United Nations
Educational, Scientific and
Cultural Organization

UNESCO
INSTITUTE
FOR
STATISTICS

Information Paper No. 46
October 2017

Mind the Gap: Proposal for a Standardised Measure for SDG 4–Education 2030 Agenda

UNESCO

The constitution of the United Nations Educational, Scientific and Cultural Organization (UNESCO) was adopted by 20 countries at the London Conference in November 1945 and entered into effect on 4 November 1946. The Organization currently has 195 Member States and 10 Associate Members.

The main objective of UNESCO is to contribute to peace and security in the world by promoting collaboration among nations through education, science, culture and communication in order to foster universal respect for justice, the rule of law, and the human rights and fundamental freedoms that are affirmed for the peoples of the world, without distinction of race, sex, language or religion, by the Charter of the United Nations.

To fulfil its mandate, UNESCO performs five principal functions: 1) prospective studies on education, science, culture and communication for tomorrow's world; 2) the advancement, transfer and sharing of knowledge through research, training and teaching activities; 3) standard-setting actions for the preparation and adoption of internal instruments and statutory recommendations; 4) expertise through technical cooperation to Member States for their development policies and projects; and 5) the exchange of specialized information.

UNESCO Institute for Statistics

The UNESCO Institute for Statistics (UIS) is the statistical office of UNESCO and is the UN depository for global statistics in the fields of education, science, technology and innovation, culture and communication.

The UIS was established in 1999. It was created to improve UNESCO's statistical programme and to develop and deliver the timely, accurate and policy-relevant statistics needed in today's increasingly complex and rapidly changing social, political and economic environments.

This paper was written by Mr Nadir Altinok, BETA, CNRS & University of Lorraine (France).

BETA, UFR Faculté de Droit, Sciences Economiques et Gestion, 13, place Carnot CO 70026 54035 Nancy cedex, France. Tel: +33 3 72 74 84 52. Email: nadir.altinok@univ-lorraine.fr

Published in 2017 by:

UNESCO Institute for Statistics
P.O. Box 6128, Succursale Centre-Ville
Montreal, Quebec H3C 3J7
Canada

Tel: +1 514-343-6880
Email: uis.publications@unesco.org
http://www.uis.unesco.org

ISBN 978-92-9189-216-7
Ref: UIS/2017/ED/TD/9
DOI: https://doi.org/10.15220/978-92-9189-216-7-en

Abstract. Monitoring Sustainable Development Goal 4 requires reliable, high quality and cross-nationally comparable data compiled at regular intervals. Launching such a global assessment scheme would be the ideal – but this will take years, perhaps decades. As a second-best alternative, we use a rigorous yet comprehensive methodology which provides globally comparable data for the proportion of students reaching the Minimum Proficiency Level (MPL) in reading and mathematics. Our approach creates indices of comparison between differing assessments where enough countries participate in both. This enables swift and efficient comparison, since no additional instruments or costs incurred in the anchoring process. Based on this methodology, we obtain an international dataset on students reaching the MPL in both primary and secondary schools for more than 160 countries/localities between 1995 and 2015 and for more than 30 Sub-Saharan African countries. We conduct a series of robustness tests and provide confidence intervals for each estimate in order to enhance reliability for our estimates, and we clearly delineate the limitations of the study. To the best of our knowledge, this study provides the largest and most internationally comparable information available for monitoring Sustainable Development Goal 4 for the education sector.

Keywords: Quality, Human Capital, Education, Database, PISA, TIMSS, SACMEQ.

JEL Classification: C8, I2, N3.

Table of contents

List of tables

List of figures

List of Appendix figures and tables

Acronyms & abbreviations

CONFEMEN	Conference of Ministers of Education of French-Speaking Countries
EAS	*Evaluation des Acquis Scolaires* (national assessment in Burkina Faso)
EFA	Education For All
GPR	gender parity ratio
IEA	International Association for the Evaluation of Educational Achievement
IIEP	International Institute for Educational Planning (UNESCO)
IRT	item response theory
LLECE	Latin American Laboratory for Assessment of the Quality of Education
MDG	Millennium Development Goals
MPL	minimum proficiency level
OECD	Organisation for Economic Co-operation and Development
ONE	*Operativo Nacional de Evaluación* (national assessment in Argentina)
OSG	OECD Standardization Group
PASEC	Programme d'Analyse des Systèmes Éducatifs
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
RPR	residence parity ratio
SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SDG	Sustainable Development Goals
SE	standard error
SERCE	Second Regional Comparative and Explanatory Study
SES	socio-economic status
SSA	sub-Saharan Africa
TERCE	Third Regional Comparative and Explanatory Study
TIMMS	Trends in International Mathematics and Science Study

1. Introduction

This paper presents globally comparable estimates of education quality for a policy-relevant application. In particular, we apply a methodology and database developed by Altinok, Angrist & Patrinos (2017) to enable monitoring of SDG 4. We present estimates for the proportion of students reaching Minimum Proficiency Levels (MPL) in both reading and mathematics covering 160 countries / territories from 1995-2015. We further provide estimates for subsamples in order to assess equity. To our knowledge, this is the first study to provide comparable data for monitoring the education-focused SDGs.

In contrast to the Millennium Development Goals (MDG) and Education for All (EFA), which focused on universal completion of basic education and reducing educational disparities linked to gender, the focus of Sustainable Development Goal 4 (SDG 4), the education goal is "inclusive and equitable quality education and lifelong learning opportunities for all". In total, 17 goals and 169 targets are included in the SDGs. SDG 4 is made up of ten targets, including three means of implementation that focus on how to achieve the outcomes described in the targets. Education is also related to other targets. For instance, education can be linked to public financing of basic services and policy/legal frameworks that provide educational opportunities and the integration of different objectives into national education policies and curricula (UIS, 2016).

Among all the indicators provided, the international community has to address critical measurement challenges within two main groups of indicators: learning outcomes and educational equality. Probably the most important challenge is establishing statistical standards and the need for a high quality of data over time and across countries. Statistical standards consist of definitions, concepts, classification systems and methodologies. At the global level, new data collections and processing may be needed in order to make categories comparable and hence to create metrics that are comparable across countries. Currently, there is lack of consistency of standards and definitions among all stakeholders and even among those international organizations which are involved in the production of education statistics (UIS, 2016).

Five of the ten education targets focus on the learning outcomes of young children, youth and adults. This is a clear shift from the MDGs which mainly focused on access, participation and completion. The SDG agenda, beyond Goal 4, highlights the need to focus on equity. Target 4.5 is the elimination of gender disparities and equal access to all levels of education and vocational training for the vulnerable, including persons with disabilities, indigenous peoples and children in vulnerable situations. Hence, education indicators should not only capture national averages but also their variation across different sections of the population. These are defined by group and individual characteristics such as sex, wealth, location, ethnicity, language or disability. The parity index is a simple ratio calculated by dividing the indicator values for one group (e.g. rural areas) by the values for a comparison group (e.g. urban areas). To calculate the parity indices needed to monitor target 4.5, many of the thematic indicators have to be disaggregated for different dimensions. As a result, equity-related measures represent about 60% of the total number of point estimates needed to complete monitoring of all targets under SDG 4.

Measuring learning is complex. It is critical to address the technical and political challenges to measuring learning and achieving SDG 4. Despite the growing number of learning assessments, there is currently no framework available to put together the various types of assessments and to produce cross-nationally comparable data. Target 4.1 of SDG 4covers the quality of primary and lower secondary education. The current global indicator for this target is the "proportion of children and young people: (i) in Grade 2 or Grade 3; (ii) at the end of primary education; and (iii) at the end of lower secondary education who achieved at least MPL in (a) reading and (b) mathematics". Large-scales assessments can be divided into two categories: school-based or household surveys. School-based assessments include two types: national assessments and cross-national initiatives, which are administered in a number of countries, based on a commonly agreed framework. In this study, we will focus on cross-national initiatives. One of the main challenges for measurement on the global level relates to the definition of what counts as meeting a "minimum competence" level in different national contexts, and thus to generate tools to describe the level of competency.

