral assessments of literacy

parently expose a country's performance vis-à-vis its neighbors and/or more economically developed countries present higher political costs to policymakers than do other types of tests—we explore what can motivate or discourage policymakers to take this risk. Second, cross-national assessments, in addition to measuring progress towards the education SDGs, create a platform for cross-country research on policies and programs that help drive improvement. To the extent these benefits accrue to all countries, actions to offset some of the costs for participating countries are justified. By exploring these benefits and costs, our paper aims to identify practical strategies for expanding the coverage and impact of existing cross-national assessments or building political support for the development of a new assessment—purpose-built for monitoring global learning progress.[5]

Research on these questions is still surprisingly limited, considering how central good data on learning are for monitoring progress and assessing the effect of system reforms. The overwhelming bulk of studies to date are focused on the OECD's PISA exam. Section 2 of this paper summarizes this literature. Section 3 explores the testing issue in Latin America and the Caribbean, with emphasis on the relative success of the region *qua* region in the gradual acceptance and use of benchmarked testing, including the widespread use of a regional test, ERCE, in the Latin American countries (though not the English-speaking Caribbean or Haiti). It includes case studies of the political decisions in Peru and Ecuador to join benchmarked tests and use results to motivate or defend education system reforms. Section 4 turns to the question of why benchmarked testing in Africa has been less widespread across countries than in Latin America, and less systematic and continuous over time in the region—with separate programs for francophone and anglophone countries, the former

and numeracy in the first two grades of primary school (EGRA for reading and EGMA for math) carried out in many African countries have generated compelling evidence of school systems' failure to impart critical basic literacy and numeracy skills, and in countries such as Kenya, have directly stimulated major reforms of the curriculum and teacher practice (Crouch and Gustafsson, 2018) (Piper et al., 2018). These three types of assessment are undeniably important, and by helping to create broader public understanding of the learning crisis, can in principle open the door to regionally or globally benchmarked tests (Trevino and Ordenes, 2017). But in this paper we focus on the potential positive impact of the more politically sensitive benchmarked tests—because they generate comparable data on learning across countries, and on different rates of learning progress over time.

[5]As recommended by Trevino and Ordenes (2017).

with more recent progress than the latter. As an input to assessing why Africa has had relatively little benchmarked testing, we report on the results of empirical analysis, across all five developing regions, of the country characteristics associated with children's learning performance. In doing so, we use a proxy for learning results created by the World Bank as input to its Human Capital Index—the harmonized test score based on the patchwork of benchmarked tests across developing countries.

2 Cross-National Tests of Learning: Some evidence they matter

Multiple different cross-national learning assessments have been introduced over the past 60 years and, globally, country participation in these programs is increasing. Yet it is still a Balkanized picture, with no international assessment covering more than 75 of the world's 223 countries, and even the largest sustained regional assessment, Latin America's ERCE, which covers 15 of the region's 20 Spanish or Portuguese speaking countries, includes none of the neighboring Caribbean countries.

Cross-national assessments by definition generate benchmarking data that are potentially useful to national policymakers and that cannot be obtained in any other way. Other countries' experience defines the "production possibility frontier" of education system performance to raise learning, generates a universe of policy and program ideas, and can supply directly comparable research evidence on their impact. But cross-national assessments have not only financial and technical but also political costs, notably the risk of exposing a country's poor performance. What causes countries to prioritize the potential benefits over the immediate costs? And how frequently do the comparable results of cross-national assessments actually influence national education policies?

Addey and Sellar (2019)[6] have been exploring the first question for several years. They

[6]Other relevant work includes Addey (2015), Addey (2017), Addey and Sellar (2018).

identify eight different rationales for participation in international large-scale assessments (ILSAs), grouped into four broad categories: *political rationales* (build support for government priorities or make the case for change), *economic rationales* (signal a commitment to skills development and competitiveness and/or mobilize additional donor aid), *technical rationales* (strengthen national assessment capacity), and *socio-cultural rationales* (conform to a growing global "assessment culture that conceives of education as a set of skills that can be measured, compared, and ranked"). In most cases, they note that motivation is a mix of these rationales. They also note that non-government actors sometimes influence country decisions to participate or not—in Paraguay a private sector coalition lobbied the government to join PISA for Development (PISA-D), while in Cambodia a donor coalition advocated against it.

Addey and Sellar (2019) group government motivations not to participate in ILSAs under the same four basic rationales: *political* (a reaction to poor performance has led countries such as Peru, Mexico, Panama, Dominican Republic, and South Africa to come in and out of PISA, TIMSS, and PIRLS over the years, and has led to more explicit statements such as Bolivia's that country rankings are antithetical to its more holistic view of education); *economic* (participation is costly if financed domestically and resources might be more productively used for a national assessment); *technical* (low-income and fragile states have difficulty meeting sampling and administration protocols); and *socio-cultural* (ILSAs are not culturally or linguistically appropriate for countries with a lower level of economic development and higher linguistic diversity).

The most comprehensive review of the impact of cross-national assessments on education policy is a recent meta-study by Fischman et al. (2018) that analyzes all major ILSAs,[7] and hundreds of studies, 71 of which were published in peer-reviewed (English-language) journals. They report that 90 percent of the global published literature on cross-national testing focuses on a single assessment—the OECD's PISA exam, which has been administered

[7]PISA, TIMSS, PIRLS, PIAAC, LAMP, STEP, and WEI-SPs. The study did not include regional assessments.

every three years since 2000. Through six rounds of results, PISA has stimulated dozens of country-level studies of policy impact in countries ranging from Germany to Kyrgyzstan as well as sub-national participants, such as Shanghai, different US and Mexican states, and Canadian and Argentine provinces.

Fischman et al. (2018) associate the global prominence of PISA with the early, highly publicized cases of countries' response to the "PISA shock" and the OECD's active program of technical assistance to countries seeking to improve their results. A key factor that emerges from our own research is the perception of middle-income countries such as Peru, Chile, Ecuador, and Brazil that PISA performance sets the global standard for education system quality and also for assessment quality. Joining PISA, although it poses the risk of looking bad, is seen as signaling readiness to compete economically, as well as educationally, with the world's most successful countries. Brazilian Finance Minister Joaquin Levy alluded to this in a 2015 speech, saying that the "new measure of investment grade for emerging markets is the PISA score."

The Fischman et al. (2018) review notes that while many studies document changes in education policy in the aftermath of PISA results, it is impossible to conclude causality. But the sheer volume of cases is impressive and Fischman et al. (2018) describe four different channels through which PISA appears to affect countries' education policy:

- Governments stimulated to implement policy change in response to results;
- Governments use results to justify and/or accelerate the implementation of pre-existing reforms;
- Governments use the results to set specific goals for improved rankings; and
- Results stimulate cross-national policy borrowing from "successful countries" and policy convergence.[8]

[8] Fischman et al. (2018) administered a small survey to 21 "experts" directly involved with PISA in their countries as well as 24 "non-experts" familiar with the test and its results. While 96 percent of respondents agreed that PISA results are "used" in country policy, 43 percent of the experts thought results were mainly "mis-used" (either misunderstood by policymakers, "decontextualized" or mainly used for "name, shame and blame" purposes).

New policies in response to results. Examples of the first channel include the German and Danish government's actions after the first "PISA shock" in 2000 to align national standards with PISA and to develop new support systems for disadvantaged and immigrant students who scored particularly low (Breakspear, 2014) (Ertl, 2014). Another common response linked to PISA in many countries is reform of the curriculum and national assessment systems. Spain simplified its national curriculum to focus on competencies, while Shanghai shifted from "transmission of content" to "real life problems." Both Germany and the Czech Republic introduced national assessment systems in response to PISA and Brazil decided to expand its assessment system from a sample basis (SAEB) to a census-based system (Prova Brasil).[9] A third response is policies to raise teacher quality, such as in-service training, evaluation, new teacher development, and standards. Studies in Spain, Poland, Japan, Brazil, and Colombia, among other countries, have documented changes in teacher-related policies in reaction to PISA results (Fischman et al., 2018).

Use of results to legitimate existing reforms. Examples of the second channel are the French government's use of PISA to justify curriculum and decentralized funding reforms that were already underway in the early 2000s; Switzerland's change in curriculum standards and monitoring; and the UK's reform of its GCSE (secondary school completion) targets. The UK's education secretary Michael Gove frequently referenced PISA to justify his education reform efforts. Research conducted for the present study found that former Peruvian education minister Jaime Saavedra frequently cited the country's last-place ranking among 65 countries on the 2012 PISA to justify staying the course on the country's teacher policy reform, expanding the national assessment system, increasing infrastructure investments, and decentralizing education funding. In trying to assess PISA's impact on this channel, one can reflect on the counter-factual. Clearly, policymakers in many of these countries were already convinced of the need to reform and had already identified priority actions. There would have been a communications strategy to "sell" reforms with or without PISA data

[9] 2019 interview with Maria Helena Castro, director of the Brazilian national testing agency in 2002.

and the comparison to higher-performing countries. The question is, could it have been as persuasive.

Goal-setting. Several countries have set national education targets for improving PISA results, including Mexico, Brazil, Australia, Denmark, Thailand, and Wales. In Mexico's case, the goal was to reach 435 in math and reading by 2012 (not achieved). In Brazil's case, the national education quality index and national learning assessment (Prova Brasil) are anchored to PISA and the goal is to reach OECD average PISA performance by 2021. More recently, Paraguay joined PISA-D and has set learning targets for 2030 linked to PISA (Addey and Sellar, 2019).

Cross-national policy borrowing. Finally, studies document the global interest in "best practices" from PISA's highest performers, including Finland, Singapore, Hong Kong, and Shanghai. All these countries have been the object of study tours, research, and the source of a growing global consensus around policies to raise learning and increase economic competitiveness—including selective recruitment and high-quality training for teachers, concern with educational equity, and a twenty-first century curriculum. Peruvian Ministry of Education officials cited a McKinsey and Co. (2007) report as an influence on policy in the Garcia administration.[10] McKinsey and Co. (2007) defined the world's highest-performing education systems as the ten highest-scoring countries on PISA.

A similar phenomenon was observed in Latin America after the 1997 and 2006 regional test results showed Cuba to be the highest performer in the region, by a substantial margin, in all grades and subjects. Carnoy et al. (2007) produced a book on Cuba's academic advantage, and Ecuadorian ministry officials report that a subsequent study tour to Cuba influenced the design of their reform program.[11]

It is logical to hypothesize that for lower-income countries that do not participate in PISA, other cross-national tests might play a similar role in benchmarking performance and

[10]Survey response from Liliana Miranda, Director of Assessment, Ministry of Education Peru, 2004-2016.
[11]Survey response from Harvey Sánchez Restrepo, former director INEVAL (National Institute for Education Evaluation), Ministry of Education Ecuador.

stimulating change. However, except for the attention brought to Cuba as a result of the Latin American regional test (ERCE), there are many fewer examples of ERCE or the Africa region tests (SACMEQ and PASEC) provoking a "shock" in countries and/or the adoption of education reforms. We review the few published cases in Section 3.2. It is plausible that these regional tests are genuinely less potent than PISA in stimulating countries to react, but the imbalance also likely reflects a global research bias towards richer countries and English language publications. It also reflects publication bias more generally—we know little about the much larger set of countries whose PISA scores provoked no clear policy reactions or those that produced failed reforms.

3 Cross-National Learning Measurement and Education Policy in Latin America

3.1 Learning Levels and Trends

Because most Latin American countries participate in the ERCE regional tests, many participate in PISA, and some also participate in TIMSS and PIRLS, the region is unique among developing regions in the high number of countries that can report globally-benchmarked learning results at all three measurement points recommended for monitoring SDG 4: Grade 3, Grade 6 (the end of primary school), and in secondary school (age 15). These data permit countries to benchmark their performance vis-à-vis that of peers in the region and with the OECD. They also allow countries to compare their rates of progress directly, and they have created a platform for cross-country research on what drives learning improvement.

The picture that emerges is of a region with average learning performance well below the OECD average, but highly heterogeneous, both in terms of current learning levels and rates of improvement.

Learning levels in lower secondary school. At the secondary level, PISA 2015 shows

Chile with the highest performance in all three subjects tested (math, reading, and science) and the Dominican Republic scoring lowest in the region on all three. The average score difference between the two countries of around 100 points (with 40 points roughly equal to 1 year of schooling) equals more than two years of schooling. All countries in Latin America still trail the OECD average on PISA, with Chile roughly two years behind the OECD average, followed by Uruguay, Costa Rica, and Trinidad and Tobago in most subjects; Colombia, Mexico, Brazil, and Peru around three years behind; and the Dominican Republic roughly four years behind the OECD. Results for the PISA-D test suggest that Ecuador is at par with Peru and Brazil, and Guatemala, Honduras, and Paraguay are in the range of the Dominican Republic.

Table 1: Latin America's Performance in PISA 2015

Country	Reading Score	Math Score
Chile	459	423
Uruguay	437	418
Costa Rica	427	400
Trinidad & Tobago	427	417
Colombia	425	390
Mexico	423	408
Brazil	407	377
Peru	398	387
Argentina (2012)	396	388
Dominican Republic	358	328
LAC Average	416	394
OECD Average	493	490

Source: PISA 2015: Results in Focus.

Table 2: Latin America's Performance in PISA-D 2018

Country	Reading Score	Math Score	Science Score
Ecuador	409	377	399
Honduras	371	343	370
Paraguay	370	326	358
Guatemala	369	334	365
LAC Average	380	345	373
Lower-Middle Income Country Average	378	368	392
OECD Average	493	490	493

Source: PISA for Development: Results in Focus.

Both PISA and PISA-D apply school-level, system-level, and student surveys that generate substantial data that are useful for research on factors impacting learning performance. Data collected from students goes well beyond age and gender to cover everything from opinions of their teachers, health, mood, school attendance, and home and family characteristics. School system data cover spending, infrastructure, calendar, testing, and teacher qualifications, among other areas. In general, in the PISA participating countries in Latin America, girls outperform boys in reading, but in math and science, the score difference in favor of boys is more than twice the OECD average. The data also show higher socio-economic segregation between schools in Latin American countries than in the OECD, and bigger gaps in infrastructure and resources between schools serving the poorest and richest 20 percent of students (OECD, 2016).

But, in every area, the data also show significant differences among Latin American countries—and particularly striking differences with the sole Caribbean country participant, Trinidad and Tobago, which is the only country in the region where girls outperform boys, not only in reading, but also in math and science, and by a wide margin (OECD, 2016). This suggests how valuable comparative research could be if the Latin American regional assessment were extended to the English-speaking Caribbean countries and Haiti.

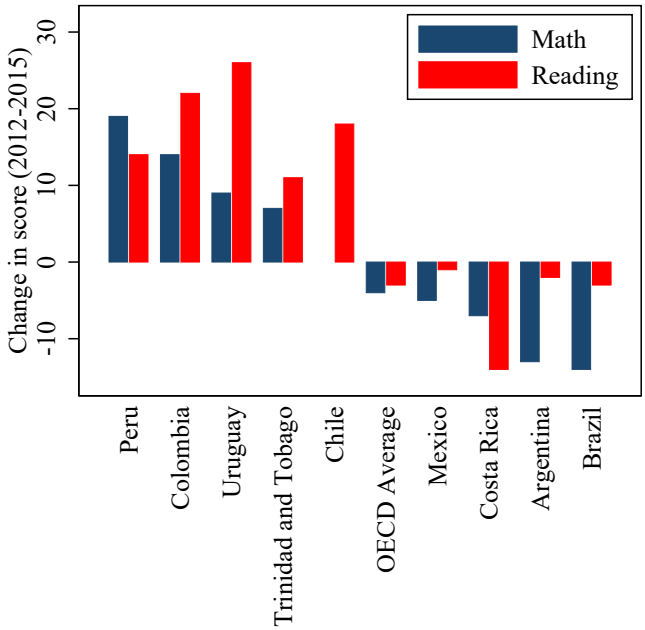
Learning levels at the end of primary school. The second SDG measurement point is captured by the Latin American regional test at the sixth-grade level. TERCE 2013 shows Chile with the region's highest learning levels in math and reading by a significant margin. The Dominican Republic had the lowest learning outcomes in the region in both math and reading.

Learning levels in the early grades (2/3). At the third-grade level, Chile was the highest performing country in both math and reading on TERCE, by significant margins, with Costa Rica and Uruguay the next strongest. The Dominican Republic had the weakest performance in both areas. Across the region, nine percent of third-graders scored below Level 1 in reading, which means they were unable to identify a single word in a sentence. In the Dominican Republic, 31 percent of students scored below Level 1.

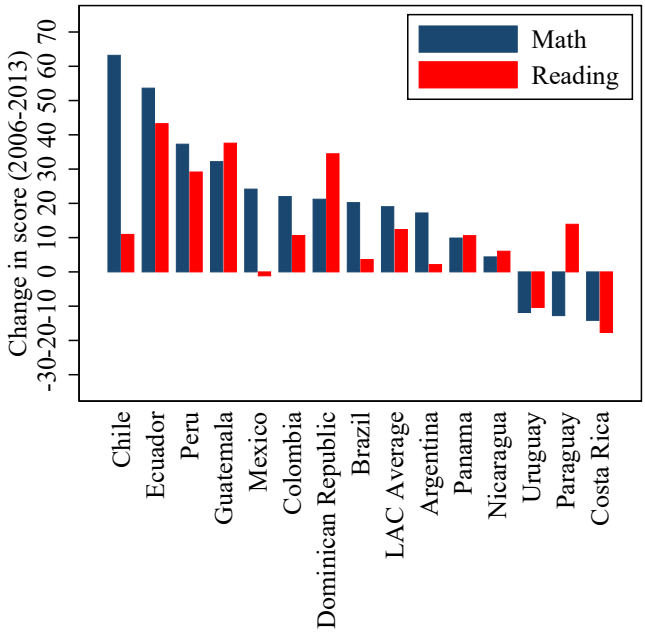
Progress in raising learning. One of the most striking patterns in cross-national learning data for Latin American countries is the rapid improvement in some countries versus stagnation and/or declines in others. Some of the countries with the highest learning outcomes at the start of the 2000s, such as Argentina, Uruguay, and Costa Rica, and more recently Brazil, have seen little gain or even declining performance on both the Latin American regional test and PISA. But another set of countries has made remarkably large and sustained improvements in learning outcomes over the same period. These countries include Chile, Peru, Ecuador, Guatemala, and the Dominican Republic (from a very low base).

Figure 1: Change in Test Scores in Latin America

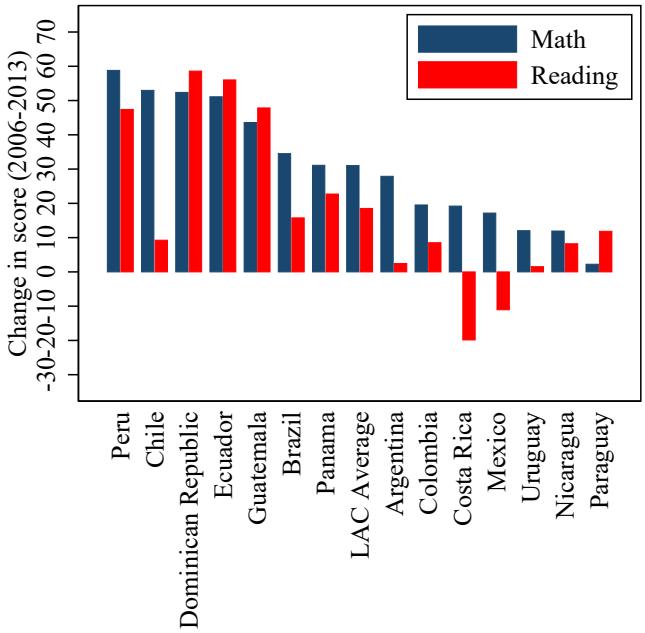
(a) PISA



(b) Latin American Regional Test (Sixth Grade)



(c) Latin American Regional Test (Third Grade)



Note: In Figure 6 (a) change for Argentina is from 2009-2012, as the country did not report scores in 2015. Change for Trinidad and Tobago is from 2009-2015, as the country did not report scores in 2012.

Source: PISA 2015: Results in Focus, PISA 2012: Results in Focus, PISA 2009: Results in Focus, and First Release of Results: TERCE.

Although the Dominican Republic continues to show the lowest performance in the region on both PISA and the Latin American regional test, at all three levels of education in all subjects, it is making large improvements in learning and is closing the gaps. Chile is the one country in the region that started from a relatively high base and has continued to improve strongly, especially in math. Large learning gains have been registered by Peru and Ecuador over the past decade, on both PISA and the regional assessment, starting from a very low base in both cases. Ecuador has only one data point for PISA, having joined only the last (PISA-D 2018) round, but its scores were the highest of the PISA-D group (most of which were lower-income countries) and are now on par with Peru and Brazil—two countries that it had trailed behind on the Latin American regional test in 2006. Guatemala showed large learning increases in both reading and math in both third and sixth grades between SERCE and TERCE, and while its performance on PISA-D was about 40 points (equal to one year of schooling) below Ecuador's, it has also been on a clear upward trend.

Chile has the region's longest and broadest history of cross-national learning measurement—it is the only country in the region to participate in TIMSS and PIRLS in addition to PISA and the Latin American regional test. But the other four countries have made a perceptible "pivot to learning" in the past decade, by increasing their participation in cross-national learning measurement, as well as strengthening their national assessment systems. Peru re-joined PISA in 2009 after dropping out of two cycles; the Dominican Republic joined PISA in 2015; and Guatemala and Ecuador participated in PISA-D in 2015. By joining PISA-D, Paraguay can also be included among countries that have opened their education systems to more extensive cross-national benchmarking in recent years. At least through TERCE 2013, there is less evidence of learning improvement in Paraguay; it will be interesting to see if the ERCE 2019 results show a different picture.

3.2 Has Cross-National Learning Measurement Informed Education Policy in Latin America?

Latin America stands out as "the first region of the [developing] world to undergo a big, concerted push to establish learning assessment systems" (Ferrer and Fiszbein, 2015). A regional assessment has served virtually all Spanish-speaking countries in the region, as well as Brazil, for over twenty years, and over this period almost all countries have developed robust national assessment systems.[12] With few exceptions, these national assessment systems today deliver reliable and standardized measurement of learning progress.

The regional assessment, LLECE, was inspired in 1994 when Ernesto Schiefelbein became Minister of Education in Chile[13] and publicized a national assessment showing that only half of Chilean fourth-grade students could comprehend a short text. UNESCO's regional office used the Chile results to open a dialogue with other countries in the region on the value of learning measurement and the Inter-American Development Bank agreed to support the development of a regional assessment.

Under UNESCO's technical leadership, LLECE was applied in 13 countries in 1997. A second-round test (SERCE) followed in 2006. A third round (TERCE) was applied in 2013, and the current round (ERCE) is being implemented in 18 countries in 2019. While a few countries have come in and out of participation, the test has served a fairly stable pool of 13-18 countries over the past twenty years (Figure 2).

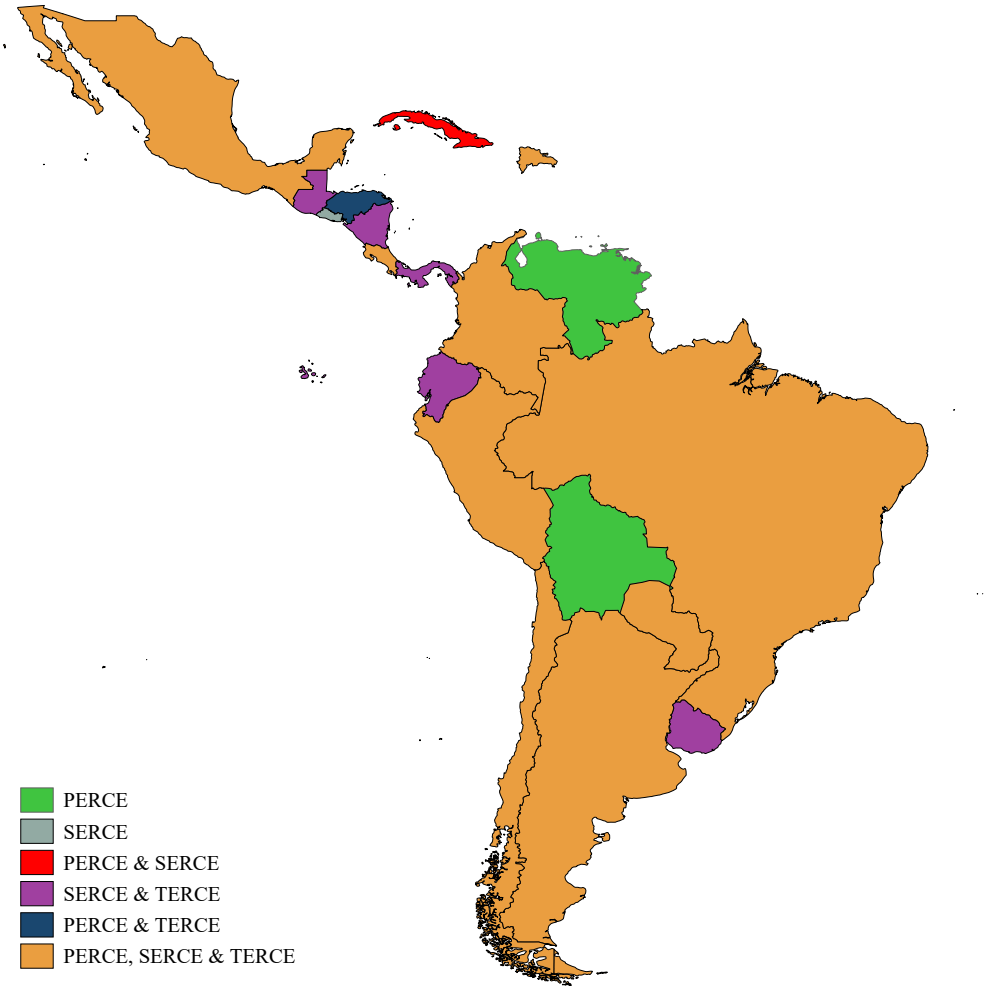
Five different studies have examined the impact of Latin America's regional test on countries in the region (McMeekin, 2003) (Ferrer, 2006) (Ravela et al., 2008) (Solano-Flores and Bonk, 2008) (Ferrer and Fiszbein, 2015). These studies concur on three positive impacts:

Stronger national capacity. UNESCO drew international experts into collaboration with national coordinators on the design of the first iteration of the regional test, PERCE,

[12]The region's 13 English-speaking Caribbean countries, however, have remained outside of this process and collaborate through the Caribbean Examinations Council. Haiti has also remained outside.

[13]From 1991-1994, Schiefelbein headed UNESCO's regional office for Latin America and wrote several research papers on education quality issues in Latin America and the Caribbean.

Figure 2: ERCE Coverage Across Latin America



which measured reading and math competencies at the third and fourth grade levels. It was applied in 13 countries between June and November 1997.[14] In 2006, a second round, SERCE, also tested reading and math but shifted to Grades 3 and 6 (Grade 6 also tested science). SERCE was applied in 13 countries and the Mexican state of Nuevo León.[15] A third round in 2013 (TERCE) tested the same grades in 15 countries (Cuba dropped out) and added measurement of writing skills.[16] A fourth round (now called ERCE) that tests language, math, science, and writing in Grades 3 and 6 is being applied in 2019, with 18 countries participating.[17]

A 2008 evaluation conducted a survey of the participating countries and concluded that the initiative "represented a first step towards the creation of a culture of testing and accountability in Latin America" where there had previously been little expertise in large-scale testing (Solano-Flores and Bonk, 2003). It is significant that LLECE developed at a time when many countries in the region did not have functioning national assessment systems. While several countries in the region began to launch national learning assessments in the 1990s, very few systems were institutionalized (Ferrer, 2006). Testing regimes were started in the early 1990s in Ecuador, Peru, Bolivia, Honduras, Guatemala, and Paraguay, but all fell into dysfunction after one or two cycles when World Bank, Inter-American Development Bank, or USAID funding ended.

In this context, the LLECE "Laboratorio" played an important role as "a funded organization with a structure providing training and experience in international achievement testing to teams from all participating countries" (Ferrer, 2006). Cross-national teams worked together to define the curriculum and learning domains to be measured, the appropriate range

[14]PERCE countries include Argentina, Bolivia, Brazil, Colombia, Costa Rica, Chile, Cuba, Honduras, Mexico, Paraguay, Peru, Dominican Republic, and Venezuela.

[15]SERCE countries include Argentina, Brazil, Chile, Colombia, Costa Rica, Cuba, Ecuador, Dominican Republic, El Salvador, Guatemala, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay, and Nuevo León (Mexico).

[16]TERCE countries include Argentina, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay, and Nuevo León (Mexico).

[17]ERCE countries include Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Paraguay, Peru, and Uruguay.

of learning levels in each domain, and collaborated on the development of test items. Through sustained engagement with the Laboratorio and "learning by doing," countries developed a critical mass of testing expertise. Regional expertise was boosted significantly further in the Latin American countries that began participating in PISA after 2000.[18] Latin American countries' national assessment instruments have become more sophisticated psychometrically over time and their sampling and administration protocols guarantee standardized measurement.

Ferrer and Fiszbein (2015) point to several other areas of "clear progress" in the institutional depth of national testing systems in Latin America. Many national systems have acquired stronger and more formal legal status (in many cases as autonomous organizations outside of direct control of the Ministry of Education). Countries such as Brazil and Colombia have innovated ways of communicating learning results more powerfully to schools and parents, and of following individual students' learning gains through university-level education.

Brazil's Basic Education Index (Índice de Desenvolvimento da Educação Básica, IDEB) combines fifth, ninth, and eleventh grade test scores with other key outcomes, such as repetition and dropout rates, into a single value on a scale of 1-10 that allows for transparent ranking of individual schools as well as Brazil's 5,000 municipal and 27 state education systems. The IDEB index is credited with stimulating states and municipalities to adopt innovations from those making faster IDEB progress, such as school-based bonus pay linked to achievement of IDEB targets, which has now diffused across a large number of states and municipalities in Brazil (Bruns et al., 2012).

Another example is Colombia's SABER assessment, administered in primary school and the final year of secondary school and repeated at the end of higher education. It is unique in the region in allowing for transparent measurement of the learning value-added from different university courses of study and different universities.

[18]PISA countries include Argentina, Brazil, Chile, Colombia, Mexico, Peru, Uruguay.

Focus on learning. To Ferrer and Fiszbein (2015), the most important progress is the impact that regional assessment data have had on public opinion and society. "Assessment results have helped to install the issue of education quality—understood in terms of learning results—in national and transnational debates in Latin America." By 2015, they observe, "most national education goals (in Latin America) include learning results as a key indicator for measuring quality and progress over time" (Ferrer and Fiszbein, 2015). World Bank and Inter-American Development Bank education projects in Latin America today routinely include improvements in learning outcomes as program goals and impact measures.

Stimulus to system reform. National coordinators surveyed by Ferrer and Fiszbein (2015) report that having a common metric for understanding how their country's educational system stacked up against its regional counterparts contributed to "identifying problem areas in the system and carrying out reforms in light of these findings." They note that LLECE also offered member states technical input to support education policy decisions and policy reforms and that a 2004 Inter-American Development Bank report detailed specific examples of reforms inspired by the first LLECE results (McMeekin, 2003).

Compared with PISA, however, there are relatively few examples of LLECE results provoking a "shock" in countries that stimulates the adoption of reforms. The one documented case of SERCE 2006 results being used actively by politicians in a communications strategy to legitimate reforms is that of Ecuador (Schneider et al., 2017) (Bruns and Luque, 2015), discussed in Section 3.3.1.

Summing up, the stable twenty-year history of the Latin American regional test has by all accounts had positive impacts on education ministries and education policy in the region. Participating countries developed assessment expertise through "horizontal" collaboration in comparing curriculum standards and learning domains, identifying different levels of learning, generating test items, and developing representative samples. The first round (PERCE) regional test results offered the first systematic data on student learning in a number of countries; it also offered those countries immediate perspective on how their systems were

performing. Over the next 20 years, virtually all the participating Latin American countries put in place technically robust and sustainable national assessment systems. The low cost of the regional program—approximately $300,000 per country per round—made participation an efficient investment. An important asset of the Latin American regional test today is that its core structure—assessing reading, math, and writing skills in Grade 3 and Grade 6—aligns directly with the first two SDG-recommended points of measurement. It is the only regional assessment that does so.

3.3 Have Cross-National Learning Data Stimulated Learning Improvement in Latin America?

We contribute to the evidence on this question by looking at the role of cross-national learning data in two cases of education reform in Latin America that have produced notable learning gains over the past decade: Ecuador and Peru.

3.3.1 The Case of Ecuador

Over a period of enormous political instability in the late 1990s—seven presidents and nine ministers of education in the ten years before 2005—Ecuador's education system became badly degraded. It was the only country in Latin America where education spending fell from 1990 to 2000, dropping from close to three percent of GDP to a shockingly low one percent of GDP. The primary completion rate was stagnant. Less than half of secondary-age students were enrolled. A sample-based national assessment system supported by the Inter-American Development Bank in the early 1990s was abandoned. In 1997, Ecuador was the only major South American country that declined to participate in the first Latin American regional test. But by 2006, with the low quality of education becoming a political issue and inspiring a national referendum, Ecuador signed on to the second regional assessment (SERCE).

In 2008, the SERCE results showed Ecuador near the bottom in every grade and subject tested, on the level of much poorer countries (Estarellas and Bramwell, 2015). The failure of

the education system to serve Ecuador's children became the political rallying cry of Rafael Correa's campaign for re-election in 2009, with the message that the country needed deep education reforms (Schneider et al., 2017). Over the next four years, Ecuador began to transform its basic education system in two key ways. First, it made learning the central goal and learning measurement the central metric of system progress. Second, it focused on teaching quality as the core strategy for raising learning.

Learning measurement. Several years of preparation culminated in the 2012 creation of Ecuador's National Institute for Education Evaluation (INEVAL) as an autonomous body in charge of all assessment processes, including teacher performance evaluation and school management. Learning is measured in Grades 3, 6, 9, and 12, and a summative evaluation at the end of high school also functions as the national university entrance exam (ENES, Examen Nacional para la Educación Superior). The national assessment is applied annually under an interesting format—administration is carefully controlled in a relatively small nationally representative sample of schools in order to generate reliable, standardized results for system monitoring. But a variant of the test is given in all schools (census basis) so that every school receives learning results that they can use to benchmark their performance and diagnose and address issues. Also unique among Latin American countries is the fact that the learning assessment is mandatory for all students and counts for 30 percent of a student's final grade in the subjects tested.

Ecuador also participated in TERCE in 2013, and in 2016, it joined PISA-D, to benchmark learning progress internationally. TERCE results showed Ecuador's learning gains were the largest in the region, and the PISA-D results released in 2018 showed that the learning levels of Ecuador's 15 year olds in reading, math, and science are now on par with Brazil and Peru—countries Ecuador had trailed badly a decade earlier. Starting from no learning measurement, there is no question that tracking and cross-national benchmarking of learning progress took on a central role in education policy after 2008.

Raising teacher quality. The Correa government's reforms to raise the quality of

teaching were politically contentious and technically bold. In 2009 the core teacher law was re-written, to eliminate teachers' union control of teacher hiring and Ministry appointments, to raise the bar for teacher quality at entry, and to make teachers in service more accountable for performance. Teacher hiring was now based on competency tests and clear standards; all teachers were subject to performance evaluation at regular intervals that included assessment of their classroom practice; promotions were based on performance evaluations rather than years of service; and dismissal was mandatory after two successive poor evaluations, notwithstanding teachers' civil service status. The union lost its role in the appointment of school directors, and directors were no longer granted lifetime tenure with no performance evaluation; they were required to compete for re-appointment every four years. These policies were strengthened further in the 2011 Education Law, which took the additional steps of closing low-quality teacher training institutes and raising the minimum university entrance score for teacher training candidates to the same level required for medical school.

The 2009 reforms provoked strong resistance from the teachers' union, but the government reacted equally strongly with actions to strip the union of automatic funding from teacher salary deductions, eliminate the right to strike with pay, reduce union influence in the Ministry, and, above all, its control over teacher appointments. Reforms were pursued with high continuity over Correa's ten-year period in office. A single team (a minister succeeded by vice minister) led education reform over the first six years of this period. Under technically strong ministers, a cohesive and technocratic implementation team was built for teacher evaluation, student testing, planning, infrastructure management, and other core functions. Ministry technical staff participated in regular sessions with the economist President that left no doubt of his personal strong commitment to an "education revolution."

Ecuador's progress was substantially helped by economic tailwinds, a commodity boom that allowed the country to increase education spending dramatically, from one percent of GDP in 2000 to more than five percent (of an expanded GDP) in 2014. With expanding resources, the government was able to build new schools, upgrade infrastructure, and expand

enrollments at all levels, while investing in books and materials, a new teacher training university, and doubling teacher salaries. These resources also allowed the government to adopt an attractive incentive for early retirement that greatly sped the impact of the reforms to raise teacher quality. Over a four-year period after 2008, more than 40,000 teachers and directors retired, making room for younger teachers selected on higher standards, and reform-minded directors—this represented a 25 percent turnover of the teaching force in the space of four years. After the house was cleaned, the government consolidated the political support of new teachers and directors by doubling salaries. Unique features of Ecuador's political and economic context greatly facilitated the reform program and make the intensity of implementation difficult to replicate. But the program is also marked by a political legacy from Rafael Correa that may undermine its ultimate sustainability.

What can we say about Ecuador's experience from the standpoint of learning measurement in general and cross-national testing in particular? Clearly, a period of disastrous education performance created a deep national appetite for educational change, which was tapped by Minister Vallejo (under President Palacios) in 2006 and President Correa after 2007—even without any national or international learning data. But interviews and Ecuador's reform trajectory suggest that the Latin American regional test, SERCE, influenced education policy through three of the channels identified by Fischman et al. (2018): i) a "shock" factor causing the country to think more actively or more ambitiously about needed reforms after 2008, ii) cross-country policy borrowing, through the report by Carnoy et al. (2007) explaining Cuba's superior SERCE performance and subsequent study tours there, and iii) a communications tool for government reformers like Rafael Correa and Ministers Gloria Vidal and Freddy Peñafiel to justify their programs.