In this paper, we propose to use all possible results from international and regional student achievement tests in order to obtain comparable results for the proportion of students reaching the MPL in both primary and secondary education. By applying the criterion that some countries took part in different assessments simultaneously, we propose to link assessments with each other by using the results of these "doubloon countries" – countries that participate in both regional and international assessments. Compared to previous research, our project brings at least three significant contributions.

First, while previous research mainly focused on mean scores, we propose a new international benchmark for tracking the students who reach the MPL. We define this new threshold by using different assessments that are more suited for developing countries. Similarly to recent initiatives which are focused on low-income countries, such as "PISA for Development" (Adams and Cresswell, 2016), we propose to use two different benchmarks for both mathematics and reading. Besides the "Standard Skills Benchmark" which is more appropriate to middle-income and high-income countries, we also provide a "Basic Skills Benchmark" for both reading and mathematics. Indeed, we show that focusing on countries with education systems that are still in development requires an additional benchmark. Indeed, although our dataset provides information for more than 160 countries/areas, the statistics are well suited for developing countries and more especially for sub-Saharan African countries.

Second, despite the fact that our methodology is mainly based on "doubloon countries", we are able to obtain a set of four different estimations for each combination between countries, education levels and skills. Our methodology is based on the "linking and equating" approaches (Kolen and Brennan, 2014). While in previous research papers only a single methodology was used, we are able to provide alternative estimations of the results of our anchoring process. These additional estimations permit us to obtain the standard errors of the linking process, alongside traditional standard errors computed within each assessment. By using both standard errors, we are able to provide confidence intervals for the estimation of students reaching MPL around the world.

Probably the most important contribution of this work is to be able to track the proportion of minimum performing students over time and to distinguish between different subsamples for equity purposes. Since we provide comparable scores both across time (between 1995 and 2015) and between different groups within each country, our international anchored dataset includes more than 16,000 combinations of results for students reaching the MPLs. Subsamples included in our dataset are mainly gender-based, or make a distinction between residence of schools, socio-economic levels of families, different languages spoken at home and immigration status.

To our knowledge, this is the first study to provide globally comparable statistics for tracking SDG 4.1.1 within the education sector. While ambitious, we remain realistic, highlighting the limitations and drawbacks of our approach as a second-best alternative. First-order to successful monitoring of the SDGs remains a clear, universal commitment of countries to robust assessment and monitoring in a locally-relevant, globally streamlined fashion.

In Section 2, we present the different assessments included in our paper and a definition of the benchmarks within each assessment. Section 3 describes the methodology used for linking assessments and obtaining a standardized dataset on children reaching MPL. Section 4 presents the main results and Section 5 highlights limitations and robustness checks. We finally conclude by proposing recommendations for future monitoring of SDGs for education.

2. Data and definition of low-performing students

2.1. International and regional student achievement tests

Our application to monitor SDG 4 is enabled by the growth of recent international and regional student achievement tests. Below, we provide a brief description of the various existing learning assessments. Table 1 presents the main characteristics of the assessments presented below. More information about student achievement tests and the definition of proficiency levels can be found in Altinok (2017).

First, we present the two best-known international assessments conducted since 1995. Historically, the International Association for the Evaluation of Educational Achievement (IEA) was the first body to measure individual learning achievement and conduct recurrent surveys for international comparative purposes as early as the early 1960s. The surveys include the highly regarded "Trends in International Mathematics and Science Study" (TIMSS) and "Progress in International Reading Literacy Study" (PIRLS).

a. **TIMSS.** The major survey series from the IEA is the Trends in International Mathematics and Science Study (TIMSS). The central goal of TIMSS is to assess pupils' performance in both subjects and to describe the environment in which they acquired these skills. With this second objective in view, those who launched TIMSS firmly took a policy-oriented approach, since pupils' scores were correlated with the various factors that affected their teaching. Five TIMSS rounds have been held to date. The first, conducted in 1995, covered 45 national educational systems and three groups

of learners.[1] The second round covered 38 educational systems in 1999, examining pupils from secondary education (grade 8). The third round covered 50 educational systems in 2003, focusing on both primary and secondary education (grades 4 and 8). In 2007, the fourth survey covered grades 4 and 8 and more than 66 educational systems while this amount increased to 77 in 2011. The last round was performed in 2015 and covered 63 countries/areas. The precise content of the questionnaires can vary quite a lot but remains systematic across countries. Each topic is given a specific weight (as for example, numbers, algebra and geometry in mathematics subjects and life sciences, physical sciences and the history of science in science subjects).

b. PIRLS. The other major IEA survey is the Progress in International Reading Literacy Study, also known as PIRLS. Up to 2011, three major rounds of PIRLS have been held: in 2001, in 2006 and in 2011. The PIRLS survey tests pupils from primary schools in reading proficiency.[2] For instance, the 2006 PIRLS survey involved 41 countries/areas, only two of which were African countries (Morocco and South Africa). This round included 4 lower-middle-income countries (Georgia, Indonesia, Moldova, Morocco) and 8 upper-middle-income countries (Bulgaria, Islamic Republic of Iran, Lithuania, Macedonia, Federal Yugoslavian Republic, Romania, Russian Federation, South Africa) which took part in PIRLS 2006. The last PIRLS round was carried out together with TIMSS (2011) and included 60 countries/areas.

In this paper, we use all recent IEA studies in two skills (mathematics and reading/literacy). The results and information are taken from official reports (Harmon et al., 1997; Martin et al., 2000; Mullis et al., 2000; Mullis et al., 2003; Mullis et al., 2004; Martin et al., 2007; Mullis et al., 2008; Mullis et al., 2009; Martin et al., 2016; Mullis et al., 2016).

c. PISA. The Organisation for Economic Co-operation and Development (OECD) is another international organization that has carried out standardized international comparisons of pupil achievement. The OECD launched its Programme for International Student Assessment (PISA) in 1997 to meet the need for readily comparable data on student performance. The basic principles underlying PISA studies are the use of an extended concept of "literacy" and an emphasis on lifelong learning. Literacy is considered more broadly because PISA studies are concerned with the pupils' capacity to extrapolate from what they have learnt and apply their knowledge to new settings. More generally, since 2000 PISA has assessed the skills of 15-year-old pupils every three years in a growing number of countries. PISA concentrates on three key areas, namely mathematics, science and literacy and all three domains are assessed in all PISA cycles. The main focus of PISA 2000 was on reading literacy, in the sense that it included an extensive set of tasks in this domain. In PISA 2003, the emphasis was on mathematical skills and in 2006 the focus was on scientific skills. The framework for evaluation remains the same across time so that one cycle's findings can

---

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

2 Similarly to TIMSS, pupils from Grade 4 were chosen.

be compared with those of the others.[3] In 2009/2010, the number of participants was equal to 75 countries/areas against 65 in 2012 and 72 in 2015. Unlike the IEA surveys, PISA assesses only 15-year-old pupils, whatever their school level, whereas the grade is the main criterion in selecting pupils for IEA assessments (and over-all student achievement tests).

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

d. SACMEQ. The Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ) grew out of a very extensive national investigation into the quality of primary education in Zimbabwe in 1991. It was supported by the UNESCO International Institute for Educational Planning (IIEP) (Ross and Postlethwaite, 1991). Keen to follow up this successful initiative, several education ministers in southern and Eastern African countries expressed an interest in the study and wished to take part in such an assessment. Planners from seven countries met in Paris in July 2004 and established SACMEQ as a special group. The 15 SACMEQ-member education ministries are those of Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, the Republic of South Africa, Swaziland, United Republic of Tanzania, United Republic of Tanzania (Zanzibar), Uganda, Zambia and Zimbabwe.