Media reports from the time indicate that Correa's communications strategy consistently used the SERCE regional test results to make the case for reform (Schneider et al., 2017). Opinion polls also indicated that Correa and team's communications strategy was successful in mobilizing public support for the reform program, with over 70 percent of respondents in

2011 expressing satisfaction with the education system, up from 40 percent in 2007 (Schneider et al., 2017). When the TERCE results in 2013 showed that Ecuador had one of the largest learning gains in the region in both reading and math, both the President and Minister Espinosa used these as evidence that the country was on the right path.[19] Consistent with their reform strategy focused on teaching quality, communications about the results gave full credit to Ecuador's teachers.[20]

It is clear that the country made a "pivot to learning"—from no learning measurement prior to 2007 to building a high-quality national assessment system by 2012. "For good policies, we need good information" was the mantra of President Correa, who took the decision to join PISA-D in 2014, at a time (before the TERCE results came out) when the assessment team believed they were taking a big risk.[21] The PISA-D results confirm that Ecuador has made very significant, sustained learning gains over the past ten years. Its improvement on TERCE, confirmed by PISA-D, represents learning gains of roughly more than a year of schooling for primary and secondary school students. Ecuador today also has one of the best developed national assessment systems in Latin America; it is one of relatively few countries to implement a census-based assessment system at several different grade levels, which allows for feedback on learning to be provided to every school, district, and region in the country.

3.3.2 The Case of Peru

Along with 11 other developing countries, Peru participated in the 2001 round of PISA (called PISA-plus). The results were abysmal. Peru's 15 year-olds performed almost 100 points lower than those in Chile, Argentina, and Mexico (approximately two and a half years of schooling behind), and 90 points below Mexico and Chile in science (more than two years behind). Peru was 200 points behind the OECD average in math (five years of

[19]Survey response from former Minister Freddy Peñafiel.
[20]Survey response from former head of INEVAL, Harvey Sánchez Restrepo.
[21]Ibid.

schooling). Peru was the only country of the 43 that participated in PISA where over half of the students (54 percent) scored below the lowest level in reading, Level 1, compared with six percent of OECD students.

These results provoked a reaction. The Prime Minister called a "Declaration of Emergency" and the Ministry of Education responded with a plan that included an "emergency program focused on literacy", actions to raise teacher quality, decentralization, and a "new pedagogical strategy focused on teaching and learning," among other areas. The Ministry also began the development of a national early grade test of reading and math to catch learning problems to be applied in monolingual schools at the end of second grade, and in bilingual schools at the end of fourth grade. Also in response to the 2001 "shock," the Ministry withdrew from PISA's 2003 and 2006 rounds, believing that the level was too demanding and that its administration would divert the Ministry's limited technical capacity away from developing the national assessment.

When Alan Garcia was re-elected President in 2006,[22] his inaugural address signaled a focus on education, calling for expanded student assessment and decentralization. In 2007, the government adopted a major reform to raise teacher quality, establishing a new teacher career path (CPM, Carrera Publica Magisterial) with higher standards, pay, and accountability for performance. All hiring was now based on test performance; teachers in service would undergo regular performance evaluations, with the possibility of dismissal for consistent poor performance; and promotions would be governed by performance rather than be automatic with years of service. The government also set higher standards for admission into teacher training colleges (ISPs). The content of Peru's teacher reform was very similar to that of Ecuador's, happening at almost the same time. But Ministry staff report that the major influence on Education Minister Chang was a McKinsey and Co. (2007) report that, drawing on an analysis of the highest performing countries in PISA, emphasized the importance of recruiting high-quality teachers.[23]

[22]Garcia served a first term as President from 1985-1990.

[23]Survey response from Liliana Miranda, UMC Director, 2004-2016.

As in Ecuador, the reforms were met with strikes and protests from the teachers' union. But Garcia also was a skilled communicator and launched an active campaign to convince the public that Peru's learning crisis demanded major reforms. By 2007, polls showed that 74 percent of people in Lima thought the government's CPM would be good for students and good for teachers, and 46 percent believed that union resistance was hurting education. Relentless communication moved public opinion even more strongly behind the government by 2012, when 62 percent believed that resistance from teachers was the main threat to education quality (Bruns and Luque, 2015).

In the face of union resistance, the CPM was made applicable only to new hires; existing teachers could opt in. But even substantially higher salaries did not persuade existing teachers to enter the new regime—over the final three years of the Garcia administration, only ten percent of existing civil service teachers submitted to the competency tests required to access the CPM.

In 2009, Peru re-joined PISA at the urging of the Ministry's learning assessment unit (UMCE, Unidad de Medición de la Calidad), which made two arguments: the team would grow professionally from the interaction with the OECD assessment experts and Peru would benefit from having comparative data to benchmark its progress. The 2009 results showed clear evidence of learning progress. Peru's reading score increased by more than 60 points (one and a half years of schooling). But the country remained in last place among the eight Latin American countries that took the test, and Ministry staff were chagrined that evidence of progress was lost in the media reports over the rankings.[24] One of the strongest critiques of cross-national tests in the international literature is that rankings, which provide little useful information for policy, are the most extensively cited and used results.

After its bitter fights with the Garcia administration over the teacher career path reform, the teachers' union mobilized on behalf of the opposition party in 2011 and helped elect Garcia's left-of-center opponent Ollanta Humala. But rather than reverse the reform, Humala in

[24]Ibid.

fact extended it—passing the LRM *(Ley de Reforma Magisterial)* in late 2012, which made the new career path mandatory for all teachers. Despite further strikes and resistance, the government made no concessions on the content of the reforms, but the LRM incorporated a further increase in salaries.

In 2013, Humala named Jaime Saavedra minister of education. Saavedra carried forward the teacher reform, put further emphasis on learning outcomes, invested in infrastructure and bilingual education to reduce disparities in school quality across regions, and put an emphasis on efficient management. Saavedra was also an excellent communicator and when the 2012 PISA results again showed Peru in last place among the 65 countries tested, he used these results in every speech to argue that Peru's "disastrous" situation called for major reforms; he added that spending on education had to increase as well and mobilized an additional one percent of GDP (a 30 percent real increase in the education budget). Saavedra's communication skills and reputation for efficient management led to a 2015 poll showing that 94 percent of Peruvians were satisfied with his performance, and in 2016, Humala's center-right successor reappointed Saavedra as education minister—an almost unheard of example of continuity in education minister across opposing administrations.

What can we say about learning measurement in general and cross-national testing in particular from Peru's experience? Interviews and Peru's reform trajectory suggest that in Peru's case, it was the PISA test that influenced education policy, through three of Fischman et al. (2018) channels: i) a "shock" factor causing the country to think more actively or more ambitiously about needed reforms after 2001, ii) cross-country policy borrowing from OECD countries, through the McKinsey and Co. (2007) report on PISA's highest performers, and iii) a communications tool for government reformers like Alan Garcia and Jaime Saavedra to justify their programs. The UMC director also reported that technical cooperation with the OECD team was very beneficial, and that alignment with PISA was an important influence on the competencies and learning standards that Peru adopted as part of its 2016 curriculum reform.

In Peru's case, the influence of the Latin American regional test was less strong, The UMC team noted that Peru's performance, in the middle of the regional distribution, did not attract as much attention in the media. Interestingly, even though Peru showed large learning gains on TERCE in 2013 for both third and sixth graders, Saavedra's political strategy focused on the negative 2012 PISA results for 15-year-olds. The view of the UMC is that the regional test's strongest impact was on capacity development in the late 1990s and early 2000s, as Peru implemented a series of sample-based national assessments before a census-based early grade assessment was launched. Collaboration around LLECE was described as more "horizontal" than with PISA, as Latin American countries played a much larger role in the definition of ERCE's learning domains, levels, and the development of test items—experiences that directly supported the development of Peru's national assessment system.[25]

4 Cross-National Learning Measurement and Education Policy in sub-Saharan Africa

4.1 Learning Measurement

The data gap against the SDG-mandated points of learning measurement is greatest in sub-Saharan Africa. Only ten of the region's 46 countries have cross-nationally comparable data for Grades 2/3 through their participation in an oral test administered at Grade 2 by PASEC (Programme d'analyse des systèmes éducatifs).[26] At the end of primary (Grade 6), there are data for the ten PASEC countries plus 15 countries that have participated in SACMEQ (Southern Africa Consortium for Monitoring Education Quality)[27] tests admin-

[25]Ibid.

[26]PASEC 2014 countries include Benin, Burkina Faso, Burundi, Cameroon, Chad, Congo, Cote d'Ivoire, Niger, Senegal, and Togo. The 2019 round of PASEC is expected to include Gabon, Guinea, Madagascar, Mali, and the Democratic Republic of the Congo as well.

[27]SACMEQ III countries include Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, South Africa, Swaziland, Tanzania (Mainland), Tanzania (Zanzibar), Uganda, Zambia, and Zim-

istered in 1995 (SACMEQ I), 2000 (SACMEQ II), and 2006 (SACMEQ III).[28] [29] For lower secondary, only two African countries to date have a cross-national measure of learning, Senegal and Zambia, which participated in PISA-D, the results of which were released in 2018.

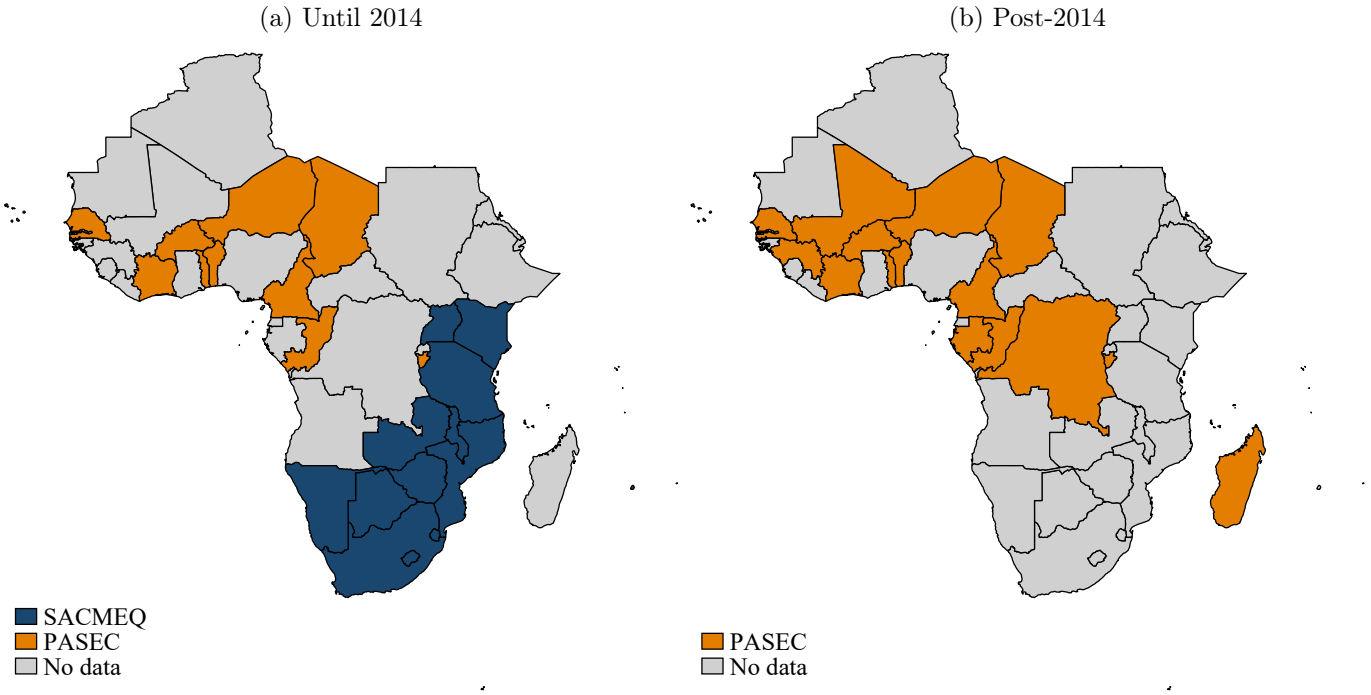
Figure 3 illustrates the extent to which countries in sub-Saharan Africa are covered by a regionally benchmarked test of learning. There are big gaps, with only about one-half of all countries covered by one or the other of two regional assessments (SACMEQ in English-speaking countries and PASEC in francophone countries), which test children in primary school grades. SACMEQ has suffered from a methodological controversy since 2016 regarding the results from its most recent round, and shows little sign of further expansion. PASEC on the other hand, is slowly expanding country coverage. Both tests are grade-based and face issues of comparability, given widely differing age distributions of children in the same grade across different African countries (Figure A.0.3 in the Appendix). For example, SACMEQ test takers in Grade 6 have tended to be older in Tanzania than in Uganda, and Tanzania has a wider distribution of age groups within the same grade. Therefore, even within PASEC and SACMEQ, there are issues of comparability across countries, highlighting one of the rationales for age-specific rather than grade-specific testing (Birdsall et al., 2016).

babwe.

[28] SACMEQ carried out a fourth round of testing in 2012-2013, but results have never been released publicly. Technical concerns have been raised about the comparability and validity of the results, including inconsistencies in the methodology used to calculate the results for 2012 round, and the possibility that weaker students were excluded from some of the results for the 2012 round (Spaull, 2016).

[29] Each SACMEQ round was administered over multiple years so we use the first year to refer to a particular round. For example, SACMEQ I was implemented from 1995-1999, SACMEQ II was implemented from 2000-2004, SACMEQ III was implemented from 2006-2011, and SACMEQ IV was implemented from 2012-2014. We refer to them as the 1995, 2000, 2006, and 2012 rounds respectively.

Figure 3: Regional Assessment Coverage of African Countries: PASEC and SACMEQ



Note: Botswana and South Africa are covered by TIMSS and PIRLS in addition to SACMEQ. Kenya, Tanzania, and Uganda are covered by Uwezo in addition to SACMEQ. Senegal participated in PISA-D, the results of which were released in 2018, in addition to PASEC. Zambia participated in PISA-D, the results of which were released in 2018, in addition to SACMEQ.

Source: Data is from SACMEQ III (2006) and PASEC 2014.

Tanzania, Kenya, and Uganda have also collaborated with Uwezo, a citizen-led assessment that measures basic literacy and numeracy. Citizen-led assessments such as ASER in South Asia and Uwezo in Africa have been successful at highlighting the problem of low learning. However, because neither the item design nor the application conditions are standardized, results are not directly comparable across countries or over time. Furthermore, since these tests are top-coded at a very basic measure of literacy and numeracy, they provide relatively limited information about learning progress over time, even within a single country.

The history of cross-national testing in Africa is actually longer than that of Latin America. Both PASEC and SACMEQ were launched in the early 1990s, inspired by the 1990

Education For All Conference in Jomtien, Thailand.[30] Both programs have also built substantial technical capacity, produced well-designed instruments, and enjoyed significant donor support. But they have not followed parallel trajectories in terms of testing priorities and country engagement, which has made collaboration difficult.

PASEC. At the 42nd Conference of Ministers of Education of French-speaking Countries (CONFEMEN) in 1990 in Bamako, ministers agreed to establish a joint program to measure student learning (Wagner, 2011). PASEC was established in 1991 to assess the performance of education systems in francophone Africa through large-scale surveys, learning measurement, and technical support to countries for the development of national assessment systems.[31] Tests were first implemented in Central African Republic, Congo, Djibouti, Mali, and Senegal from 1991-1995. In many cases, tests were applied at the beginning and end of a single school year, permitting the measurement of teacher value-added at the classroom level (i.e., controlling for differences in student ability).

Over PASEC's first two decades, it was not strongly focused on producing a standardized cross-national assessment of learning. The core mission was to provide technical support for developing national assessments, and PASEC extended support to francophone countries outside of West Africa, including Djibouti, Lebanon, Cambodia, Vietnam, and Laos. In 2012, CONFEMEN refocused PASEC on development of a regional assessment to provide standardized measurement of learning in math and language at the second grade and sixth grade levels. PASEC's first round test was applied in ten countries in francophone West Africa in 2014 and is being applied to an expanded sample of 15 countries in 2019. PASEC 2019 also introduced an in-depth teacher survey and tests of teacher knowledge in reading, mathematics, and pedagogy. PASEC also surveys school directors and collects data on the school and classroom environment and teaching conditions. PASEC is administered in both national languages and French (and English in Cameroon). The cost is approximately

[30] World Declaration on EFA adopted in Jomtien in 1990: "Every person—child, youth and adult—shall be able to benefit from educational opportunities designed to meet their basic learning needs."

[31] Further details about PASEC's origin can be found here.

600,000 euros per country, with CONFEMEN financing 50 percent and countries responsible for the other 50 percent. One-quarter of this amount supports national capacity building.

SACMEQ. SACMEQ began in Zimbabwe in 1991, when the Minister of Education and Culture requested support from UNESCO's International Institute for Educational Planning (IIEP) to develop a sample-based assessment to measure the quality of primary education. Several other ministries of education in Eastern and Southern Africa took interest in the initiative and in 1994, Zimbabwe invited education planners from several neighboring countries to collaborate in writing a policy report based on the test results. The unusual experience of planning officers involved in writing a report on another country generated a lot of interest, and in 1994 seven countries[32] established a formal group called the Southern Africa Consortium for Monitoring Education Quality (SACMEQ). Inspired by the Jomtien goals, SACMEQ's mission was to build local capacity for evaluation of education quality and to generate information to guide policy (Murimba, 2005). The consortium is also known as SEACMEQ to reflect the participation of East African countries.

With support from UNESCO's IIEP and Australia's ACER, SACMEQ developed a high-quality sampling framework and instruments for measuring pupil characteristics, teacher characteristics (including teacher knowledge), characteristics of school leaders, infrastructure, equity across regions and schools, and students' reading and math achievement. Math items from TIMSS were included in the 2000 and 2006 testing rounds to allow SACMEQ's scale to be linked to that of TIMSS.

SACMEQ's first-round report in 1998 expressly avoided cross-country comparisons when results were released, primarily due to political sensitivity (Murimba, 2005). By 2004, however, the 14 countries which participated in SACMEQ II (2000-2004) accepted the presentation of cross-national results. SACMEQ III (2006-2011) and SACMEQ IV (2012-2014) also involved 14 countries and provide some basis for comparing trends as well as levels of performance. But relatively little of this has been done. Because of controversies over the

[32]SACMEQ I countries included Kenya, Malawi, Mauritius, Namibia, Tanzania (Zanzibar), Zambia, and Zimbabwe.

handling of SACMEQ IV data, official results have never been released. IIEP, a long-time technical partner of the program, is no longer involved and the four-to-five year cycle has been suspended. There are reports that SACMEQ is planning a new round of testing in 2021, which is welcome. However, the ten-year hiatus in cross-national testing results for a large and important share of countries in the region has been unfortunate.

In short, even with PASEC's expansion and the participation of Senegal and Zambia in PISA, without recent SACMEQ data, fully two-thirds of sub-Saharan African countries have no cross-national measures of learning in primary education and virtually all of the region lacks a cross-national measure of learning at the secondary level.

4.2 Has Cross-National Learning Measurement Informed Education Policy in sub-Saharan Africa?

Have data from cross-national assessments stimulated education policy reform in sub-Saharan Africa? Given that PASEC's first round of fully comparable cross-country learning data is quite recent and only offers one data point, it is perhaps not surprising that we could find no published cases of reform stimulated by the PASEC test.

But there is evidence that the SACMEQ regional test played a role similar to that of Latin America's ERCE in providing countries in the region with transparent comparative data on their neighbors' education performance that stimulated policy change. Gustafsson (2019) documents that the 2003 release of the 2000 SACMEQ results "shocked" South Africa by revealing how much worse the reading and math scores of South Africa's Grade 6 students were relative to countries such as Kenya and Swaziland, both of which spent much less per student. He notes that this was the first time concrete evidence on learning outcomes were available at an earlier point in the schooling cycle than Grade 12 examinations, and he links the SACMEQ shock to several key shifts in education policy, including a major curriculum reform in the early grades, national workbooks, a new national assessment system, and changes in rules governing teachers, school principals and charter schools, among other areas.

South Africa also joined the PIRLS and TIMSS international assessments and both tests provided evidence of significant gains in learning over the second half of the 2000s decade; PIRLS Grade 4 reading results improved by more than one year's worth of learning between 2006 and 2011, and Grade 9 mathematics results on TIMSS improved by one standard deviation between 2002 and 2015. Interestingly, Gustafsson has observed that South Africa's relatively poor results on TIMSS and PIRLS results did not have the same policy shock value because the other participating (mainly OECD) countries are much wealthier.[33] Seemingly consistent with this, we also found no reported evidence of education reforms in Ghana and Mauritius linked to their participation in TIMSS and PIRLS.

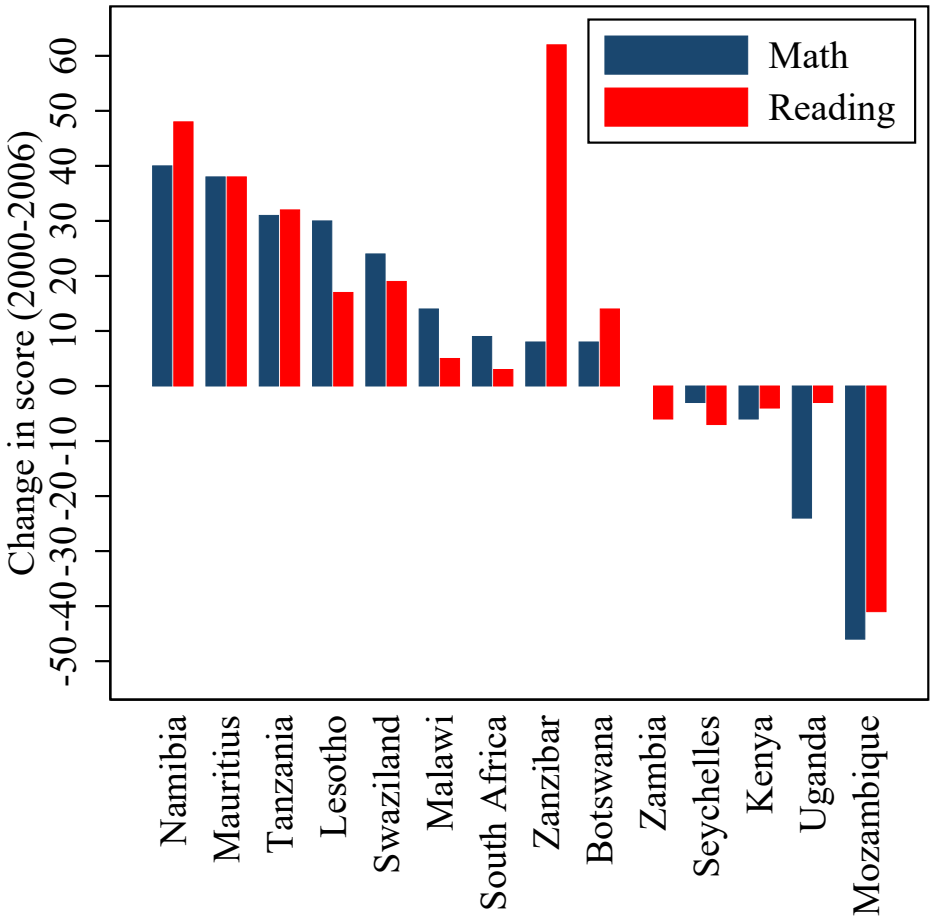
In contrast, there are several other documented cases of SACMEQ's impact on education policy. Murimba (2005) and Makuwa and Maarse (2013) find that Namibia's declining results between the 1995 and 2000 rounds of SACMEQ stimulated a reform agenda focused on greater equity in the allocation of education spending across regions and targeted support for language instruction. They also argue that these polices can be linked to Namibia's improved performance on the 2006 SACMEQ test. Leste (2005) documents actions by the Seychelles government to reverse its streaming policies after the 2006 SACMEQ results exposed one of the biggest gender gaps in the sub-region (with girls better than boys) in Grade 6 reading and math. Other countries reported to have used SACMEQ results to inform their education policies include Kenya, Zambia, and Malawi (Murimba, 2005).

4.3 What Do We Know About Learning in sub-Saharan Africa?

What about learning itself? Gross primary school enrollment rate in sub-Saharan Africa has expanded from 68 percent in 1970 to above 100 percent in 2010 (WDR, 2018); by 2017 net primary enrollment was 78 percent.[34] Despite this massive success in getting children to school, a significant proportion of African children go through years of schooling without acquiring basic skills in literacy and numeracy. The World Bank estimates that less than 50

[33]Email correspondence from Martin Gustafsson on November 11, 2019.
[34]Data accessed from UNESCO website on February 5, 2019.

Figure 4: Change in SACMEQ Scores



Note: Zimbabwe did not participate in the 2000 round of SACMEQ.

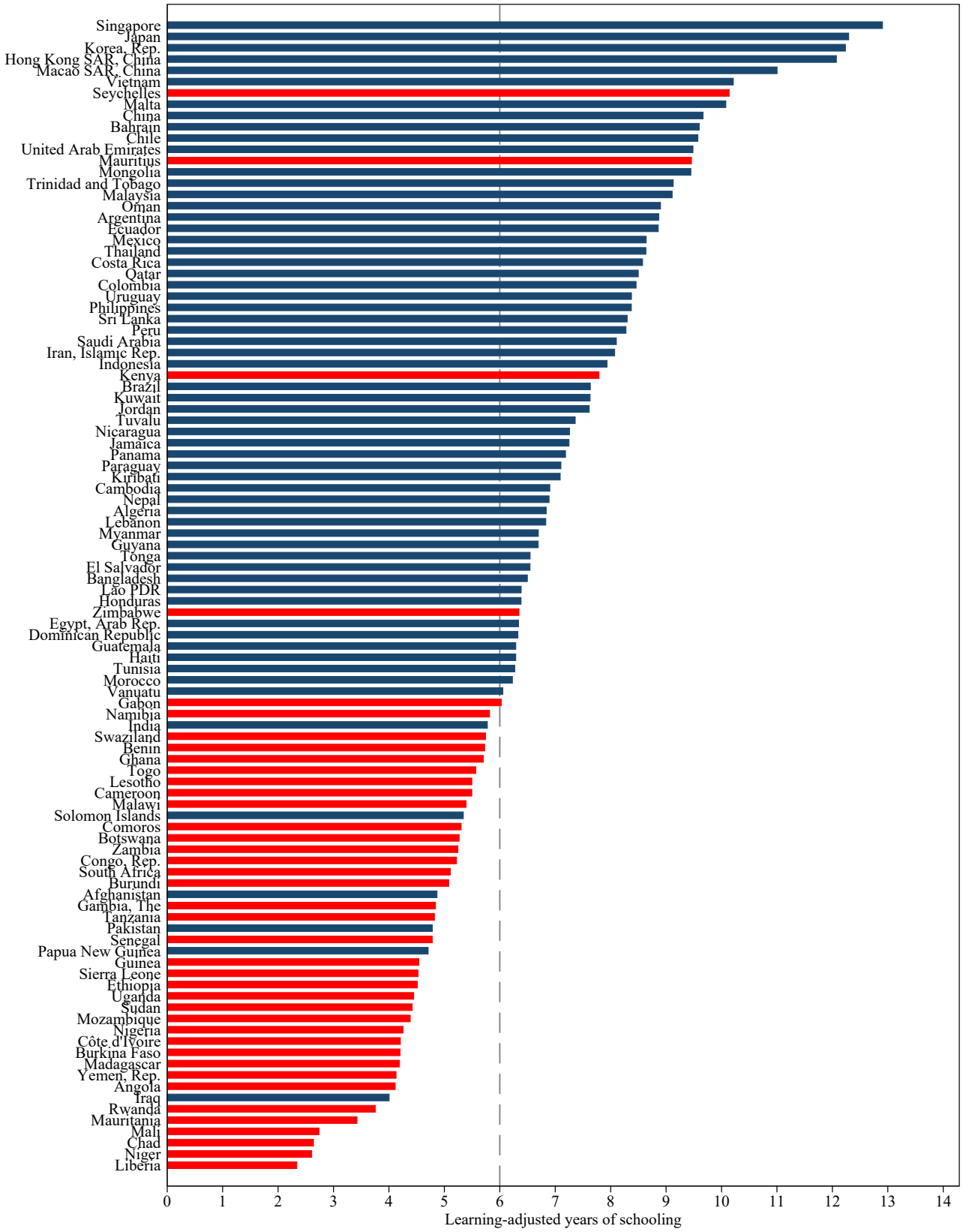
Source: SACMEQ Policy Issues Series, 2010.

percent of Grade 6 students in Southern and East Africa are able to do more than simply decipher words and less than 40 percent can go beyond basic numeracy (WDR, 2018). In francophone West Africa, only one out of four children completing Grade 5 reaches minimum literacy or numeracy benchmarks (Lilenstein, 2018). The fact that there is a learning crisis in sub-Saharan Africa is clear from the results summarized in Table A.0.2 in the Appendix.

Figure 5 uses learning-adjusted years of schooling[35] from the World Bank's Human Capital Index (HCI) data for all countries (excluding OECD, and Europe and Central Asia) and paints a similarly grim picture. The red bars represent African countries. A significant majority of African countries have below six learning-adjusted years of schooling on average—implying that most children in Africa do not receive the equivalent of a primary education. No Latin American country falls below this threshold.

[35]Learning-adjusted years of schooling combine both quantity and quality of schooling into a single metric and take into account the fact that productivity of a year of schooling varies widely across different countries. They are calculated by multiplying the estimates of "Expected Years of School" by the ratio of most recent "Harmonized Test Score" to 625, where 625 corresponds to advancement attainment on the TIMSS (Trends in International Mathematics and Science Study) test (See more details on the World Bank website). The World Bank defines complete education as 14 learning-adjusted years of schooling, which serves as a benchmark against which performance can be compared (Filmer et al., 2018).

Figure 5: Learning-Adjusted Years of Schooling



Source: Authors' calculations based on HCI data from the World Bank. Sample excludes countries classified as belonging to OECD or Europe & Central Asia.

We use available data on learning (described in Section 4.3.1) to address two sets of questions about the extent of the learning crisis: first, what does learning look like in sub-Saharan Africa compared to other regions, given GDP per capita, schooling levels of the parent generation, quality of governance, and other country-specific conditions; and second, what factors explain differences in learning among countries within Africa. In comparing sub-Saharan Africa to other regions, we focus especially on Latin America, where there has been much greater use of cross-national learning measures in the last two decades, and where there is suggestive evidence that consistent attention to measuring learning has contributed in the political arena to countries' focus on improving learning and willingness to adopt systemic education reforms.

4.3.1 Is Africa Different? Benchmarking sub-Saharan Africa's Performance

The limited available and directly comparable data shows that learning levels, on average, are lower in sub-Saharan Africa than in Latin America. In this section we explore what accounts for differences in learning results across regions, including between sub-Saharan Africa and Latin America and the Caribbean, using harmonized test score data from the World Bank's Human Capital Index (HCI). The data set includes 164 countries. We take the 2017 values for the harmonized learning score of a country.[36] Harmonized learning outcomes are produced using a conversion technique utilizing results from various international and regional assessments including PISA, TIMSS, PIRLS, SACMEQ, LLECE, PASEC, and EGRA, which are linked using "ratio linking" to compare scores on a similar scale (Patrinos and Angrist, 2018).[37]

We repeat the same exercise using Item Response Theory (IRT) equated test score data. We use fixed parameter IRT-equated scores from Sandefur (2018)—the paper uses IRT to

[36] We drop all countries classified as belonging to OECD or Europe and Central Asia from our regression sample.

[37] While psychometric linking is valid for formal, standardized cross-national assessments, it is problematic for EGRA, which is an oral test and not standardized (e.g., comparable over time). For ten countries, including Nigeria and Ethiopia, accounting for over 30 percent of Africa's population, the "harmonized test score" estimate in the World Bank database is entirely based on EGRA data points.

equate math scores for SACMEQ from 14 anglophone African countries and puts them on the international TIMSS scale, a globally benchmarked test implemented in 57 countries in 2015.[38] The results of this analysis are broadly comparable, but rely on a smaller sample—they are included in Section D.2 in the Appendix.[39]

The average harmonized test score from the World Bank's HCI for the high-income OECD countries is 520. Across regions, scores (in the range of 300 to 500 in sub-Saharan Africa, compared to the OECD average of 520) are associated with the average GDP per capita of countries in each region—a finding that is consistent with available international test results, such as PISA (Table B.0.2 in the Appendix).

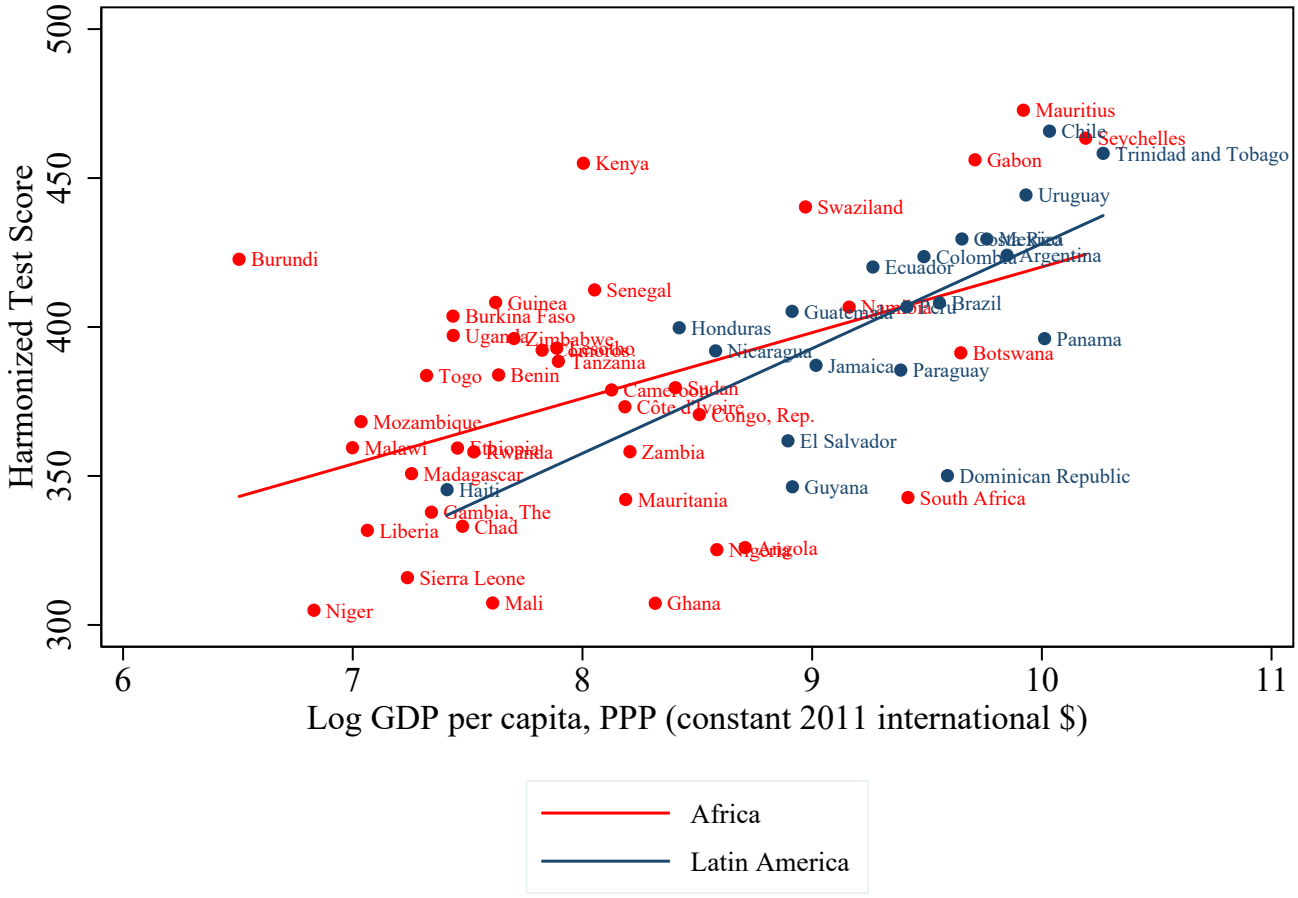
However, GDP per capita accounts for relatively little of the differences across countries within the two regions. Figure 6 provides a rough comparison of the harmonized test scores from HCI in sub-Saharan Africa and Latin America and the Caribbean for given per capita GDP. Higher GDP per capita is associated with the difference in average scores between the two regions, but little of the differences *among* countries, particularly among sub-Saharan Africa's relatively lower-income countries. That is even more the case in Figure 7, which shows the relationship between scores in the two regions and median income (the measure commonly used in Latin America) or consumption (the measure commonly used in other developing regions including sub-Saharan Africa). For most households in any distribution, consumption will be lower than income but at the low levels of consumption for many countries in Africa shown in Figure 7, measured household consumption is close to and can even

[38] In his paper, Sandefur (2018) uses three different approaches for equating scores: equipercentile equating, fixed parameter equating, and mean-sigma equating. While each method hosts its own set of challenges, we use the results from fixed parameter equating method, which relies on overlapping items across tests and uses IRT. Equipercentile equating relies on existence of data from multiple tests for a common population of pupils and does not rely on IRT, and mean-sigma equating relies on IRT but is commonly applied to link subsequent rounds of testing regimes. The fixed parameter equating method relies on the inclusion of TIMSS items in the SACMEQ test. The linking is possible because South Africa and Botswana participated in both TIMSS and SACMEQ and the 2000 and 2006 SACMEQ rounds embedded a number of items from TIMSS. We use equated pupil scores for SACMEQ III and TIMSS 2003 (Grades 7 and 8 scale). For South Africa and Botswana, we use the original TIMSS scores.

[39] We drop all countries classified as belonging to OECD or Europe and Central Asia from our regression sample.

exceed measured levels of income (Birdsall and Meyer, 2014).[40] Median income/consumption appears to explain even less of the differences across countries in harmonized test scores and especially less within low-income Africa.