The first SACMEQ round took place between 1995 and 1999. SACMEQ I covered seven different countries and assessed performance in reading at grade 6. The participating countries were Kenya, Malawi, Mauritius, Namibia, United Republic of Tanzania (Zanzibar), Zambia and Zimbabwe. The studies, albeit mainly national in scope, had an international dimension and shared many common features (research issues, instruments, target populations, sampling and analytical procedures). A separate report was prepared for each country. In the second round, which was held between 2000 and 2002 and covered 14 countries and one territory (Zanzibar), performance in mathematics and reading was assessed. The target cohort consisted of grade 6 pupils, as under SACMEQ I. The participating countries were Botswana, Kenya, Lesotho, Malawi, Mauritius, Mozambique, Namibia, Seychelles, the Republic of South Africa, Swaziland, United Republic of Tanzania, United Republic of Tanzania (Zanzibar), Uganda and Zambia.

Several SACMEQ II items were replicated from the TIMSS survey to secure comparable results. The questionnaires were used to collect information on educational inputs, the educational environment and issues relating to the fair allocation of human and material resources. Information about the socio-economic context was gleaned from the pupils' questionnaires. More generally, SACMEQ II included items selected from four previous surveys, namely the *Indicators of the Quality of Education* (Zimbabwe) study, SACMEQ I, TIMSS and the 1985-94 IEA *Reading Literacy Study*.

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

The third SACMEQ round (SACMEQ III) covers the same countries as in 2002 (plus Zimbabwe) and focuses on the achievement levels of grade 6 pupils. The latest round began in 2013 but results are not yet available.

e. PASEC. Surveys under the "Programme d'Analyse des Systèmes Éducatifs" (PASEC, or "Programme of Analysis of Education Systems") of the Conference of Ministers of Education of French-Speaking Countries (CONFEMEN) have been conducted in the French-speaking countries of sub-Saharan Africa. This assessment contains results for primary school performance in mathematics and in French. In both CP2 (the second grade in primary school) and CM1 (grade 5), between 2,000 and 2,500 young learners in about 100 schools, along with their teachers and school heads, were surveyed in each of the countries evaluated. Some countries have taken part in the PASEC survey several times. In contrast to other assessments, the PASEC study was not conducted simultaneously in all countries. Therefore, the participation of countries has varied considerably since 1994.[4] It should be noted that the findings of the first four assessments are not available because data relative to assessments are not available[5]. Moreover, the recent participation of Asian countries such as Cambodia and Lao PDR was carried out with a very different framework, making it impossible to anchor with the remaining countries. A significant modification of the PASEC assessment was conducted in 2014, where 10 countries took part at the same moment in an assessment of their pupils from grades 2 and 6. These tests are not directly comparable with previous PASEC items.

In order to simplify the analysis, we will consider three different rounds of PASEC: the first round includes assessments carried out between 1996 and 2003; PASEC II takes into account evaluations between 2004 and 2010. The latest round of PASEC (PASEC III) was conducted in 2014. Moreover, as scores are not directly and fully comparable between each assessment, an anchoring of major items has been made to allow for international comparability.[6] Currently, the inclusion of PASEC III results is not possible, due to the absence of the release of SACMEQ IV results, which may be available in 2018.

f. LLECE. The network of national education systems in Latin American and Caribbean countries, known as the Latin American Laboratory for Assessment of the Quality of Education (LLECE), was formed in 1994 and is coordinated by the UNESCO Regional Bureau for Education in Latin America and the Caribbean. The main aim of this survey is to collect information on pupil performance and performance-related factors likely to help policymakers to design better educational policies. For this purpose, the LLECE seeks to answer the following questions: What do pupils learn? At what level is learning achieved? What skills are developed? When does learning occur? Under what circumstances does it occur? (Casassus et al., 1998).

---

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

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

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

Assessments conducted by the LLECE therefore focused on learning achievement in reading and mathematics in grades 3 and 4 in 13 countries of the subcontinent (Casassus et al., 1998, 2002), namely Argentina, Bolivia, Brazil, Chile, Columbia, Costa Rica, Cuba, Dominican Republic, Honduras, Mexico, Paraguay, Peru and the Bolivarian Republic of Venezuela (Casassus et al., 1998). In each country, samples of about 4,000 pupils in grade 3 (ages 8 and 9) and grade 4 (ages 9 and 10) were assembled. These surveys covered over 50,000 children, amounting to at least 100 classes per country. In 2006, the second round of the LLECE survey was initiated in the same countries as LLECE I. Data between the two rounds are therefore not directly comparable. Moreover, grades tested partly changed compared to the first study: pupils from grade 3 and grade 6 took part in the Second Regional Comparative and Explanatory Study (SERCE). The latest LLECE round, the Third Regional Comparative and Explanatory Study (TERCE), was done in 2013 in both grades 3 and 6 and included 15 Latin American and Caribbean countries. Our analysis will include both SERCE and TERCE results, since the grade tested is the last grade in all countries.

2.2. Definition of benchmarks for reading and mathematics

In this section, we provide some useful information about the monitoring of indicator 1 which is "Percentage of children/young people (i) in grades 2/3; (ii) at the end of primary; and (iii) at the end of lower secondary achieving at least MPL in (a) reading and (b) mathematics". Since the assessments used in our analysis are based on several different grades, it appears to be possible to evaluate the proportion of pupils/students reaching the low international benchmarks for both primary and lower secondary education levels.

In the SDGs, there is a reference to the proportion of pupils who reach the MPL in mathematics and reading. Although a student mean score can provide a basis for comparison between students, it does not provide a guide to the student's strengths and weaknesses. Item response theory (IRT) permits us to distinguish the level of difficulty of items and hence provides a description of the characteristics of groups of students according to their proficiency levels. As shown in Table 2, the definition of "low performing students" varies greatly among assessments.

In PISA, student results are scaled and items are divided into proficiency levels according to how many students answered each item correctly. There are six levels, from Level One (which is the most basic level), to Level Six (which is the most advanced level). Besides PISA, other assessments also define proficiency levels. TIMSS and PIRLS identify four points along the achievement scale to use as international benchmarks of achievement. These are "Advanced International Benchmark" (with a threshold of 625 points), "High International Benchmark" (550), "Intermediate International Benchmark" (475) and "Low International Benchmark" (400). For SACMEQ, a total of eight different proficiency levels are provided for both reading and mathematics. Rasch IRT was used to establish the difficulty level for each test item. Similar processes were used in LLECE assessment where four different proficiency levels are defined.[7] No specific benchmark was defined in PASEC before the 2014 study. However, a level equal

[7] A different analysis was chosen to obtain SERCE/TERCE comparable benchmarks. This is the reason why results provided in our analysis are not directly comparable to the results presented in official reports.

to 40 points is considered as the minimum according to analyses done on this study (Michaelowa, 2001). In PASEC 2014, while three different levels were created in mathematics, the PASEC team preferred to define four proficiency levels in reading.

For PISA, pupils reaching at least Level 2 can be considered to have reached the minimum level. In reading, this means that pupils' scores should be higher than 407; however, this threshold is equal to 420 in mathematics. Moreover, a report on low-performing students by the OECD defines similar thresholds for the two skills (OECD, 2016a).

In TIMSS assessment, four different proficiency levels are defined and named respectively "Low International Benchmark", "Intermediate International Benchmark", "High International Benchmark" and "Advanced International Benchmark". We propose to define as a minimum level the threshold of "Low International Benchmark". This means that pupils who achieve scores of at least 400 points in mathematics in both grades 4 and 8 can be considered as reaching the minimum level in the considered level. Since the methodology is the same for PIRLS, we propose to use the same threshold (i.e. 400 points).

The new PASEC 2014 assessment defines different proficiency levels. While four different levels are provided for reading, only three proficiency levels are present for mathematics. In reading, the PASEC 2014 international report defines Level 3 for reading and Level 2 for mathematics as a minimum level. In this way, pupils reaching at least Level 3 in reading and Level 2 in mathematics may be considered as reaching the minimum level of proficiency.