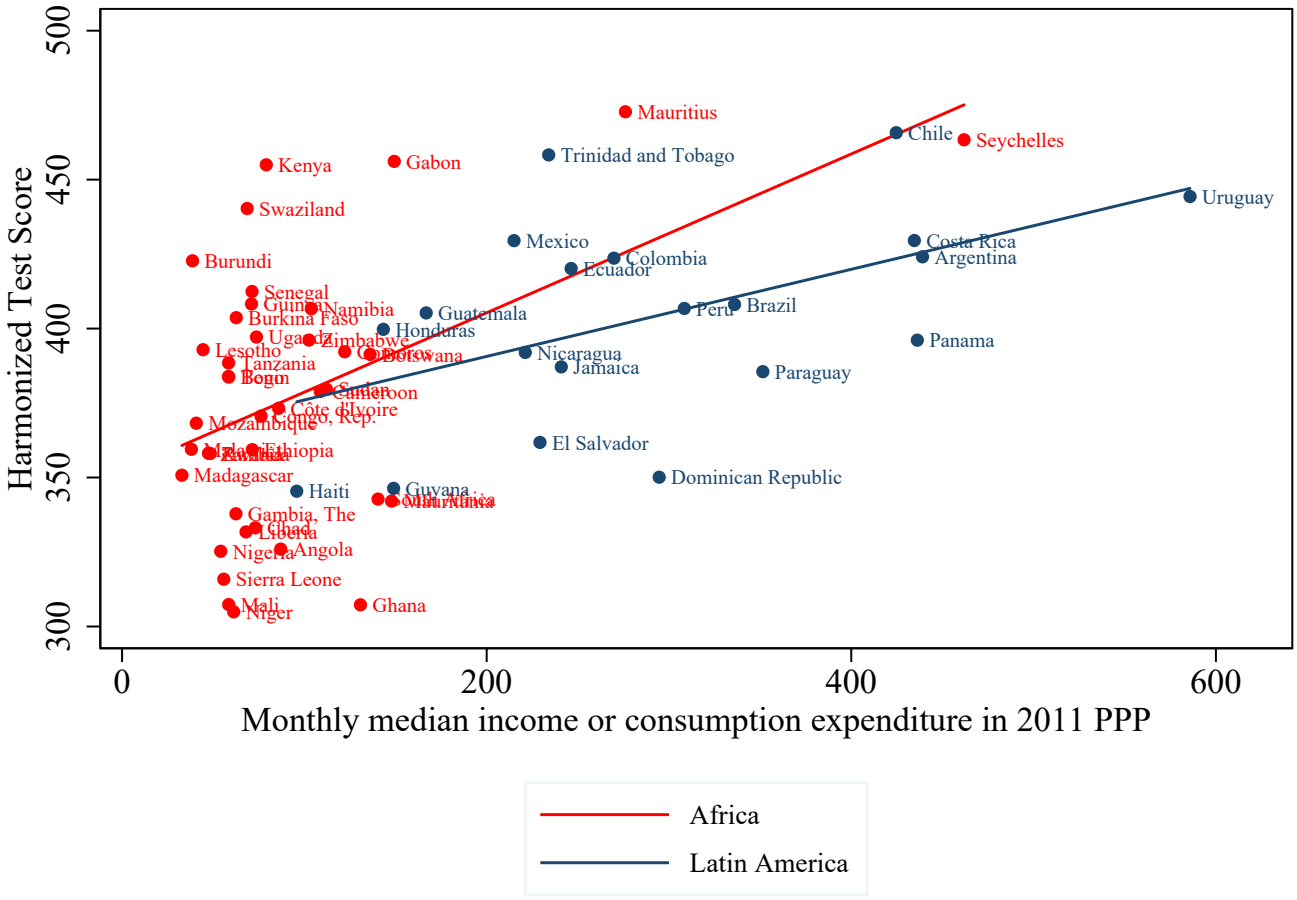
Figure 6: Harmonized Test Scores and GDP Per Capita: sub-Saharan Africa vs. Latin America



Source: Authors' calculations based on harmonized test score data from the World Bank's HCI and GDP per capita data from the World Bank's World Development Indicators. GDP per capita (constant 2011 international dollars and PPP adjusted) is for the most recent available year. We take natural log of the GDP figure. We drop all countries classified as belonging to OECD or Europe and Central Asia from our sample.

[40] Figure A.2 in the Appendix of Birdsall and Meyer (2014).

Figure 7: Harmonized Test Scores and Median Consumption/Income: sub-Saharan Africa vs. Latin America



Source: Authors' calculations based on harmonized test score data from HCI. Median consumption/income data are for available year in World Bank's PovcalNet in terms of 2011 PPP. We use consumption data when both income and consumption data are available for a country, and we use figures for "urban" when data is only available dis-aggregated by "urban" and "rural."

Table 3 shows the regression results of countries' harmonized test scores for 87 countries in five regions of the developing world, as a function of GDP per capita and other variables. Across all regions (with the Latin America and the Caribbean region the omitted dummy), GDP per capita does matter for scores; indeed Africa has on average higher scores taking into account Africa's relative poverty (Column 2), including compared to South Asia and the MENA region, and particularly taking into account its relatively lower expected years

of schooling (Column 3).[41] Results using the median consumption/income measures are similar.[42]

Table 3: Harmonized Test Scores Across Regions

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4	(5) Column 5
East Asia	61.849***	51.896***	52.719***	44.195***	42.843***
	(19.872)	(13.641)	(13.582)	(13.159)	(13.318)
MENA	-3.425	-25.130**	-13.019	-11.723	-14.192
	(13.970)	(10.268)	(11.315)	(10.406)	(10.369)
S. Asia	-39.765***	-6.231	-1.398	-7.018	-8.578
	(11.151)	(11.481)	(11.339)	(12.087)	(13.054)
SSA	-26.116**	18.785*	36.116**	24.765	12.854
	(11.327)	(11.250)	(15.266)	(16.142)	(13.891)
Log GDP Per Capita (2011 PPP)		34.832***	24.654***	15.979*	16.770**
		(5.061)	(6.955)	(8.149)	(8.091)
HCI Expected Years of School			7.049**	5.596	
			(3.498)	(3.639)	
BL Avg Schooling (20-29)					1.577
					(3.651)
BL Avg Schooling (40-49)			1.644	0.685	0.996
			(2.812)	(2.710)	(3.595)
Gov Effectiveness				19.661**	22.009***
				(7.880)	(8.242)
No. of countries	87	87	87	87	87
R-squared	0.32	0.60	0.64	0.66	0.65

The dependent variable is harmonized test scores from the World Bank's HCI. Regression sample drops countries classified as belonging to OECD or Europe and Central Asia, and is restricted to countries that have data on all the covariates included in the regression model. "East Asia," "MENA," "South Asia," and "SSA" are regional dummies that take the value of 0 or 1 depending on whether the country belongs to the region. Latin America and the Caribbean is the omitted category. GDP per capita (constant 2011 international dollars and PPP adjusted) is for most recent available year from the World Bank's World Development Indicators. We take natural log of GDP figure. "HCI Expected Years of School" are calculated using repetition-adjusted enrollment rates by school level to proxy for age-specific enrollment rates up to age 18 and taken from World Bank's HCI data. "BL Avg Schooling (20-29)" and "BL Avg Schooling (40-49)" are from Barro and Lee 2010, and are weighted by population. Government effectiveness is the 2017 figure from the World Bank's Governance Indicators. Asterisks (∗, ∗∗, ∗∗∗) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

[41] Expected years of school are taken from the World Bank's HCI data set and are calculated using repetition-adjusted enrollment rates by school level to proxy for age-specific enrollment rates up to age 18.

[42] Median income/consumption is more statistically robust (Columns 2-5) than GDP per capita (Table D.1.1 in the Appendix).

We include proxies for the possible effect of Africa's less educated older generations (more or less parents and grandparents) to test their effect as a handicap on children's scores relative to the Latin America and Caribbean region (that would disappear with near-universal primary school enrollment in Africa);[43] they do not seem to be relevant (Column 4). Africa's better performance disappears with inclusion of the government effectiveness variable (Column 5)—its country average is the worst of the five regions—suggesting the possibility that in African countries, education systems "perform" relatively well compared to those in other regions with overall more effective government institutions.

In short, controlling for its relatively lower income, lower current expected years of school, and weaker government effectiveness, sub-Saharan Africa is not particularly "different" from Latin America and the Caribbean in terms of the performance of its school children on test scores. Indeed, the differences across all the other regions in their absolute scores relative to Latin America and the Caribbean is small in Table 3, except for the consistently better performance of East Asia. Once factors over which education leaders have minimal short-term control are taken into account, the difference for sub-Saharan Africa compared to Latin America is just 13 points (Column 5), which is 3 percent of the average harmonized test score across this sample of countries.[44]

In Table 4 we add a country index of fragility and country index of (with-in country) linguistic diversity[45] to the right-side variables in Table 3. The effect of linguistic diversity is particularly notable. Among the regions, countries in Latin America and the Caribbean have the lowest average by far for linguistic diversity—at 0.17 compared to 0.67 for sub-Saharan Africa; the next highest diversity is East Asia at 0.47 (Table B.0.2 in the Appendix). Taking into account Africa's high linguistic diversity, with its negative effect on scores, suggests African countries' average "performance" on learning outcomes compared to Latin America

[43]The grandparent average for schooling years in sub-Saharan African is 4.89 in our regression sample, much lower than the average of 8.12 for Latin American countries, as shown in Table B.0.2 in the Appendix.

[44]The average is 404 and the standard deviation is 58, as shown in Table B.0.1 in the Appendix.

[45]The linguistic diversity index reports the probability that any two randomly selected individuals from a country would have different mother tongues. A value of 0 indicates that everyone has the same mother tongue, while a value of 1 indicates that no two people have the same mother tongue.

and the Caribbean is better (in the conditional sense), by 41 points. In magnitude of harmonized test score points, that more than doubles its conditionally better performance, from 16 to 41 points—Columns 2 and 4. Linguistic diversity also robs GDP per capita of its robust positive effect (Columns 4 and 5). Average "diversity" across the full sample is 0.50, so linguistic diversity reduces scores across all countries by about 19 points[46]—again relatively small but highly robust statistically.

[46] 0.50 times the coefficient of 0.38 (Table B.0.2 in the Appendix and Column 5 in Table 4).

Table 4: Harmonized Test Scores Across Regions

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4	(5) Column 5
East Asia	55.731***	51.755***	52.694***	62.290***	63.222***
	(20.578)	(15.111)	(14.592)	(15.868)	(15.527)
MENA	-3.425	-23.645**	-7.005	8.855	9.049
	(13.986)	(10.304)	(11.022)	(13.123)	(13.340)
S. Asia	-39.765***	-8.526	-2.023	14.457	16.030
	(11.163)	(11.360)	(11.339)	(15.428)	(15.328)
SSA	-26.116**	15.712	31.529**	41.111**	32.535**
	(11.339)	(11.452)	(15.414)	(16.467)	(15.154)
Log GDP Per Capita (2011 PPP)		32.448***	18.208**	12.337	12.083
		(5.431)	(7.272)	(8.071)	(8.152)
HCI Expected Years of School			6.246*	4.244	
			(3.438)	(3.654)	
BL Avg Schooling (20-29)					0.085
					(3.844)
BL Avg Schooling (40-49)			3.682	2.433	3.655
			(2.763)	(2.586)	(3.863)
Gov Effectiveness				11.410	11.829
				(13.267)	(14.109)
Language Diversity				-38.076**	-38.448**
				(16.461)	(17.368)
Fragility Index				-0.237	-0.354
				(0.487)	(0.518)
No. of countries	84	84	84	84	84
R-squared	0.29	0.55	0.60	0.65	0.65

The dependent variable is harmonized test scores from the World Bank's HCI. Regression sample drops countries classified as belonging to OECD or Europe and Central Asia, and is restricted to countries that have data on all the covariates included in the regression model. "East Asia," "MENA," "South Asia," and "SSA"‘ are regional dummies that take the value of 0 or 1 depending on whether the country belongs to the region. Latin America and the Caribbean is the omitted category. GDP per capita (constant 2011 international dollars and PPP adjusted) is for most recent available year from the World Bank's World Development Indicators. "HCI Expected Years of School" are calculated using repetition-adjusted enrollment rates by school level to proxy for age-specific enrollment rates up to age 18 and taken from the World Bank's HCI data. "BL Avg Schooling (20-29)" and "BL Avg Schooling (40-49)" are from Barro and Lee 2010, and are weighted by population. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5." Linguistic diversity index is Greenberg's diversity index from Ethnologue (data accessed from website on February 7, 2019). The index reports the probability that any two randomly selected individuals from a country would have different mother tongues. A value of 0 indicates that everyone has the same mother tongue, while a value of 1 indicates that no two people have the same mother tongue. The computation of the diversity index is based on the population of each language as a proportion of the total population. Measure of fragility is taken from Fund for Peace (accessed on March 1, 2019) and is for the year 2018. The fragile states index is based on different factors countries face that impact their level of fragility. Details about the methodology can be found here. Asterisks (∗, ∗∗, ∗∗∗) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

To the extent sub-Saharan Africa is different from Latin America in a way that clearly affects countries' scores, Africa's substantially greater within-country linguistic diversity matters almost as much as other factors, including its lower average GDP per capita, in explaining

its performance. We return to this issue in the next section on differences across countries *within* Africa.

We repeat the same regression exercise using the IRT-equated scores developed by Sandefur (2018). Results (on a smaller sample of countries, and with only Chile—the highest performing country in Latin America and an OECD country—representing Latin America and the Caribbean and having the necessary data to compute the IRT score) are essentially the same. Again, African countries' performance, conditional on GDP per capita and other factors, especially linguistic diversity, is better than Latin America's (Tables D.2.1 and D.2.2 in the Appendix).

4.3.2 Differences in Harmonized Test Scores Across Countries *Within* Africa

It is clear from a cursory glance at Figures 6 and 7 above that there are substantial differences among countries within sub-Saharan Africa in measured harmonized test scores—for example, between Kenya's score of over 450 (higher than Uruguay's and close to that of Chile in Figure 6) and Ghana's of just over 300. The harmonized test score of Burundi, one of the poorest countries in Africa, is greater than that of most other countries in the region, including Ethiopia and much richer Cote d'Ivoire and South Africa, and is similar to that of Colombia and Argentina. Figure 7, in which scores are mapped against median consumption, illustrates the minimal role median consumption has in accounting for these relatively big differences in scores within the region. Even among African countries that have similar levels of linguistic diversity, such as Nigeria and Gabon, there is large variation in performance (Figure 8).

Table 5 shows the result of repeating our basic descriptive regressions across a sample restricted to African countries only. Linguistic diversity again seems to matter for test scores (Column 5). The same exercise using median consumption instead of GDP per capita (Table D.3.1 in the Appendix) is consistent, though language diversity loses statistical significance.

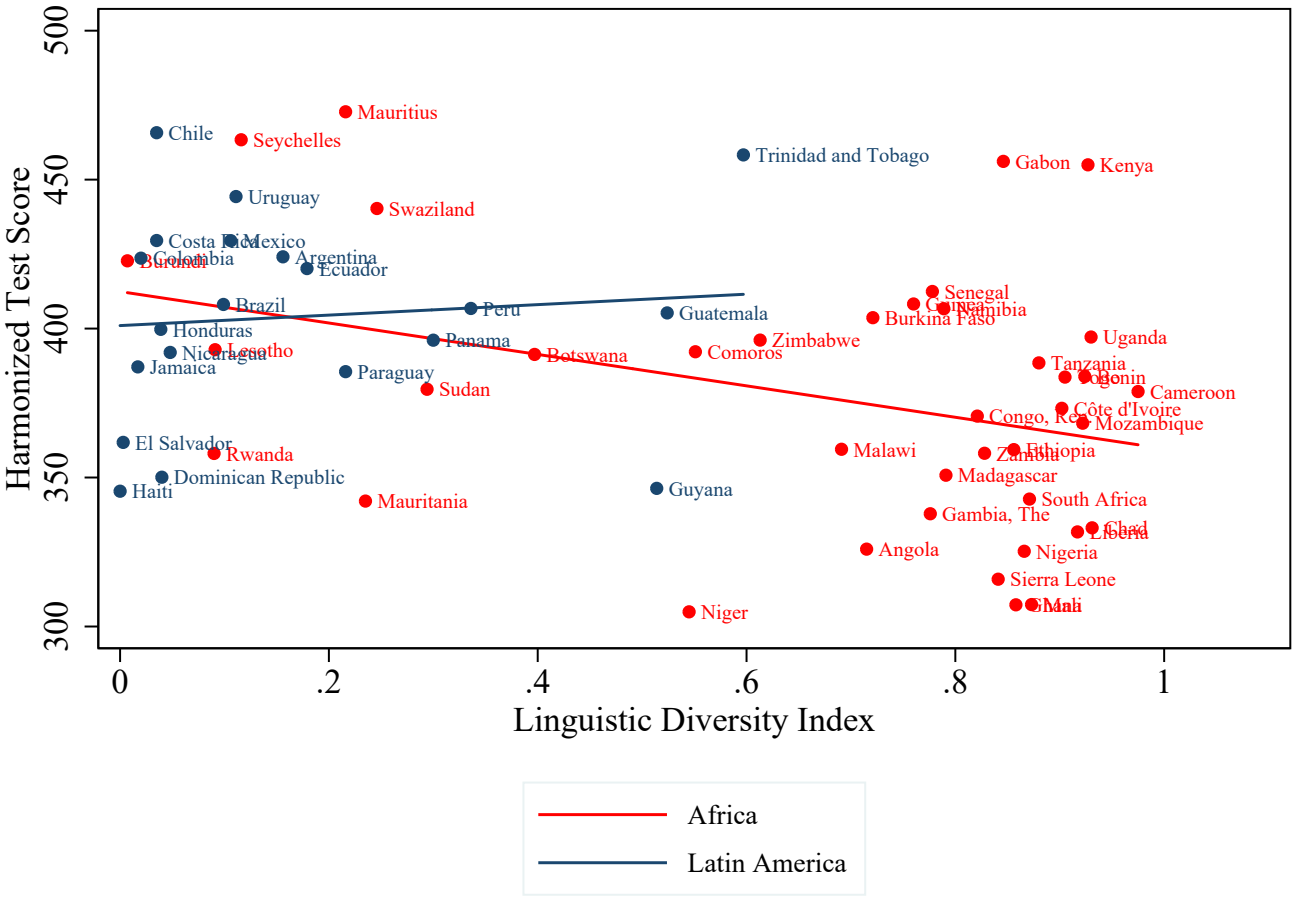
Table 5: Harmonized Test Scores Within sub-Saharan Africa

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4	(5) Column 5
Log GDP Per Capita (2011 PPP)	22.050**	19.231**	14.800	15.330	11.966
	(8.288)	(7.726)	(9.248)	(9.563)	(10.844)
Language Diversity		-40.096*	-37.802	-38.409	-38.702*
		(23.050)	(23.013)	(23.575)	(21.679)
Fragility Index			-0.461	-0.561	-0.166
			(0.557)	(0.805)	(0.844)
Gov Effectiveness				-4.042	-5.045
				(19.765)	(18.076)
HCI Expected Years of School					6.698
					(5.119)
No. of countries	39	39	39	39	39
R-squared	0.20	0.26	0.28	0.28	0.33

The dependent variable is harmonized test scores from the World Bank's HCI. Regression analysis runs on a restricted sample of African countries only. GDP per capita (constant 2011 international dollars and PPP adjusted) is for most recent available year from the World Bank's World Development Indicators. We take natural log of GDP figure. Linguistic diversity index is Greenberg's diversity index from Ethnologue (data accessed from website on February 7, 2019). The index reports the probability that any two randomly selected individuals from a country would have different mother tongues. A value of 0 indicates that everyone has the same mother tongue, while a value of 1 indicates that no two people have the same mother tongue. The computation of the diversity index is based on the population of each language as a proportion of the total population. Measure of fragility is taken from Fund for Peace (accessed on March 1, 2019) and is for the year 2018. The fragile states index is based on different factors countries face that impact their level of fragility. Details about the methodology can be found here. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5." "HCI Expected Years of School" are calculated using repetition-adjusted enrollment rates by school level to proxy for age-specific enrollment rates up to age 18 and taken from World Bank's HCI data. Asterisks (∗, ∗∗, ∗∗∗) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

Figure 8 suggests that it is the negative relationship between scores and linguistic diversity in sub-Saharan Africa that possibly drove the results in Table 4 above (probably for South Asia as well), compared with Latin America, where most countries have relatively low language diversity.

Figure 8: Harmonized Test Scores and Linguistic Diversity: sub-Saharan Africa vs. Latin America



Source: Authors' calculations based on harmonized test score data from the World Bank's HCI. Linguistic diversity index is Greenberg's diversity index from Ethnologue (data accessed from website on February 7, 2019). The index reports the probability that any two randomly selected individuals from a country would have different mother tongues. A value of 0 indicates that everyone has the same mother tongue, while a value of 1 indicates that no two people have the same mother tongue. The computation of the diversity index is based on the population of each language as a proportion of the total population.

The results in sub-Saharan Africa (of a negative relationship of language diversity and test scores) are consistent with studies suggesting that early instruction in the child's mother tongue is beneficial for learning. Bashir et al. (2018) note that students in countries where the language of both testing and instruction is the national language—for example, Kirundi in Burundi and Kiswahili in Tanzania—outperform students in countries where the language of

the test is an international language, such as English. Other evidence from Piper et al. (2016) and Seid (2016) indicates that mother tongue instruction can effectively improve literacy as well as enrollment. However, implementing mother tongue instruction policies is complicated in contexts with a variety of languages, as it is costly to produce adequate learning materials and there may be a dearth of local language teachers. Nigeria, for example, has more than 40 languages.

By the same logic, language diversity has implications for standardized learning measurement, both nationally and cross-nationally, particularly in the early grades, with its substantially higher costs of preparing and managing testing. To the extent that good data on learning can stimulate system reforms that improve learning, then language diversity is a further handicap, and is a far greater one in sub-Saharan Africa than in Latin America.

Other handicaps include low government effectiveness and fragility, which complicate designing and administering tests of learning (as well as managing school systems). Africa performs poorly compared to other regions on both these measures (Table B.0.2 in the Appendix).[47] Consistent with the regression results across regions in Table 4, fragility within Africa has the expected negative signs (Table 5 and Table D.3.1 in the Appendix), though it is not statistically significant in this small sample. At similar levels of fragility and government effectiveness, there is considerable heterogeneity in learning performance among African countries (Figures C.0.1 and C.0.2 in the Appendix).

Tying up our findings *across* regions and *within* sub-Saharan Africa, some of the unconditional difference between the higher average of country test scores in Latin America and Caribbean compared to sub-Saharan Africa is accounted for by Africa's lower average income and greater linguistic diversity. Indeed, taking those into account, Africa's "performance" reflected in the harmonized test scores, is no worse than Latin America's, in the conditional sense. Its lower expected years of schooling, lower government effectiveness, and greater fragility also probably matter, but their robustness is reduced depending on the inclusion

[47]More African countries tend to be fragile—sub-Saharan Africa is home to 19 of the 36 countries the World Bank describes as "fragile or conflict affected," according to the World Bank, IFC.

of other variables that are probably highly correlated with them and with GDP per capita. To the extent they do matter independently, they are likely to hamper the development of cross-national assessments by creating implementation hurdles and raising the costs of managing testing.

We emphasize in particular language diversity as a challenge for Africa in managing regionally benchmarked testing; high linguistic diversity (also a challenge in some countries of South Asia), raises the implementation costs of developing and implementing tests of learning, and may have discouraged even standardized national assessments. In order to participate in the international fourth grade reading test, PIRLS, South Africa translated the test into 11 different local languages—something no OECD country had to do. Language diversity also matters in managing learning itself, particularly in the early grades; the example of Burundi suggests that there are benefits to young children in establishing a base of learning in their own language, but that conclusion is contingent on many factors specific to each country, and abstracts from questions regarding at what grade or age to introduce a common language, and at what age or grade to introduce a European language in countries that unlike Burundi (and Tanzania) do not have a widely spoken national language.

Other factors at the regional level also distinguish sub-Saharan Africa from Latin America (though not the Caribbean sub-region). For the many relatively small economies in sub-Saharan Africa, the fixed costs of testing are high, and the absolute large number of countries (46) makes it more difficult to cooperate on any region-wide initiative. The region's cultural and political diversity—an outcome of its longer and more difficult colonial history—may also slow positive program "contagion" of the benefits of testing and of education policy change across countries. All these conditions make cross-national testing more technically challenging and expensive than in Latin America and have contributed to differences in the politics of testing and its evolution in the two regions.

A bottom line from a practical point of view is the following: If sub-Saharan Africa is to benefit from regionally and globally benchmarked testing, as have at least some countries

in Latin America, as input to policies aimed at raising learning, two constraints have to be addressed. One, expanding the reach and impact of cross-national testing in Africa will require higher commitment from African governments (even more than was necessary in Latin America), and is most likely to come as it did in Latin America, in the form of a purposeful regional initiative with leadership from one or more of Africa's regional institutions. Two, it will require greater willingness of international donors to collaborate with governments and regional programs in support of benchmarked testing.

5 Conclusions

"Assess learning to make it a serious goal:" measuring learning is the first step in raising education quality according to the World Bank's WDR (2018). This paper finds substantial support for that recommendation from the experience of the Latin American region over the past 20 years. Latin America's experience suggests that the decision of a national government to measure and report the results of tests of children's learning can contribute to better programs and policy, and more learning. The trajectory Latin America followed in doing measurement and using results has lessons for sub-Saharan Africa—most importantly, that region-wide agreement on a common test contributes to the development of capacity and processes at the national level and eases the politics of education reforms within countries.

What has been called "a culture of testing" is broadly diffused across the Latin American region, with virtually all countries now measuring learning nationally and cross-nationally at regular intervals through technically robust programs of assessment. It is harder to compare or quantify the intensity of education reform activity, but a striking number of countries have implemented large-scale reforms aimed at raising learning (Chile, Ecuador, Peru, Colombia, Mexico, Brazil), and in many cases have seen important gains. In a sense, even those cases where reforms proved politically unsustainable attest to how ambitious Latin America's efforts to improve education quality have been. It is difficult to imagine politicians assuming

the risks of major reforms without public awareness of, and dissatisfaction with, student learning levels.

Latin America's "culture of testing" developed around the experience of a critical mass of countries working together to build and apply a regionally benchmarked test. Indeed, an important contribution of a regional test is that it can be implemented in countries that do not yet have national assessment capacity, as was the case for many countries when they joined the first Latin American regional assessment in the late 1990s. A regional effort lowers the costs of initial capacity development, as it allows countries with similar levels of educational development to leverage technical expertise that is in scarce supply globally.

In two-thirds of the countries in sub-Saharan Africa today, there has been no cross-nationally comparable measurement of learning since 2014. Our analysis suggests that the issues may be less on the demand side than the supply side. Two major cross-national testing initiatives in Africa were launched with countries' demand for better, and comparable, measures of learning in the early 1990s—even before Latin America began to establish its regional test. Over three rounds of implementation, SACMEQ appears to have had real impact on education policy by presenting countries with transparent evidence that some of their neighbors were performing better, similar to the impact of the ERCE regional test in LAC. Our analysis (using the World Bank's harmonized test score proxy for formal test scores) shows that Africa's learning performance is no worse than the performance of countries in Latin America and the Caribbean, once African countries' greater poverty and much greater diversity of languages are taken into account.

The major difference in sustained support for regional testing seems to be on the supply side, driven by differences in the regions *qua* regions—small country size, internal language diversity, fiscal constraints—which raise the relative and absolute costs of regional cooperation in designing and administering a benchmarked test in Africa. In Latin America, the "horizontal" effort among Spanish-speaking countries, plus Brazil, was confined to a single language bloc and subsidized by the Spanish government and the Inter-American Develop-

ment Bank. It never extended to the 20 English-speaking countries of the Caribbean or Haiti. In Africa, the existence of two large language blocs, plus countries that have not aligned easily with either, has led to two different regional testing efforts and many countries outside of them.

International support is critically important for overcoming these supply-side constraints to regionally benchmarked assessment in sub-Saharan Africa. Equally important is institutional coordination to lower the costs of cross-national measurement. All the regional and international testing agencies could gain from agreements to harmonize the grades tested, for example, aligning around a paper-based test at third grade, as well as a sixth grade test, and contributing to and drawing from the common global bank of test items being developed by UNESCO's Institute for Statistics.

Finally, the value of a regional assessment—for participating countries and education policy—goes up with the number and diversity of participating countries. International assessments, such as PISA, TIMSS, and PIRLS have had major influence on global education policy because of the window they provide into the learning performance of education systems with very different institutional features and policies. Fifty-two countries in Africa have now unified behind the African Continental Free Trade Agreement (AfCFTA), designed to promote regional integration. A parallel effort to unite the region in a single, Africa-wide learning assessment that builds upon the instruments, capacity, and experience of PASEC and SACMEQ, and extends participation to all countries in the region would powerfully advance the same goal.

Four years after the SDG targets were adopted, the world still has no baseline measure of learning levels at third grade, sixth grade, and the end of lower secondary for about half of the countries in the world and for most of the poorest countries in the world—the countries whose children will benefit most from schools that deliver learning. Even a goal of establishing such a baseline by 2021 seems audacious, given the slow progress of the past four years. Latin America's experience provides a clear example of how coordinated

and efficient application of resources can produce significant strides in building countries' capacity, strengthening their focus on learning, supporting better research, and stimulating countries to adopt and diffuse reforms that raise learning. Hopefully, it will inspire other regions too.

6 Bibliography

Addey, C. (2015). *Participating in International Literacy Assessments in Lao PDR and Mongolia: A Global Ritual of Belonging*. Cambridge University Press.

Addey, C. (2017). Golden Relics & Historical Standards: How the OECD is Expanding Global Education Governance through PISA for Development. *Critical Studies in Education*, 58:311–325.

Addey, C. and Sellar, S. (2018). *Why do Countries Participate in PISA? Understanding the Role of International Large-scale Assessments in Global Education Policy*. London, Bloomsbury.

Addey, C. and Sellar, S. (2019). Is It Worth It? Rationales for (Non) Participation in International Large-Scale Learning Assessments. *UNESCO Working Paper Series*.

Akmal, M. and Pritchett, L. (2019). Learning Equity Requires More than Equality: Learning Goals and Achievement Gaps between the Rich and the Poor in Five Developing Countries. *Center for Global Development Working Paper Series*.

Bashir, S., Lockheed, M., Ninan, E., and Tan, J. (2018). Facing Forward: Schooling for Learning in Africa. Technical report, Washington, DC: World Bank.

Beatty, A., Berkhout, E., Bima, L., Coen, T., Pradhan, M., and Suryadarma, D. (2018). Indonesia Got Schooled: Fifteen Years of Rising Enrolment and Flat Learning Profiles. *RISE Working Paper Series*.

Birdsall, N., Bruns, D., and Keller, J. (2016). Learning Data for Better Policy: A Global Agenda. *Center for Global Development Working Paper Series*.

Birdsall, N. and Meyer, C. (2014). The Median Is the Message: A Good Enough Measure of Material Well-Being and Shared Development Progress. *Center for Global Development Working Paper Series*.

Breakspear, S. (2014). How Does PISA Shape Education Policy Making? How Do We Measure Learning Determines What Counts in Education. *Seminar Series Paper No. 240. Centre for Strategic Education. East Melbourne, Australia.*

Bruns, B., Evans, D., and Luque, J. (2012). *Achieving World-Class Education in Brazil.* Washington, DC: World Bank.

Bruns, B. and Luque, J. (2015). *Great Teachers: How to Raise Student Learning in Latin America and the Caribbean.* Washington, DC: World Bank.

Carnoy, M., Gove, A., and Marshall, J. (2007). *Cuba's Academic Advantage: Why Students in Cuba Do Better in School.* Stanford University Press. Palo Alto, CA.

Crouch, L. and Gustafsson, M. (2018). Worldwide Inequality and Poverty in Cognitive Results: Cross-sectional Evidence and Time-based Trends. *RISE Working Paper Series.*

Education Commission (2017). The Learning Generation: Investing in Education for Changing World. Technical report, Education Commission.

Ertl, H. (2014). How Does PISA Shape Education Policy Making? How Do We Measure Learning Determines What Counts in Education. *Oxford Review of Education*, 32(5):619–634.

Estarellas, C. and Bramwell, D. (2015). *Ecuador, 2007-14: Attempting a Radical Educational Transformation.* Simon Schwartzman, London.

Ferrer, G. (2006). *Educational Assessment Systems in Latin America: Current Practice and Future Challenges.* PREAL.

Ferrer, G. and Fiszbein, A. (2015). What Has Happened with Learning Assessment Systems in Latin America? Lessons From the Last Decade of Experience. *Washington, DC: World Bank.*

Filmer, D., Rogers, H., Angrist, N., and Sabarwal, S. (2018). Learning-Adjusted Years of Schooling (LAYS): Defining A New Macro Measure of Education. *Washington, DC: World Bank*.

Fischman, G., Topper, A., Silova, I., Joebel, J., and Holloway, J. (2018). Examining the Influence of International Large-Scale Assessments on National Education Policies. *Journal of Education Policy*, 34:470–499.

Gustafsson, M. (2019). Pursing equity through policy in the schooling sector 2007-2017. In Spaull, N. and Jansen, J., editors, *South African Schooling: The Enigma of Inequality*. Springer.

Leste, A. (2005). Streaming in Seychelles: From SACMEQ Research to Policy Reform. *Conference Paper, International Invitational Educational Policy Research Conference, Paris, France*.

Lilenstein, A. (2018). Integrating Indicators of Education Quantity and Quality in Six Francophone African Countries. *Working Papers 09/2018, Stellenbosch University, Department of Economics*.

Makuwa, D. and Maarse, J. (2013). The Impact of Large-Scale International Assessments: A Case Study of How the Ministry of Education in Namibia Used SACMEQ Assessments to Improve Learning Outcomes. *Research in Comparative and International Education*, 8:349–358.

McKinsey and Co. (2007). How The World's Best-Performing School Systems Come Out On Top. Technical report, McKinsey and Co.

McMeekin, R. (2003). Networks of Schools. *Washington, DC: Inter-American Development Bank*.

Murimba, S. (2005). The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ): Mission, Approach and Projects. *Prospects*.

OECD (2016). PISA 2015 Results (Volume I): Excellence and Equity in Education. Technical report, OECD Publishing, Paris.

Patrinos, H. and Angrist, N. (2018). Global Dataset on Education Quality: A Review and Update (2000-2017). *Washington, DC: World Bank*.

Piper, B., Destefano, J., Kinyanjui, E., and Ong'ele, S. (2018). Scaling Up Successfully: Lessons from Kenya's Tusome National Literacy Program. *Journal of Educational Change*, 19:293–321.

Piper, B., Zuilkowski, S., and Salome, O. (2016). Implementing Mother Tongue Instruction in the Real World: Results from a Medium-Scale Randomized Controlled Trial in Kenya. *Comparative Education Review*, 60:776–807.

Pritchett, L. (2013). *The Rebirth of Education*. Center for Global Development.

Ravela, P., Arregui, P., and Valverde, G. (2008). The Assessments that Latin America Needs. *Working paper No. 40. PREAL. Inter-American Dialogue. Washington DC*.

Sandefur, J. (2018). Internationally Comparable Mathematics Scores for Fourteen African Countries. *Economics of Education Review*, pages 267–286.

Schneider, B., Estarellas, P., and Bruns, B. (2017). The Politics of Transforming Education in Ecuador: Confrontation and Continuity, 2006-17. *RISE Working Paper Series*.

Seid, Y. (2016). Does Learning in Mother Tongue Matter? Evidence from a Natural Experiment in Ethiopia. *Economics of Education Review*, 55:21–38.

Solano-Flores, G. and Bonk, W. (2003). Evaluation of the Latin American Laboratory for the Evaluation of Educational Quality (LLECE). *UNESCO Internal Oversight Service*.

Solano-Flores, G. and Bonk, W. (2008). Evaluation of the Latin American Laboratory for the Evaluation of Educational Quality (LLECE). Technical report, UNESCO.

Spaull, N. (2016). Serious Technical Concerns about SACMEQ IV Results Presented to Parliament.

Trevino, E. and Ordenes, M. (2017). Exploring Commonalities and Differences in Regional and International Assessments. *UNESCO Information Paper No. 48*.

UNESCO (2018). SDG 4 Data Digest 2018: Data to Nurture Learning. Technical report.

Wagner, D. (2011). Smaller, Quicker, Cheaper: Improving Learning Assessments for Developing Countries. Technical report, UNICEF.

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

Appendix

A Existing Evidence

Figure A.0.1: Summary of Existing Evidence from Regional Assessments in sub-Saharan Africa

Subject	Test	Country	Finding
Mathematics	SACMEQ 6	Botswana, Eswatini, Kenya, Lesotho, Malawi, Mauritius, Namibia, Seychelles, South Africa, Zambia, Zimbabwe, Uganda	In most countries, less than 50% of children reached minimum performance
Mathematics	PASEC 6	Burundi, Benin, Burkina Faso, Cameroon, Chad, Congo, Rep., Cote d'Ivoire, Mali, Niger Senegal, Togo	In most countries, less than 50% of children reached minimum performance
Mathematics	SDI 4	Kenya, Tanzania, Togo, Mozambique, Nigeria, Senegal, Uganda	In all countries, less than 50% of children reached minimum performance
Mathematics	PASEC 2	Benin, Burundi, Burkina Faso, Cameroon, Chad, Congo, Rep., Cote d'Ivoire, Niger, Mali, Senegal, Togo	In most countries, less than 75% of children reached minimum performance
Mathematics	Uwezo	Kenya, Tanzania, Uganda	Less than 40% of children know how to do simple math by Grade 3 (Akmal & Pritchett, 2018)
Reading	SACMEQ 6	Botswana, Eswatini, Kenya, Lesotho, Malawi, Mauritius, Namibia, Seychelles, South Africa, Uganda, Zambia, Zimbabwe	In most countries, greater than 75% of children reached minimum proficiency
Reading	PASEC 6	Benin, Burkina Faso, Cameroon, Chad, Congo, Rep., Niger, Togo	In all countries, less than 75% of children reached minimum proficiency
Reading	SDI 4	Kenya, Tanzania, Togo, Mozambique, Nigeria, Senegal, Uganda	In most countries, less than 50% of children reached minimum proficiency
Reading	PASEC 2	Benin, Burkina Faso, Burundi, Cameroon, Chad, Congo, Rep., Cote d'Ivoire, Mali, Niger, Senegal, Togo	In most countries, less than 50% of the children reached minimum proficiency
Reading	Uwezo	Kenya, Tanzania, Uganda	Less than 50% of children know how to read by Grade 3 (Akmal & Pritchett, 2018)

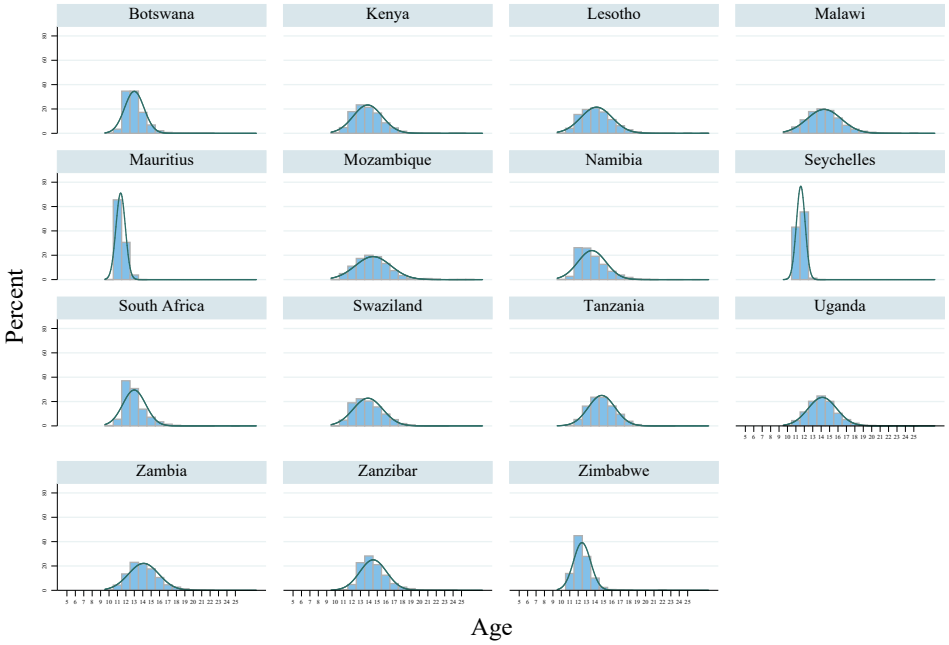
Source: Figures are from Bashir et al. (2018).