Despite the definition of four different proficiency levels, the TERCE assessment does not provide any reference to the minimum level of competency. However, the analysis of competencies acquired by pupils reaching Level II permits us to define this level as the minimum one. Indeed, our study proposes to use Level II for each skill (mathematics and reading) and both grades (grades 3 and 6).[8]

The greatest number of competency levels is provided by the SACMEQ study. In total, eight different levels are present. Among them, Level 3 appears to be the minimum level, since pupils reaching this level are considered as reaching the competency of "basic reading" in reading and "basic numeracy" in mathematics.

It should be noted that all these benchmarks are not aligned and thus a direct comparison may lead to estimation bias. Moreover, there is no indication that the thresholds defined within each assessment are reliable across assessments for the countries that took part in several assessments. Thus, we have to define a homogenous threshold that will be used in our anchored dataset. Since our analysis mainly deals with developing countries, we can use a benchmark as defined in a regional assessment such as SACMEQ or TERCE. We chose the SACMEQ benchmark since this is the only assessment that provides comparable results since 1995 and where there is a rigorous analysis of competences for both reading

[8] Since our analysis aims to provide trends in schooling performance, we used the SERCE benchmarks which are different from the new TERCE benchmarks. Therefore, results provided in our paper may differ from the results published in the TERCE reports, based on the new benchmarks.

and mathematics. On the other hand, SACMEQ is only conducted at the primary level and for grade 6 pupils only. Therefore, it is impossible to obtain an anchored dataset for secondary education. One drawback of using the SACMEQ benchmark is the fact that most pupils from middle-income and high-income countries reach this threshold. In order to better assess the performance of these countries, our study proposes to use two complementary benchmarks for primary education. While the "basic literacy" and "basic numeracy" benchmarks are more suited for education systems that are still being developed, the benchmarks based on TIMSS/PIRLS "Low International Benchmark" can be considered to be more appropriate for middle-income and high-income countries. This point is discussed in Section 5 below. As shown in Figures A.5 and A.6, the most difficult benchmarks are the TIMSS benchmark for mathematics and the PIRLS for reading in primary education.

One way of defining a threshold for secondary education is either by using existing TIMSS or PISA benchmarks. We preferred to use PISA benchmarks, since they are defined in the three skills included in our database, while there is no specific analysis for reading in secondary education in the TIMSS study. Moreover, since more and more countries are taking part in PISA assessment, it appears to be a more relevant study than TIMSS. Unfortunately, we failed to find a specific anchoring for including students from grade 2 in our dataset, since there is currently no international assessment undertaken at this specific grade. Detailed information for these benchmarks is provided in Table 3.

What can students do to reach these benchmarks? Since we refer to SACMEQ/IEA in primary education and to PISA in secondary education, we use the competencies acquired by pupils by means of these assessments in our database. In grade 6, pupils reaching the MPL in reading can "interpret meaning (by matching words and phrases, completing a sentence, matching adjacent words) in a short and simple text by reading forwards or backwards" (Hungi et al., 2010). In mathematics, students can "translate verbal information (presented in a sentence, simple graph or table using one arithmetic operation) in several repeated steps". Moreover, he/she "translates graphical information into fractions, interprets place value of whole numbers up to thousands and interprets simple common everyday units of measurement" (Hungi et al., 2010). Competencies acquired by the IEA benchmark are more important and suppose for mathematics that "students have some basic mathematical knowledge. Students can add and subtract whole numbers. They have some recognition of parallel and perpendicular lines, familiar geometric shapes, and coordinate maps. They can read and complete simple bar graphs and tables." (Mullis et al., 2016). In reading, students can "locate and retrieve an explicitly stated detail. When reading Informational Texts, students can locate and reproduce explicitly stated information that is at the beginning of the text" (Mullis et al., 2012).

Concerning students enrolled in secondary education, we base our competency analysis on information provided in PISA reports, which provide several competency levels. In reading, student can typically do several basic tasks. For instance, "some tasks at this level require the reader to locate one or more pieces of information, which may need to be inferred and may need to meet several conditions. Others require recognizing the main idea in a text, understanding relationships, or construing meaning within a limited part of the text when the information is not prominent and the reader must make low level inferences" (OECD, 2016b). In mathematics, students can typically carry out the following tasks: they

"can interpret and recognize situations in contexts that require no more than direct measure. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions to solve problems involving whole numbers. They are capable of making literal interpretations of the results" (OECD, 2016b).

Since we have defined an international benchmark and highlighted the need for obtaining an anchored dataset, in the next section we present the methodology used to obtain such comparable scores.

3. A methodology of anchoring student achievement tests

Given the diversity of existing achievement tests, there is no single and comparable measure of pupil achievement over all tests. On the contrary, as shown in the previous section, international and regional assessments differ greatly in the definition of what a pupil should know in the respective skill tested. Therefore, we build on the methodology and database presented in Altinok, Angrist & Patrinos (2017) to create comparable estimates across various international and regional assessments. By computing adjusted scores, we provide below the methodology used to obtain the proportion of students reaching the MPL. The basic idea behind the methodology used in this paper is the fact that some countries took part in several assessments. By using the results obtained in these assessments, we are able to obtain anchored achievement tests. This is indeed a quick and efficient method: it does not require any additional assessment with linking items and is based on a clear and basic idea according to which similar participation of several countries in different assessments may be used as anchoring countries.[9] However, for some assessments, we also need to use external assessments that provide comparable results over time. Since these assessments provide trends over time, we anchor these assessments on the unadjusted assessments by using the results of the countries for which national assessments are available. First, we present the different methodologies which can be used for anchoring assessments. Then, we show how we obtained the anchored dataset.

3.1. Presentation of linking methodologies

Alternative methodologies can be used for linking assessments. The procedures used in this paper are based on theory developed around the notion of "equating". Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably (Kolen and Brennan, 2014). The purpose of equating relies on the possibility of adjusting for differences among assessments that are built to be similar in difficulty and content. In our case, assessments are not directly comparable since difficulty and content may differ. Other processes that are similar to equating will then be used and can be referred to as "scaling to achieve comparability" in the terminology of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) or "linking", in the terminology of

[9] Our methodology is quite similar to a linking strategy where a group of similar pupils take part in different tests. By using the results of these pupils, we are able to make a linking across different tests. However, in our strategy, countries do not include exactly the same population, which may lead to estimation bias. We discuss this point in Section 5.

Holland and Dorans (2006), Linn (1993) and Mislevy (1992). As Kolen and Brennan (2014) explain, similar statistical procedures are used in linking and equating, although their purposes are different. Here, we will use the terminology of linking instead of equating since the tests used are purposefully built to be different. Another important difference with the equating/linking process is that instead of comparing the results of similar pupils across assessments our methodology is more focused on the fact that we compare similar countries. Therefore, our approach is quite close to a standard linking theory with the difference that we anchor similar countries (which also include different pupils) instead of similar pupils (which are included in similar countries) as in most cases. This leads us to accept the idea that, although pupils may be different across similar countries including in the linking process, the results are not biased. However, since our methodology relies on the possibility of obtaining different alternative estimations of anchored results, we are able to obtain some confidence intervals which may include the true value of pupils' performance.

Let us suppose that a population of pupils, sampled from the target population *T*, takes two different assessments *X* and *Y*. Here, we suppose that any difference in the score distributions on *X* and *Y* can be attributed entirely to the assessments themselves, since group ability is assumed to be constant.[10] In this simpler context, the traditional linking functions include mean, linear, hybrid, coefficient and equipercentile linking. The linking functions are categorized as straight-linear (i.e. linear), including identity, mean, coefficient, hybrid and linear linking, and curvilinear (i.e. nonlinear), including equipercentile and circle-arc linking. While the straight-line types differ from one another in intercept and slope, the curvilinear lines differ in the number of coordinates on the lines that are estimated, whether all of them or only one.