Figure A.0.2: Summary of Existing Evidence from International Assessments in sub-Saharan Africa and Latin America

Subject	Test	Country	Finding	Comparison to LAC
Science	PISA+ (2009)	Mauritius	28.9% of students reached Level 2 (baseline level of proficiency)	37.9% reached Level 2 in Costa Rica and 29.4% in Miranda-Venezuela
Science	TIMSS (2015) Grade 8	Botswana, South Africa	51% and 32% of children reached the low international benchmark in Botswana and South Africa respectively	75% of children reached the low international benchmark in Chile
Mathematics	PISA+ (2009)	Mauritius	24.9% of students reached Level 2 (baseline level of proficiency)	27.8% reached Level 2 in Costa Rica and 23.6% in Miranda-Venezuela
Mathematics	TIMSS (2015) Grade 8	Botswana, South Africa	47% and 34% of children reached the low international benchmark in Botswana and South Africa respectively	63% of children reached the low international benchmark in Chile
Reading	STEP Adult (2016)	Kenya, Ghana	More than 50% of adults with upper secondary were at Level 1 or below	More than 50% in Bolivia and more than 20% in Colombia with upper secondary were at Level 1 or below
Reading	PISA+ (2009)	Mauritius	29.9% reached Level 2 (baseline level of proficiency)	34.7% reached Level 2 in Costa Rica and 27.8% in Miranda-Venezuela

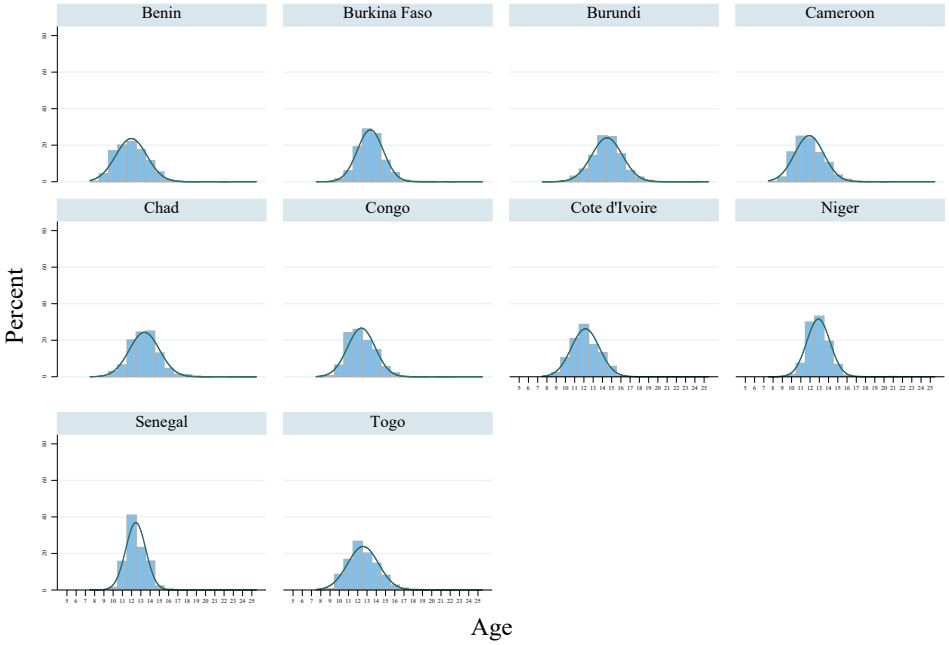
Figure A.0.3: Age Distribution of Test Takers in sub-Saharan Africa by Country

(a) SACMEQ (Grade 6)



Source: Authors' calculations based on data from SACMEQ II (2000) and III (2006).

(b) PASEC (Grade 6)



Source: Authors' calculations based on data from PASEC 2014.

B Summary Statistics

Table B.0.1: Summary Statistics for Key Variables (HCI)

Row name	Mean	SD	Max	Min	No.
Harmonized Test Score	403.76	58.37	580.87	304.92	87
LAC	.23	.42	1	0	87
SSA	.34	.48	1	0	87
East Asia & Pacific	.2	.4	1	0	87
MENA	.16	.37	1	0	87
South Asia	.07	.25	1	0	87
Log GDP Per Capita	8.96	1.18	11.67	6.51	87
Median Consumption/Income	197.86	183.77	1045.03	38.01	76
Expected Years of School	10.49	2.28	13.89	4.41	87
BL Avg Schooling (20-29)	8.31	2.79	14.34	1.99	87
BL Avg Schooling (40-49)	6.65	2.86	13.12	1.19	87
Gov Effectiveness	-.24	.79	2.21	-2.06	87
Language Diversity	.5	.33	.99	0	86
Fragility	75.63	19.01	112.67	30.42	84

Table B.0.2: Regional Means for Key Variables (HCI)

Row name	LAC	SSA	East Asia & Pacific	MENA	South Asia
Harmonized Test Score	403.98	377.86	465.83	400.55	364.22
Log GDP Per Capita	9.32	8.03	9.6	9.94	8.36
Median Consumption/Income	291.11	86.71	338.27	206.41	113.06
Expected Years of School	11.86	8.29	12.07	11.32	10.54
BL Avg Schooling (20-29)	9.54	6.34	10.08	8.95	7.57
BL Avg Schooling (40-49)	8.12	4.89	8.45	6.93	4.86
Gov Effectiveness	-.2	-.61	.39	-.09	-.6
Language Diversity	.17	.67	.47	.55	.65
Fragility	65.53	85.25	67.02	71.72	90.39

Table B.0.3: Summary Statistics for Key Variables (IRT Equated)

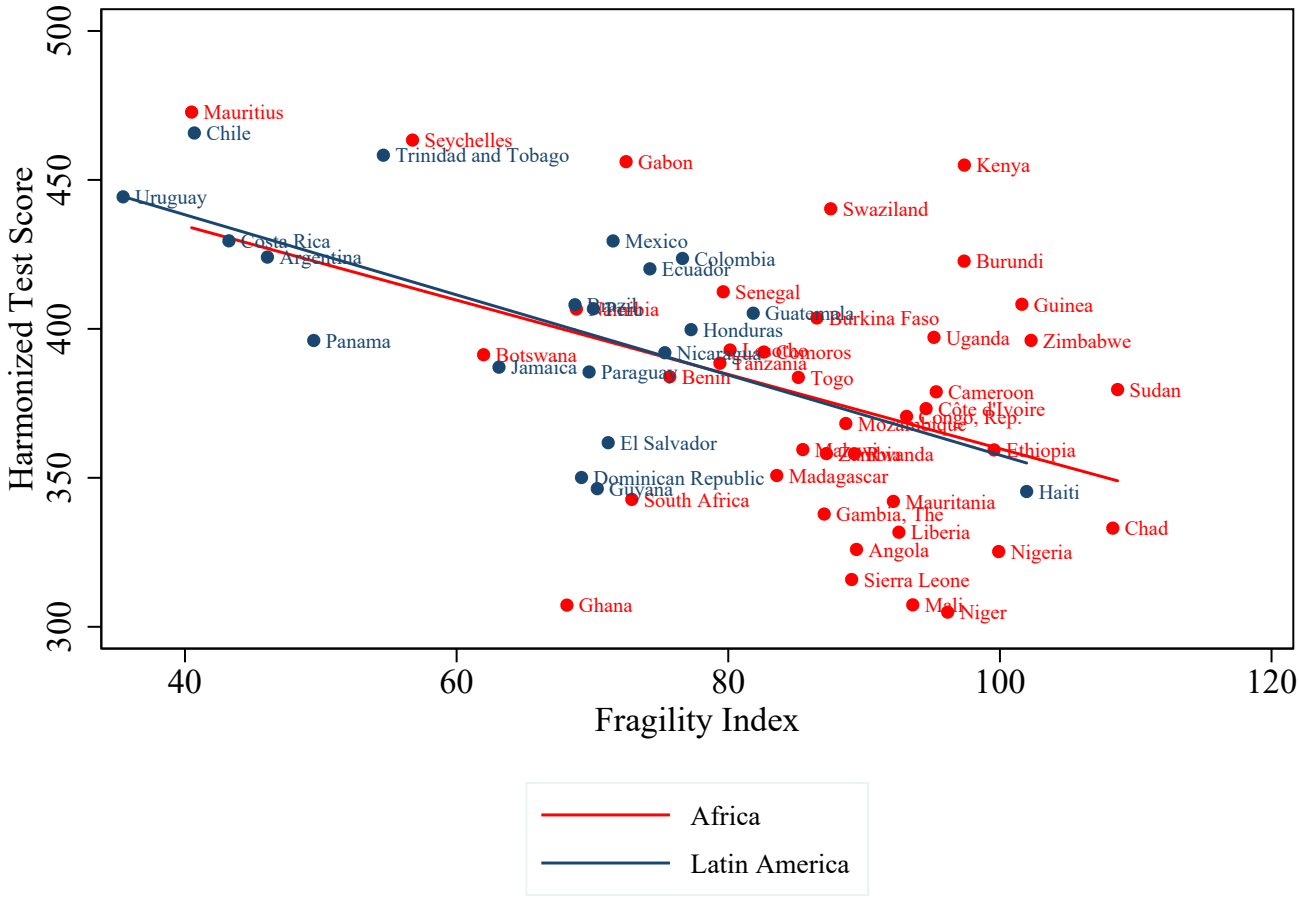
Row name	Mean	SD	Max	Min	No.
Equated Test Score	395.66	95.28	605.45	263.61	29
LAC	.03	.19	1	0	29
SSA	.48	.51	1	0	29
East Asia & Pacific	.24	.44	1	0	29
MENA	.24	.44	1	0	29
Log GDP Per Capita	9.18	1.2	11.36	7	29
Median Consumption/Income	229.2	268.43	1045.03	38.01	25
BL Avg Schooling (20-29)	9.38	2.83	14.34	1.99	29
BL Avg Schooling (40-49)	7.56	2.84	13.12	1.19	29
Gov Effectiveness	.12	.84	2.21	-1.19	29
Language Diversity	.56	.31	.93	.01	29
Fragility	71.74	19.69	102.29	30.42	28

Table B.0.4: Regional Means for Key Variables (IRT Equated)

Row name	LAC	SSA	East Asia & Pacific	MENA
Equated Test Score	386.88	333.46	521.03	395.96
Log GDP Per Capita	10.03	8.33	10.26	9.69
Median Consumption/Income	424.68	95.72	555.29	237.75
BL Avg Schooling (20-29)	11.49	7.69	12.42	9.44
BL Avg Schooling (40-49)	10.39	6.15	10.79	6.76
Gov Effectiveness	.85	-.33	1.09	-.07
Language Diversity	.04	.66	.49	.49
Fragility	40.69	79.68	53.67	75.78

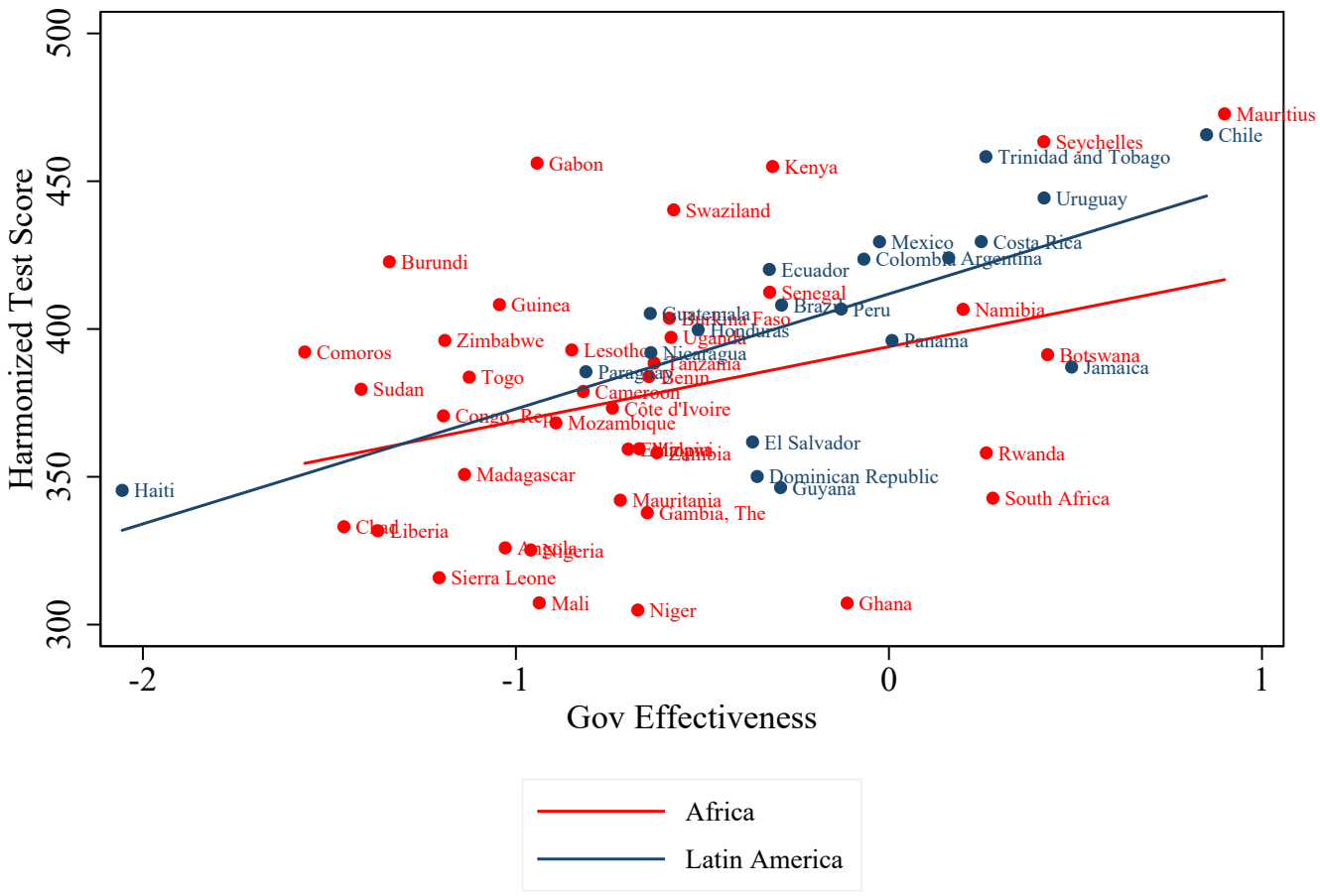
C Correlation Plots

Figure C.0.1: Harmonized Test Scores and Fragile States Index: sub-Saharan Africa vs. Latin America



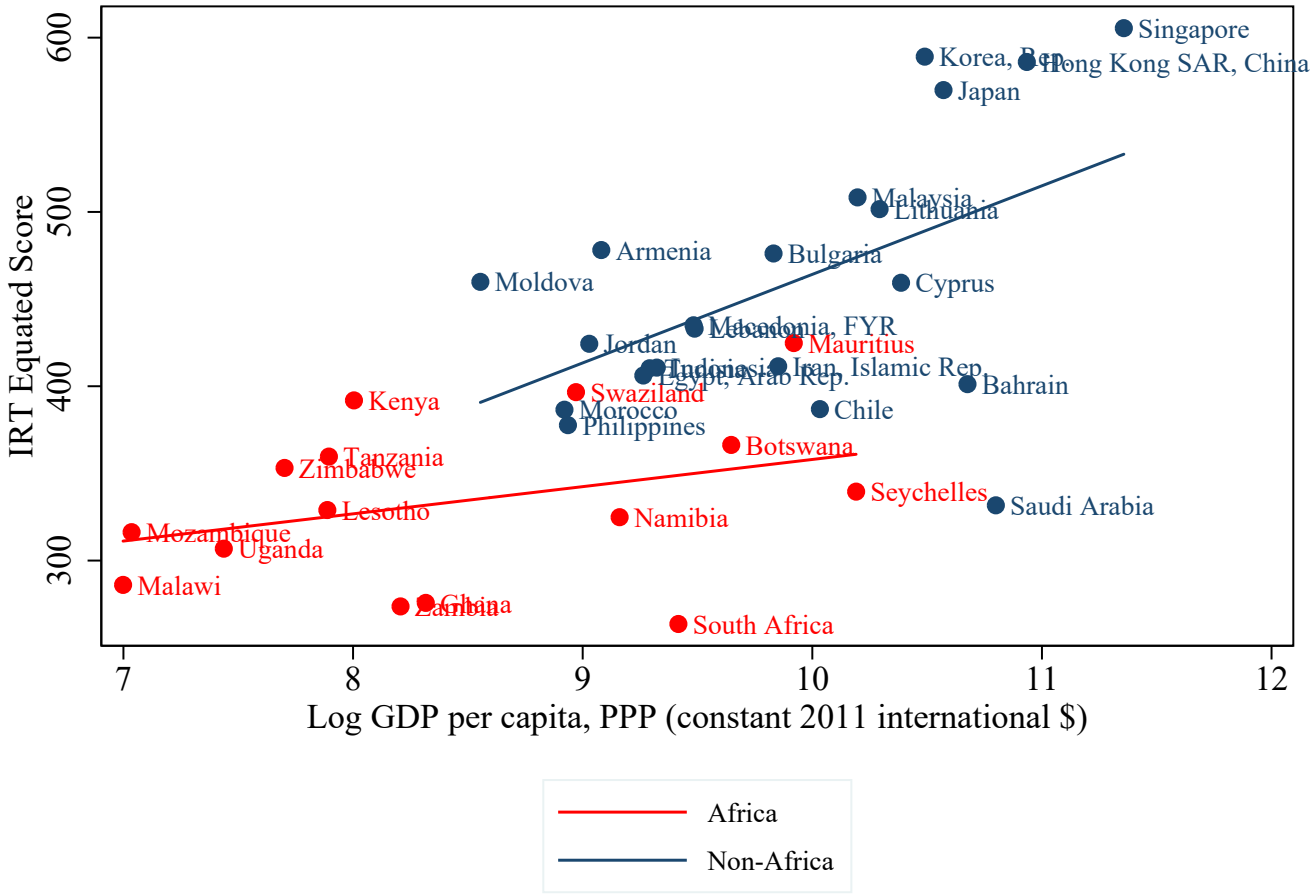
Source: Authors' calculations based on harmonized test score data from the World Bank's HCI. Measure of fragility is taken from Fund for Peace (accessed on March 1, 2019) and is for the year 2018. The fragile states index is based on different factors countries face that impact their level of fragility. Details about the methodology can be found here.