The goal of linking is to summarize the difference in difficulty between the two tests *X* and *Y*. We would like to link test *X* on the scale of test *Y*, which is called a *Reference Test*, while test *X* is called an *Anchored Test*. For instance, we would like to link a test like PISA 2003 on another assessment like TIMSS 2003. Therefore, PISA 2003 will be the *Anchored Test X* while TIMSS 2003 will be considered as the *Reference Test Y*.

All linking methodologies have different properties and there are discussed in Altinok, Angrist and Patrinos (forthcoming). In our estimation strategy, we will thus use four different linking methods: linear linking (hereafter named *l*), pseudo-linear linking (named *c*), equipercentile linking (named *e*) and presmoothed equipercentile linking (named *p*). Since we are able to obtain several estimations for the same anchored assessment, we also provide additional standard errors based on the difference between these different estimation techniques.[11] Although the true value of the proportion of low performing pupils may be outside the range of the confidence interval provided in our analysis, to the best of our knowledge, this is the first study that provides additional information about the significance of the estimation results. Our preferred method is presmoothed equipercentile linking since this is a methodology based on each percentile of the distribution of scores. Our aim is not to anchor mean

[10] We discuss this assumption in Section 5.

[11] We compute the standard errors by using the following formulae: $s.e. = \sigma/\sqrt{n}$ where σ is the standard deviation of the values obtained for each methodology and n=4 since we are using 4 different linking methodologies.

scores across assessments, but instead to use specific proportions of pupils reaching a given benchmark. Focusing on only mean and standard deviation may not be accurate for our current study.

3.2. Application of the methodology

In the equating/linking theories, the anchoring process is mainly done by either adjusting results from the same population between two tests, or with the same items used in different tests. In our case, since we would like to anchor different student achievement tests, we suppose that the tests can be anchored by looking at countries which took part in different tests at the same time. By applying the different linking methods presented above, we are able to obtain relationships between anchored tests and reference tests. In this paper, we use IEA assessments as reference tests, since these are the only tests which began before 1995 and that include both developed and developing countries. The basic idea of our methodology is thus to look at countries which took part in both IEA assessments (i.e. *Reference Tests X*) and the other assessments that need to be anchored (i.e. *Anchored Tests Y*). However, since IEA assessments are not fully comparable over time, and given the fact that some assessments do not include countries that took part in an IEA assessment, we also use an alternative methodology in these cases.

Let us suppose that we would like to anchor PISA assessment on TIMSS assessment. In this case, anchored assessments are the ones where we have countries that took part in a reference assessment (i.e. IEA assessment). This includes TIMSS (after 1995), PIRLS (after 2001), PISA (after 2003 for mathematics), LLECE and SACMEQ assessments. When within a given assessment, we do not have countries taking part in several assessments, we need to use either an external anchor or another anchor within another regional assessment. This includes the PASEC study where we do not have countries that took part in both PASEC and another international assessment. We are able to link PASEC with IEA assessment by using results from Mauritius, which participated in both SACMEQ and PASEC assessments. Indeed, in the case of PASEC I & II, the linking process is made in two steps: first, we anchor PASEC assessment on SACMEQ assessment. Second, we use the linking equations computed between SACMEQ and TIMSS/PIRLS in order to obtain anchored PASEC results.

While in previous papers, only a pseudo-linear linking method was used, we propose to use different linking methods to adjust assessments with reference assessments and to focus mainly on presmoothed equipercentile linking methodology which is more suited to providing the anchored proportion of students reaching the MPL. We thus obtain different estimations of anchored assessments, which allows us to make comparisons. It should be noted that in order to carry out all linking methods, we need to obtain the micro data of these assessments. Another important point relies on the threshold of the MPL. Since each method provides a different distribution of scores, the threshold of the MPL varies among different linking strategies, although its value is quite similar when a direct comparison is made.

Since we used countries that took part in both an *Anchored* and a *Reference Test*, our linking equations are based on the anchoring process of these countries. Moreover, we distinguished between subpopulations and not only the whole population. However, our linking equations are similar across subsamples in

order to preserve the comparability of results between original and anchored assessments. Therefore, our linking equations will depend on different dimensions (years, levels, skills). For instance, let us suppose that we would like to anchor PISA 2000 with TIMSS 1999 for mathematics. We first have to specify which countries took part in both assessments. In Table 4, we provide the list of the 21 countries that took part in both kind of assessments. However, PISA 2000 and PISA 2003 assessments are not directly comparable for maths scores.[12] In order to preserve the comparability across assessments, we still need to link PISA 2003 and TIMSS 2003 assessments. The number of "doubloon countries" is approximately the same for a PISA 2003 – TIMSS 2003 linking. Based on the results of these countries, the linking equations can be obtained according to the different methods highlighted above. In Table 5, we provide the different parameters used for each linking methodology.[13] Another example of linking between a regional and an international assessment can be obtained by linking SACMEQ III and PIRLS 2006 assessments. Since South Africa took part in both assessments, we can rely on these assessments by comparing the results of South African pupils with the different linking approaches. We obtain different relationships according to each linking method and the anchored results of SACMEQ.

Besides the proportion of low-performing students, we also computed standard error for the estimation of this proportion. In our case, we add two different errors of estimation: the first is from the computation of this proportion which is directly related to each assessment, while the second is obtained when we try to anchor each assessment to the IEA assessment. This implies that IEA assessment and the PISA test in reading do not have an additional standard error relative to the linking process, in contrast to the remaining assessments. Since the two estimation techniques are independent, the standard errors can be added to provide a combined standard error.[14] We used this combined standard error to provide confidence intervals for the proportion of low-performing students. While the computation of standard errors from the unanchored proportion of low-performing students relies on the definition provided in the technical reports of each assessment, there is no specific methodology for the computation of the standard error of the linking process. We therefore used all four estimations of low-performing students from each methodology (linear, pseudo-linear, equipercentile and presmoothed equipercentile linking) in order to obtain a standard error for our linking methodology.

4. Results

Our anchored database includes comparable data for more than 100 countries/areas around the world between 1995 and 2015[15]. Moreover, since we are able to compute the proportion of students reaching MPL for several subsamples, more than one observation for each country and year is available (gender, type of residence, language spoken at home, immigrant status and socio-economic status).

---

[12] The design of PISA assessment in every cycle is focused on a given skill. While in 2000 the major skill was reading, mathematics was chosen in 2003. Therefore, maths scores are directly comparable in PISA between 2000 and 2003, but only between 2003 and the remaining cycles.

[13] Since the equipercentile and the presmoothed equipercentile-linking methodologies are based on each percentile, it is not possible to provide all the parameters.

[14] We compute the global standard error as the square root of the sum of both standard errors.

[15] While the purpose of the current project is to propose some new results for the monitoring of SDG 4, another project from Altinok, Angrist and Patrinos (forthcoming) is more based on the provision of comparable data on education, including additional assessments and alternative proficiency levels. See Altinok, Angrist and Patrinos (forthcoming) for more information.

In order to give some general highlights of the results we first focus on the most recent data available for the countries and then, our study analyses the trends over time.

4.1. Cross-country comparison

In this section, we base our analysis on the most recent information available for all countries. Descriptive statistics for the standard skills benchmarks are presented in Table 6. It should be recalled that the year of data availability is not the same for all countries. While for most countries, we have comparable data for 2014 or 2015, this is not the case for others. For instance, since for primary schools, only PIRLS 2011 results are currently available, we were able to provide comparable scores for 2011 as the latest year for reading in primary education, with the exception of TERCE countries for which the most recent data is 2013. Similarly, since SACMEQ IV results are not currently available, the inclusion of the latest regional student achievement tests in Africa (i.e. PASEC 2014 and SACMEQ IV) is not currently possible. Comparisons between countries should thus be made with caution. For these countries, the data included in our analysis are mainly based on the period 2000-2007. On average, approximately 60% of children reach the MPL in mathematics or reading at the primary level, while this share increases slightly to around 65% at the secondary level. Comparable data are available for more than 120 countries/areas for both levels, although this number is lower for reading at the secondary level. This is mainly due to the fact that PISA assessment is the only assessment for providing comparable results for reading at the secondary school level. As expected, the economic level of countries explains the difference in performance toward achieving the MPL: while more than 85% of pupils reach the MPL in each skill for high income countries, less than 16% is the case for pupils from low-income countries. Sub-Saharan Africa (SSA) is probably the least performing region, while Arab States and Latin America perform lower than expected if we control for their economic level. In contrast, countries from Central Asia and East Asian countries are performing higher than other countries with a similar economic level. Below, we provide additional descriptive statistics by using a regional analysis. In Figures 3 and 4, we present the proportion of pupils reaching the MPL for mathematics and reading in primary education. Results are presented for each region, according to the UN regional classification of regions. Since the number of countries included in each region does not correspond to the total number of countries within each region, results are not completely representative of each region and should be only used for a general overview of the results.[16]

Both basic and standard proficiency levels are presented for each skill. While in most developed countries, students reach both proficiency levels, this is far from the case in countries from SSA and the Arab States. In North American and Western European countries, approximately 93% of students reach the MPL. Only one-fifth of pupils from SSA reach the standard mathematics benchmark. The focus on the basic numeracy benchmark appears to be more accurate for this region with about 54% of pupils reaching this benchmark. For reading, results are quite similar, although the numbers of pupils reaching the standard reading benchmark are significantly lower for SSA countries.

[16] Moreover, in order to obtain the regional means, one should weight for each country population to obtain more accurate results. However, the aim of our analysis is only to highlight the general results. A more detailed analysis should undertake more robust computations to allow for inter-regional analyses.

The same analysis is done for secondary level in Figure 3. Approximately 82% of students reach the MPL in both maths and reading in countries from North America and Western Europe, while this proportion is reduced to 70% for Central and Eastern European countries. Arab States and SSA countries are the lowest performing regions in both skills. The results for SSA should be used with caution, since we have comparable data for less than five countries in each skill. Less than half of enrolled pupils actually reach the MPL in both skills in these two regions, although the performance in reading is somewhat higher in SSA countries.

The SDGs not only require analysis of the proportion of children reaching the MPL but also reduction of inequalities between subpopulations. Therefore, we computed the gender parity ratio (GPR) by dividing the proportion of boys reaching the MPL by the proportion of girls reaching this level. It would then mean that a ratio lower than 1 would indicate that girls are outperforming boys in the given countries. Results are provided in Figure 4. In most regions, gender parity is not achieved. This is especially the case for Arab States where the gender parity ratio is equal to 0.677 for reading, suggesting that girls are likely to reach the MPL by about 47% more than boys. For instance, the gender parity ratio for mathematics is equal to 0.93 in primary education in Tunisia, which means that fewer girls tend to reach the MPL than boys at this level. However, the focus on secondary education provides the exact opposite effect (Figure 5): the GPR is equal to 1.1 which means that girls perform better at the MPL than boys. It may be possible that these results are biased since we do not control for school access and completion in both primary and secondary education. For instance, it is possible that girls have less access to secondary education and this selection effect may decrease the real proportion of girls reaching the MPL, especially in secondary schools. In other developing countries like Togo or Senegal, boys are outperforming girls in primary education (GPR higher than 1.10), suggesting that we still have a long way to go to reduce such inequalities.

Our database not only provides gender parity ratios, but also residence parity ratios and socio-economic parity ratios. Residence parity ratios (RPR) are computed as the ratio of students reaching the MPL from urban areas divided by the proportion of students reaching the MPL who are located in rural areas. In theory, the ratio should be higher than 1.00, suggesting that students from urban areas perform better in the MPL than students from rural areas. This is the case in almost all countries from all regions Figure 6). For instance, the RPR is equal to 1.35 in reading in the Arab States, which means that students from urban areas perform 1.35 times better in the MPL than students from urban areas. In general, inequality indexes based on RPR are higher than the indexes based on GPR. These inequalities do exist in North American and Western European countries, although their preponderance is lower than the remaining countries. For instance, while the situation is quite close to parity in a country such as the Netherlands, the situation is worse in a country like France. Students from urban areas of this country perform about 1.14 times better in reading than their colleagues who are located in rural areas. In developing countries, the situation is often in favour of urban areas. For instance, students from urban areas in Indonesia perform 1.5 times better than their colleagues in rural areas. The analysis of inequalities based on socio-economic status (SES) is not possible for all countries, since this measure is not present in most assessments. However, we are able to provide comparable results for more than 120 countries/areas. Results are presented in Figure 7. Inequalities

based on SES exist in all regions and are clearly greater than inequalities based on other dimensions (Figures 8 and 9). For instance, students from the Arab States with the highest SES (i.e. the wealthiest quintile of this index) perform more than 1.9 times better in reading than students with the lowest SES (i.e. the poorest quintile of this index). Even in a country known for its policy focused on equity issues such as Finland, while a situation close to equity is present at primary level, this is not the case at secondary level. In secondary schools, students with a high SES status perform 1.17 times better than students with the lowest SES status in both reading and mathematics. In other countries, inequalities are even higher (Figure 9). For instance, in Tunisia, the SES parity index is equal to 3 in reading for secondary level. These results are based on the latest data available. It may be interesting to also analyse trends over time within each country.

4.2. Trends over time

Since our application provides comparable data for a broad number of countries since 1995, it appears to be possible to analyse trends over time between 1995 and 2015 for some countries. It should be noted that since our methodology of anchoring increases the standard error of estimation of the proportion of students reaching the MPL, the trends analysis should be made over a long period of time, to avoid the potential issue of measurement error. It is indeed possible that trends observed in short terms such as five years are biased due to our linking strategy. We therefore propose to analyse trends for countries for which we have comparable data in both five year (1995/2000) and three year (2013/2015) intervals. As we will show in Section 5, the linking methodology used in our paper is not perfect and comparisons should be made with caution.

First, we propose to focus on five different countries to show how trends can be analysed for the proportion of students reaching the MPL. Given the fact that mathematics is the most widely used skill for analysing education quality, we use this skill for our example. In Figure 10, we provide trends on the proportion of pupils from primary schools reaching the MPL in mathematics for Chile, the Czech Republic, Germany, South Africa and Thailand. Trends are not available for all years, due to data availability. While the increase in the share of students reaching the MPL is increasing quite constantly in countries like Chile and South Africa, this is not the case for countries like Thailand. In the Czech Republic and Germany, the tendency is not clear, and we can only conclude that there is stagnation in the performance of these countries. The same exercise is carried out for secondary education, a level for which we have more comparable data over time (Figure 11). It is possible to compare trends between 1995 and 2015, which represents a 30-year period. Trends are declining for countries like the Czech Republic and Thailand, while a clear and significant increase is obtained in Chile and South Africa. Similarly to primary education, no clear change can be detected in Germany.