Figure C.0.2: Harmonized Test Scores and Government Effectiveness: sub-Saharan Africa vs. Latin America



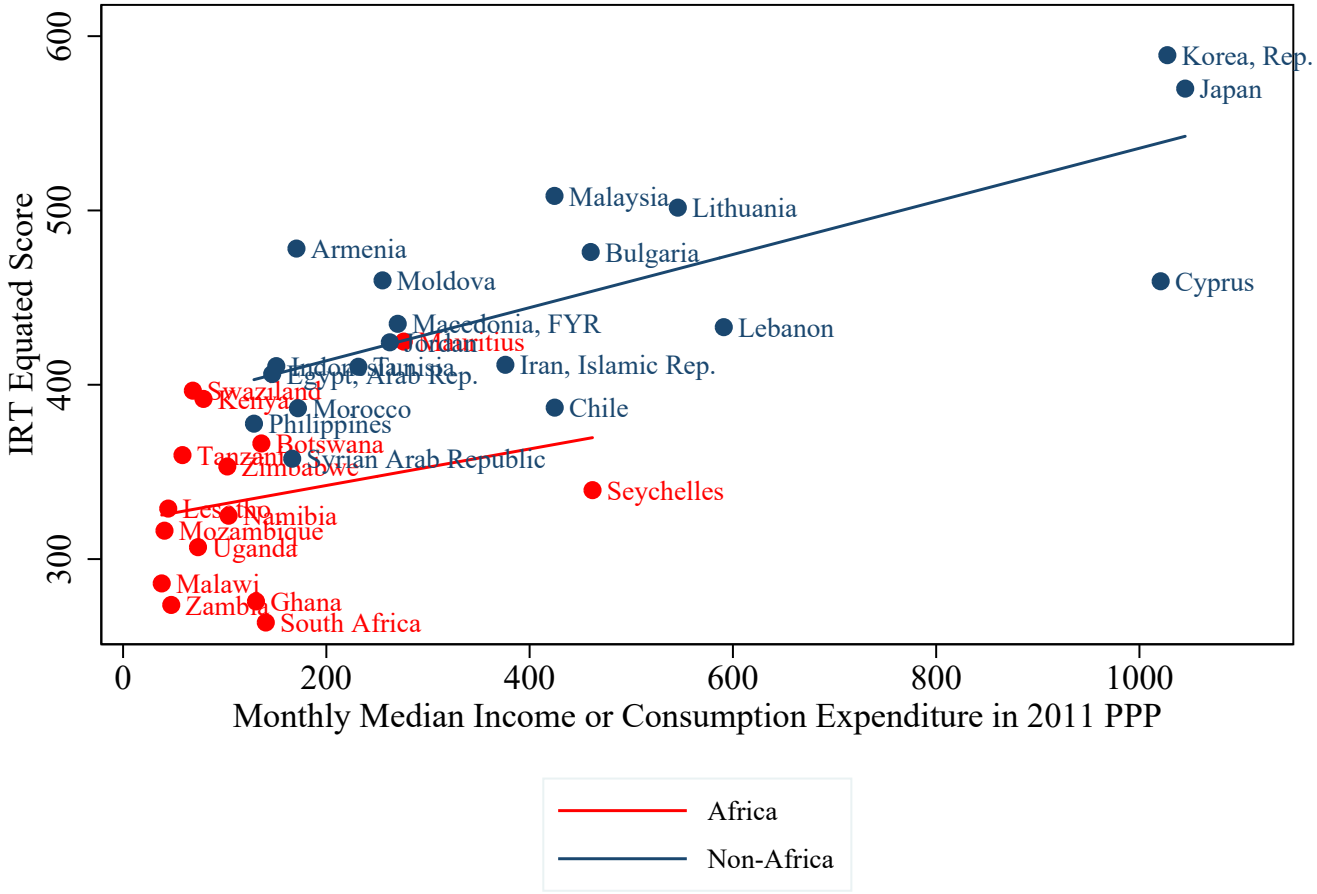
Source: Authors' calculations based on on harmonized test score data from the World Bank's HCI. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5."

Figure C.0.3: IRT-Equated Scores and GDP per capita: sub-Saharan Africa vs. others



Source: Authors' calculations based on learning data from IRT-equated scores from Sandefur (2018) for SACMEQ III and TIMSS 2003 (Grades 7 and 8 scale), and GDP per capita data from the World Bank's World Development Indicators. GDP per capita (constant 2011 international dollars and PPP adjusted) is for most recent available year. We take natural log of the GDP figure. We drop all OECD countries from our sample.

Figure C.0.4: IRT-Equated Scores and Median Consumption/Income: sub-Saharan Africa vs. others



Source: Authors' calculations based on learning data from IRT-equated scores from Sandefur (2018) for SACMEQ III and TIMSS 2003 (Grades 7 and 8 scale). Monthly median consumption/income data are for available year in World Bank's PovcalNet in terms of 2011 PPP. We use consumption data when both income and consumption data are available for a country, and we use figures for "urban" when data is only available dis-aggregated by "urban" and "rural." We drop all OECD countries from our sample.

D Regressions

D.1 Regressions Using Harmonized Test Scores *Across* Regions

Table D.1.1: Harmonized Test Scores Across Region

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4	(5) Column 5
East Asia	40.569*	31.851**	32.613***	30.845**	30.461**
	(20.414)	(12.523)	(11.898)	(12.103)	(12.950)
MENA	-28.006**	-12.345	-0.063	-0.230	-8.103
	(13.956)	(9.596)	(9.858)	(9.772)	(9.206)
S. Asia	-37.874***	-4.955	-1.004	-3.176	-4.695
	(12.202)	(11.288)	(8.562)	(9.271)	(11.240)
SSA	-26.116**	11.675	34.954**	32.183*	16.484
	(11.377)	(11.564)	(16.796)	(16.995)	(12.551)
Median Income/Consumption		0.185***	0.142***	0.132***	0.130***
		(0.018)	(0.023)	(0.025)	(0.028)
HCI Expected Years of School			7.534*	7.044*	
			(3.835)	(3.843)	
BL Avg Schooling (20-29)					2.200
					(4.441)
BL Avg Schooling (40-49)			1.592	1.017	1.532
			(2.665)	(2.835)	(3.938)
Gov Effectiveness				7.127	9.560
				(7.187)	(8.772)
No. of countries	76	76	76	76	76
R-squared	0.25	0.55	0.61	0.61	0.59

The dependent variable is harmonized test score data from the World Bank's HCI. Regression sample is restricted to countries that have data on all the covariates included in the regression model. "East Asia," "MENA," "South Asia," and "SSA" are regional dummies that take the value of 0 or 1 depending on whether the country belongs to the region. Latin America and the Caribbean is the omitted category. Monthly median consumption/income data are for available year in World Bank's PovcalNet in terms of 2011 PPP. We use consumption data when both income and consumption data are available for a country, and we use figures for "urban" when data is only available dis-aggregated by "urban" and "rural." "HCI Expected Years of School" are calculated using repetition-adjusted enrollment rates by school level to proxy for age-specific enrollment rates up to age 18 and taken from World Bank's HCI data. "BL Avg Schooling (20-29)" and "BL Avg Schooling (40-49)" are from Barro and Lee 2010, and are weighted by population. Government effectiveness is the 2017 figure from the World Bank's Governance Indicators. Asterisks (∗, ∗∗, ∗∗∗) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

Table D.1.2: Harmonized Test Scores Across Region

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4	(5) Column 5
East Asia	46.287**	36.084***	37.529***	46.583**	49.908***
	(20.997)	(12.517)	(11.803)	(17.862)	(17.833)
MENA	-28.006**	-12.633	-0.260	7.463	4.247
	(13.963)	(9.670)	(9.718)	(12.556)	(12.731)
S. Asia	-37.874***	-5.559	-0.405	6.987	8.934
	(12.208)	(11.357)	(8.702)	(13.787)	(14.283)
SSA	-26.116**	10.980	32.703*	39.728**	29.216**
	(11.382)	(11.641)	(16.866)	(16.891)	(14.472)
Median Income/Consumption		0.181***	0.132***	0.105***	0.096***
		(0.019)	(0.023)	(0.031)	(0.032)
HCI Expected Years of School			6.405*	5.555	
			(3.801)	(3.936)	
BL Avg Schooling (20-29)					1.465
					(4.379)
BL Avg Schooling (40-49)			2.787	3.010	3.775
			(2.651)	(2.842)	(3.693)
Gov Effectiveness				6.955	6.070
				(12.471)	(13.896)
Language Diversity				-28.211	-30.126
				(18.064)	(18.614)
Fragility Index				0.107	-0.066
				(0.541)	(0.568)
No. of countries	75	75	75	75	75
R-squared	0.28	0.56	0.62	0.64	0.63

The dependent variable is harmonized test score data from the World Bank's HCI. Regression sample is restricted to countries that have data on all the covariates included in the regression model. "East Asia," "MENA," "South Asia," and "SSA" are regional dummies that take the value of 0 or 1 depending on whether the country belongs to the region. Latin America and the Caribbean is the omitted category. Monthly median consumption/income data are for available year in World Bank's PovcalNet in terms of 2011 PPP. We use consumption data when both income and consumption data are available for a country, and we use figures for "urban" when data is only available dis-aggregated by "urban" and "rural." "HCI Expected Years of School" are calculated using repetition-adjusted enrollment rates by school level to proxy for age-specific enrollment rates up to age 18 and taken from World Bank's HCI data. "BL Avg Schooling (20-29)" and "BL Avg Schooling (40-49)" are from Barro and Lee 2010, and are weighted by population. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5." Linguistic diversity index is Greenberg's diversity index from Ethnologue (data accessed from website on February 7, 2019). The index reports the probability that any two randomly selected individuals from a country would have different mother tongues. A value of 0 indicates that everyone has the same mother tongue, while a value of 1 indicates that no two people have the same mother tongue. The computation of the diversity index is based on the population of each language as a proportion of the total population. Measure of fragility is taken from Fund for Peace (accessed on March 1, 2019) and is for the year 2018. The fragile states index is based on different factors countries face that impact their level of fragility. Details about the methodology can be found here. Asterisks (*, **, ***) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

D.2 Regressions Using IRT-Equated Scores *Across* Regions

Table D.2.1: IRT-Equated Scores Across Regions

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4
East Asia	134.154**	127.327**	123.766*	117.442*
	(64.104)	(58.516)	(60.288)	(57.893)
MENA	9.079	19.478	30.811	56.809
	(64.104)	(58.602)	(65.580)	(64.642)
SSA	-53.424	-1.727	5.640	8.861
	(62.068)	(60.358)	(63.049)	(60.451)
Log GDP Per Capita (2011 PPP)		30.328**	18.887	-7.834
		(12.309)	(17.256)	(22.697)
BL Avg Schooling (20-29)			6.411	7.601
			(14.703)	(14.108)
BL Avg Schooling (40-49)			0.580	-4.678
			(13.943)	(13.707)
Gov Effectiveness				56.252
				(32.726)
No. of countries	29	29	29	29
R-squared	0.65	0.72	0.73	0.76

The dependent variable is IRT-equated scores from Sandefur (2018) for SACMEQ III and TIMSS 2003 (Grades 7 and 8 scale). Regression sample drops OECD countries and countries from Europe and Central Asia, and is restricted to countries that have data on all the covariates included in the regression model. Latin America is the omitted category. Chile is the only Latin American country in the sample. "East Asia," "MENA," and "SSA" are regional dummies that take the value of 0 or 1 depending on whether the country belongs to the region. GDP per capita (constant 2011 international dollars and PPP adjusted) is for the most recent available year from the World Bank's World Development Indicators. We take natural log of the GDP figure. "BL Avg Schooling (45-49)" and "BL Avg Schooling (20-29)" are from Barro and Lee 2010, and are weighted by population. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5." Asterisks (∗, ∗∗, ∗∗∗) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

Table D.2.2: IRT-Equated Scores Across Regions

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4
East Asia	123.318*	120.113*	117.598*	153.141**
	(64.265)	(59.335)	(61.301)	(58.837)
MENA	9.079	18.851	30.425	82.111
	(63.606)	(58.867)	(66.143)	(63.461)
SSA	-53.424	-4.844	2.693	38.318
	(61.586)	(60.728)	(63.698)	(58.921)
Log GDP Per Capita (2011 PPP)		28.499**	17.972	-12.061
		(12.534)	(17.443)	(21.553)
BL Avg Schooling (20-29)			5.490	-1.117
			(14.875)	(13.877)
BL Avg Schooling (40-49)			1.081	-2.223
			(14.076)	(12.903)
Gov Effectiveness				95.511*
				(54.521)
Language Diversity				-111.708**
				(48.732)
Fragility				1.461
				(1.780)
No. of countries	28	28	28	28
R-squared	0.61	0.68	0.69	0.79

The dependent variable is IRT-equated scores from Sandefur (2018) for SACMEQ III and TIMSS 2003 (Grades 7 and 8 scale). Regression sample drops OECD countries and countries from Europe and Central Asia, and is restricted to countries that have data on all the covariates included in the regression model. Latin America is the omitted category. Chile is the only Latin American country in the sample. "East Asia," "MENA," and "SSA" are regional dummies that take the value of 0 or 1 depending on whether the country belongs to the region. GDP per capita (constant 2011 international dollars and PPP adjusted) is for the most recent available year from the World Bank's World Development Indicators. We take natural log of the GDP figure. "BL Avg Schooling (45-49)" and "BL Avg Schooling (20-29)" are from Barro and Lee 2010, and are weighted by population. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5." Linguistic diversity index is Greenberg's diversity index from Ethnologue (data accessed from website on February 7, 2019). The index reports the probability that any two randomly selected individuals from a country would have different mother tongues. A value of 0 indicates that everyone has the same mother tongue, while a value of 1 indicates that no two people have the same mother tongue. The computation of the diversity index is based on the population of each language as a proportion of the total population. Measure of fragility is taken from Fund for Peace (accessed on March 1, 2019) and is for the year 2018. The fragile states index is based on different factors countries face that impact their level of fragility. Details about the methodology can be found here. Asterisks (∗, ∗∗, ∗∗∗) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

Table D.2.3: IRT-Equated Scores Across Regions

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4
East Asia	104.268	77.773*	69.672	72.579
	(62.369)	(44.228)	(43.976)	(45.515)
MENA	12.522	52.855	31.865	38.835
	(61.497)	(44.069)	(45.864)	(49.836)
SSA	-53.424	13.309	7.518	10.234
	(58.933)	(43.702)	(44.126)	(45.597)
Median Income/Consumption		0.203***	0.229***	0.220***
		(0.042)	(0.054)	(0.060)
BL Avg Schooling (20-29)			14.555	13.012
			(9.522)	(10.428)
BL Avg Schooling (40-49)			-16.480	-15.986
			(10.278)	(10.578)
Gov Effectiveness				8.170
				(19.762)
No. of countries	26	26	26	26
R-squared	0.57	0.80	0.82	0.82

The dependent variable is IRT-equated scores from Sandefur (2018) for SACMEQ III and TIMSS 2003 (Grades 7 and 8 scale). Regression sample drops OECD countries and countries from Europe and Central Asia, and is restricted to countries that have data on all the covariates included in the regression model. Latin America is the omitted category. Chile is the only Latin American country in the sample. "East Asia," "MENA," and "SSA" are regional dummies that take the value of 0 or 1 depending on whether the country belongs to the region. Monthly median consumption/income data are for available year in World Bank's PovcalNet in terms of 2011 PPP. We use consumption data when both income and consumption data are available for a country, and we use figures for "urban" when data is only available dis-aggregated by "urban" and "rural." "BL Avg Schooling (45-49)" and "BL Avg Schooling (20-29)" are from Barro and Lee 2010, and are weighted by population. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5." Asterisks (∗, ∗∗, ∗∗∗) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

Table D.2.4: IRT-Equated Scores Across Regions

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4
East Asia	104.268	77.773*	69.672	93.170*
	(62.369)	(44.228)	(43.976)	(49.669)
MENA	12.522	52.855	31.865	46.387
	(61.497)	(44.069)	(45.864)	(50.632)
SSA	-53.424	13.309	7.518	22.669
	(58.933)	(43.702)	(44.126)	(45.330)
Median Income/Consumption		0.203***	0.229***	0.171**
		(0.042)	(0.054)	(0.069)
BL Avg Schooling (20-29)			14.555	4.255
			(9.522)	(10.766)
BL Avg Schooling (40-49)			-16.480	-10.503
			(10.278)	(10.542)
Gov Effectiveness				61.327
				(37.981)
Language Diversity				-86.149*
				(42.406)
Fragility				2.009
				(1.392)
No. of countries	26	26	26	26
R-squared	0.57	0.80	0.82	0.86

The dependent variable is IRT-equated scores from Sandefur (2018) for SACMEQ III and TIMSS 2003 (Grades 7 and 8 scale). Regression sample drops OECD countries and countries from Europe and Central Asia, and is restricted to countries that have data on all the covariates included in the regression model. Latin America is the omitted category. Chile is the only Latin American country in the sample. "East Asia," "MENA," and "SSA" are regional dummies that take the value of 0 or 1 depending on whether the country belongs to the region. Monthly median consumption/income data are for available year in World Bank's PovcalNet in terms of 2011 PPP. We use consumption data when both income and consumption data are available for a country, and we use figures for "urban" when data is only available dis-aggregated by "urban" and "rural." "BL Avg Schooling (45-49)" and "BL Avg Schooling (20-29)" are from Barro and Lee 2010, and are weighted by population. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5." Linguistic diversity index is Greenberg's diversity index from Ethnologue (data accessed from website on February 7, 2019). The index reports the probability that any two randomly selected individuals from a country would have different mother tongues. A value of 0 indicates that everyone has the same mother tongue, while a value of 1 indicates that no two people have the same mother tongue. The computation of the diversity index is based on the population of each language as a proportion of the total population. Measure of fragility is taken from Fund for Peace (accessed on March 1, 2019) and is for the year 2018. The fragile states index is based on different factors countries face that impact their level of fragility. Details about the methodology can be found here. Asterisks (∗, ∗∗, ∗∗∗) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.

D.3 Regressions Using Harmonized Test Scores *Within* sub-Saharan Africa

Table D.3.1: Harmonized Test Scores Within sub-Saharan Africa

	(1) Column 1	(2) Column 2	(3) Column 3	(4) Column 4	(5) Column 5
Median Income/Consumption	0.266***	0.221***	0.156**	0.157**	0.091
	(0.061)	(0.069)	(0.062)	(0.062)	(0.084)
Language Diversity		-31.646	-31.250	-31.279	-34.979
		(23.518)	(23.932)	(24.630)	(21.793)
Fragility Index			-0.583	-0.589	-0.274
			(0.498)	(0.844)	(0.879)
Gov Effectiveness				-0.218	-1.495
				(19.385)	(17.530)
HCI Expected Years of School					6.572
					(5.669)
No. of countries	39	39	39	39	39
R-squared	0.21	0.24	0.27	0.27	0.31

The dependent variable is harmonized test scores from the World Bank's HCI. Regression analysis runs on a restricted sample of African countries only. Monthly median consumption/income data are for available year in World Bank's PovcalNet in terms of 2011 PPP. We use consumption data when both income and consumption data are available for a country, and we use figures for "urban" when data is only available dis-aggregated by "urban" and "rural." Linguistic diversity index is Greenberg's diversity index from Ethnologue (data accessed from website on February 7, 2019). The index reports the probability that any two randomly selected individuals from a country would have different mother tongues. A value of 0 indicates that everyone has the same mother tongue, while a value of 1 indicates that no two people have the same mother tongue. The computation of the diversity index is based on the population of each language as a proportion of the total population. Measure of fragility is taken from Fund for Peace (accessed on March 1, 2019) and is for the year 2018. The fragile states index is based on different factors countries face that impact their level of fragility. Details about the methodology can be found here. Government effectiveness data is taken from the World Bank's Governance Indicators for the year 2017. According to the World Bank definition, the indicator captures "perceptions of the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.5." "HCI Expected Years of School" are calculated using repetition-adjusted enrollment rates by school level to proxy for age-specific enrollment rates up to age 18 and taken from World Bank's HCI data. Asterisks (*, **, ***) denote coefficients that are significantly different from zero at the 10, 5, and 1% levels, respectively. Robust standard errors are reported in parentheses.