A more detailed trend analysis should be done to better assess to what extent we can conclude that some countries improved their performance over time, based on the proportion of their children reaching the MPL. Although this is beyond the score of this paper, we propose to undertake such analysis in some way. In Table A.1., we compile results for countries with data for both 2003 and 2015 for primary education and for both the 1999/2000 and 2015 periods for secondary education. The number of countries for

which comparable data over time is available is highly reduced and cross-regional analyses must be made with caution. While we have comparable data for 21 countries for the primary level, there are more than 50 countries/areas for the secondary level. Since we have standard errors that are computed by combining original standard errors and the ones from the linking methodology, it is possible to say to what extent the trends observed over time are significant or not. Similarly to a significant student test, we supposed that when the absolute change was higher than twice the mean standard error of the MPL, it could be considered as significant (indicated as either "+" or "i" in the last column of each schooling level). When the change is lower than this threshold, results cannot be considered as significant (indicated as "o" in the last column of each schooling level). At the primary level, we can conclude that trends are significantly declining for five countries/regions (Flemish Belgium, Ontario, Canada, Hungary, the Netherlands and New Zealand) and an increase is observed in nine countries/regions (Quebec, Canada, Cyprus, England, Iran, Japan, Morocco, Norway, Russian Federation and Slovenia). Although changes are detected in the remaining countries, they cannot be considered as significant. At the secondary level, positive trends can be observed in only 12 countries/regions, while comparable data are available for more than 50 countries. For instance, a significant increase of the share of students reaching the MPL in mathematics for secondary education can be observed in countries like Brazil, Italy and South Africa. In contrast, a significant decline can be found in nearly 30 countries/areas, including the Czech Republic, Indonesia, the Netherlands and Tunisia. For countries like Morocco and Israel, despite the fact that a positive trend is observed over time, it is not significantly different from zero, due to the large standard errors computed for these two countries.

5. Robustness checks and limits of the study

The database obtained by using the methodology presented in the previous section may include severe estimation bias, since some assumptions may be not valid. This is the reason why we provided standard errors to highlight the uncertainty surrounding the findings. Below, we present these assumptions and provide some discussion about the validity of our approach.

5.1. Limits related to the methodology

In order to validate our methodology based on "doubloon countries", we have to accept some assumptions to keep intact countries' results of their student achievement tests. These assumptions are mainly based on the fact that we suppose that the populations tested and instruments used are similar across assessments. More generally, we can consider at least four strong differences between achievement tests which may explain why comparability between these assessments should be made with caution.

a. **Differences in score distribution across assessments.** First, there is no reason why the original distribution of scores in each assessment should coincide among themselves. For instance, it may be possible that the distribution of scores in the anchored test for a doubloon country may be different from the distribution of scores of the same country in the reference test. When we use the mean linking, we are supposing that the distribution of scores across assessments is similar. Each

assessment uses its own psychometric methodology and hence the items included within each test are different for each assessment. This means that the degree of difficulty of items may also differ and thus, the distribution of scores may not be exactly the same between assessments. For instance, the items included in the SACMEQ study may be easier than the ones in TIMSS and therefore, the distribution of scores may be more positively skewed for TIMSS and negatively skewed for SACMEQ results. This difference may lead to different thresholds of proficiency levels and perhaps different results for countries that took part in several assessments simultaneously. In our study, since we define an international standardized threshold, we potentially overcome this difficulty. However, in order to verify the accuracy of this assumption, we compared normality of score distributions for each assessment by focusing on "doubloon countries" (Table 7). In theory, the distribution of scores for these countries should be similar in order to proceed to a mean or a pseudo-linear linking. We computed four different measures for testing this normality (mean, standard deviation, skewness and kurtosis). The mean is usually used for testing the central tendency for quantitative variables, while the stamdard deviation (SD) is the most widely used measure of dispersion. Normality is generally evaluated with two additional statistics that are known as skewness and kurtosis. Skewness is a measure of whether a distribution trails off in one direction or another.[17] Kurtosis measures the thickness of the tails of a distribution.[18] As shown in Table 7, there are some differences between anchored and reference assessments. For instance, while the skewness is positive and close to 1 in SACMEQ (anchoring number 5), the skewness is close to 0 in the TIMSS assessment. The comparison of kurtosis allows us to measure the thickness of the tails of a distribution. For using an anchoring which does not take into account the variability of the distribution across assessments, the kurtosis may be similar for all countries used as "doubloon countries". If we still focus on SACMEQ countries, it appears that kurtosis is very different compared to TIMSS assessment. Indeed, while kurtosis is close to a normal distribution in TIMSS, its value is higher than 5 in SACMEQ assessment, indicating that main scores are concentrated in the middle and thus fail to capture very high and very low skill levels. However, the comparison between other anchors does not show very strong differences and hence permit us to perform a linking approach based on several methodologies. In general, we find that our methodology is well suited for all assessments with an exception for SSA studies like PASEC and SACMEQ. One potential solution to this issue is to use either equipercentile or presmoothed equipercentile linking methods which take into account the distribution of results from each assessment. Instead of using only mean scores, these linking methodologies match each percentile from anchored and reference tests and thus provide a one-to-one percentile matching which avoids the potential difference in the distribution of scores. We can then compare adjusted results between each methodology and thus obtain standard errors of the estimation which are as high as the differences between each linking approach. If we focus on the differences between regional and international assessments, some differences for anchored low-performing students are obtained across different linking methodologies. For instance, while we find that 16.4% of pupils from Botswana reached the MPL in 2007 according to the coefficient linking methodology, this proportion is equal

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

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

to 31.2% according to the presmoothed equipercentile linking methodology. Such differences are mainly observed in sub-Saharan African countries where few pupils reach the MPL based on IEA assessments. Given the fact that we are trying to obtain a measure which depends greatly on the distribution of scores, our preferred estimation will be the presmoothed equipercentile linking, since this methodology matches all existing percentiles between assessments, instead of only focusing on mean scores or standard deviation.

b. **Estimation bias may also occur when populations tested differ across assessments used for the linking**. The most evident difference can be obtained between PISA and other assessments. While PISA is an assessment based on the age of the student, the remaining tests focus on the grade tested. This distinction can lead to strong differences in countries where repetition and/or drop-out rates are important. The focus on a single grade may exclude a proportion of students who repeated classes, while assessments based on the age of students may include these groups. Since we consider that populations are similar and comparable across assessments, this difference may lead to estimation bias. It is possible to assess to what extent our results may be distorted by this difference by comparing results between TIMSS and PISA for countries that took part in both assessments. In Table 8, we compare the original results for countries that took part in both PISA and TIMSS assessments, in both maths and science. We run a linear regression to test to what extent results in PISA can explain student performance in TIMSS assessment. Dummy variables were included for both skills and years to control for potential external factors related to these variables. We computed the mean grade tested in each doubloon country and each assessment. In column 1, we regressed PISA results on TIMSS results. While the R squared is very high (approximately 0.8), we find that PISA results are underestimated compared to TIMSS results, regardless of the grade difference. The most interesting point is to control for grade difference and hence test to what extent this grade difference may impact the results of the linking process. When we include both dummies for grade difference (column 2), the overall difference between PISA and TIMSS remains quite similar. However, the dummy for a 2-year difference is not significant, which means that the differences found between PISA and TIMSS are not fully due to grade difference. Despite the fact that a significant and very high amplitude of effect is found on the 3-year difference in grades tested, this concerns only two countries (Malta and New Zealand). The correlation is very high, suggesting that the anchoring between the two assessments is possible. However, for specific countries, we observe diverging results. This is especially the case for the Russian Federation and Kazakhstan where TIMSS results appear to be overestimated. We tried the same specification with the proportion of students reaching the minimum level, in order to see if our previous results remain similar for each point of the distribution of scores (columns 3 and 4). A significant and negative coefficient is found for a 2-year difference between PISA and TIMSS, indicating that countries where most of the students tested are in grade perform less well than the remaining countries. However, this difference is very small since it is lower than 3 percentage points. Similarly to the previous estimation, only two countries are included in the 3-year difference dummy. In order to find which countries diverge between PISA and TIMSS, we plotted in Figures A.3 and A.4 the residuals obtained by using the specification in columns 2 and 4. For a small number of countries, we detect significant differences between PISA and TIMSS scores.

This is the case of South Korea, Bulgaria and Kazakhstan where student performance appears to be higher in TIMSS than in PISA. On the contrary, we find opposite results for Qatar, Norway and the Netherlands. For the majority of countries, the difference is lower than 40 points in residuals, suggesting that the comparison between the two assessments is valid.

c. **The content tested may also vary among assessments.** While in assessments such as PISA and PASEC III, items are more focused on competency skills, in all remaining tests, items are more based on common curricula of countries. This distinction may indeed lead to significant differences in countries that are more based on content knowledge rather than competence knowledge. This is especially true of most developing countries but may also include some developed economies. It is possible to test for this difference by focusing on countries that took part simultaneously in TIMSS and PISA assessments with approximately the same grade. Although grades are not exactly similar, we selected countries that took part in both assessments and where the mean grade tested in PISA was grade 9. This represents a lower number of countries than the number of doubloon countries. It is clear that our estimations are not robust since other factors may explain the differences found between the two results, but this analysis presents at least some robustness analysis which is often lacking in previous studies. Results are provided in columns 5 to 8 in Table 8. If we focus on mean scores, PISA scores are significantly different from TIMSS scores by about 0.8 score points, which is very low. As expected, the restriction to countries where the difference between grades tested is the lowest reduces the difference between PISA and TIMSS scores. When the estimation is made for countries with higher grade differences, the coefficient is higher, but still with a small amplitude (column 6). Results using the low-performing students instead of mean scores are quite similar (columns 7 and 8). We can then conclude that the characteristics directly related to assessments may not bias estimation results, at least when we compare TIMSS and PISA assessments. For the remaining assessments, since we do not have enough doubloon countries, the estimations cannot be performed. But we can reasonably assume that the differences are greater since the education systems of these countries are still developing and thus any difference between assessments may lead to performance divergence across them.

d. **Hypothesis of absence of country-specific factors.** Our methodology supposes that linking equations computed for the doubloon countries are mainly due to the differences between assessments and are independent of country-specific factors. For instance, when we anchor SACMEQ and PIRLS assessments by using South Africa as a doubloon country, we are supposing that the differences in score distribution are only due to the specific characteristics of these two assessments and are independent of the education system in South Africa. Obviously, by using only one or two doubloon countries, our methodology includes severe estimation bias, since within-country specific factors may explain differences found between anchored and reference assessments. When the number of doubloon countries is high, this bias may be lowered. It is possible to test whether the anchoring of PISA assessment on TIMSS assessment is valid by comparing linking equations between different rounds of these assessments. For instance, since PISA and TIMSS were conducted at several points in time almost simultaneously, we can use the link-

ing equations for the doubloon countries for each combination in order to test for the stability of the relationship between our TIMSS and PISA linking approach. We perform two different tests. First, we compare the coefficients between two different anchorings (PISA 2003/TIMSS 2003 and PISA 2015/TIMSS 2015). Second, instead of using all doubloon countries, we divide the sample of these countries into two parts in order to test for the stability of the anchoring process. We therefore adjust scores for each subsample of countries and then compare standardized results for each subsample. We should expect to obtain approximately the same adjusted scores for all countries by using either the full sample of doubloon countries or only half of this sample, regardless of the countries included in each subsample. In the meantime, cutting the sample into two parts reduces the information used for linking assessments and thus reduces the quality of the linking process. Results are presented in Table 9 where we present the standardized scores of the USA according to each anchoring. It is important to note that our standardization is primarily based on scores rather than benchmarks. Therefore, the most important methodological issues may be related to this standardization. In the full sample of doubloon countries, 16 countries were included; however, only half of them were selected in each subsample for the robustness analysis.[19] Theoretically, there should be no difference between each subsample. As shown in Table 9, some differences exist, although their amplitude is often very low. Results obtained from the equipercentile methodology often provide lower scores for PISA 2003 anchoring while we find the opposite in the PISA 2015 anchoring. The difference between the lowest score and the highest score is close to 30 score points in the PISA 2003 anchoring while it is reduced to approximately 10 score points in the PISA 2015 anchoring, suggesting that the increase of the doubloon countries has a clear impact on the accuracy of the linking process.[20] This analysis gives us two main results: first, it is important to highlight that country-specific factors are included in the linking process. These are not fully explained by the achievement tests themselves. This means that our estimation strategy is biased due to these country-specific factors. The second main result relates to the number of doubloon countries. The increase of these countries is highly important in order to reduce the bias related to country-specific factors in the linking process. This is probably the way we should go in the future, given the increase of countries' participation in learning achievement tests.

5.2. Limits related to the choice of the benchmark

a. **Subsample comparisons between assessments.** Our database provides new results for each subsample where data are available. By doing this, we consider that subsamples from different assessments are similar and comparable. The definition of the stratified population may also lead to testing different populations. While PISA assessment is more focused on schools rather than classes, assessments such as TIMSS or SACMEQ stratify both schools and classes.

[19] The full list of countries are Australia, Hong-Kong China, Hungary, Indonesia, Italy, Japan, Korea, Latvia, Netherlands, New Zealand, Norway, Russian Federation, Slovakia, Sweden, Tunisia and the USA. While the first eight countries were included in the first subsample, the remaining eight countries were included in the second subsample.

[20] In total, 30 countries/localities participated to both PISA and TIMSS achievement tests in 2015: Australia, Canada, Chile, Chinese Taipei, Georgia, Hong-Kong China, Hungary, Ireland, Israel, Italy, Japan, Kazakhstan, Jordan, Korea, Lebanon, Lithuania, Malaysia, Malta, New Zealand, Norway, Qatar, Russian Federation, Singapore, Slovenia, Sweden, Thailand, UAE, Turkey, USA and Buenos Aires (Argentina).

Indeed, in TIMSS assessment, preference is given to testing student from intact classes, while in PISA, classes are not stratified at all and students are chosen randomly inside each school. The comparison of subpopulations may be altered by differences between assessments which are dependent on the definition given to each subsample. For instance, differences between students with different socio-economic backgrounds may not lead to the same performance results, even if the items are exactly the same across assessments. These differences rely mostly on contextual questionnaires which may differ greatly between assessments. We tried to assess for this potential difference by comparing the share of specific subsamples between the assessments for the doubloon countries. In theory, there is no specific reason why the share of urban areas should differ among assessments. But, since the definition given to the type of location of schools is different between assessments, differences may occur. In Figures A.3. and A.4. we compare the unweighted share of girls and pupils who live in urban areas in both PISA and TIMSS assessments. Since we are able to compare for several years, countries' names may appear more than once. In general, the difference between PISA and TIMSS is very small. However, in some countries such as Slovenia, the share of pupils living in urban areas is higher in the PISA tests than in TIMSS. In Israel and Romania, the share of girls is higher for PISA than TIMSS. The main reasons for these differences may be due to the fact that PISA and TIMSS assessments are not testing exactly the same population.

b **The definition of the threshold for the minimum level benchmark.** As highlighted in the introduction, we propose two different benchmarks for primary education, since we suppose that some countries may need to focus on an intermediate benchmark which is more suitable for their students' attainment. Indeed, in a number of countries, pupils are not yet completely enrolled in either primary or lower secondary education. Since 2011, the IEA and the OECD have also prepared specific assessments for these countries (i.e. PIRLS literacy, TIMSS numeracy, PISA for development). In order to better assess the proportion of low-performing students, our study uses two complementary benchmarks for primary education: basic numeracy/literacy and standard mathematics/reading. While the "basic skills benchmark" is based on the SACMEQ benchmark, the "standard skills benchmark" uses the TIMSS/PIRLS benchmark. Given the fact that our linking methodology permits us to measure the difficulty between assessments, we used the linking equation from the hybrid method in order to compare the anchored values of each benchmark for all assessments. Results are provided in Figures A.5 and A.6. As already shown, the TIMSS assessment is the most difficult one. The low international benchmark from this study is approximately equivalent to Level 6 in SACMEQ, Level III in LLECE and Advanced Level in PASEC before 2014.[21] On the contrary, the minimum benchmark from SACMEQ provides a more realistic benchmark for low-income countries since this threshold is roughly equivalent to Level 1 in LLECE and the intermediate benchmark in PASEC before 2014. The results are quite similar for reading and suggest that the availability of two complementary benchmarks is more suited for all countries (Figure A.6).

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

In Figure A.7., we present the relationship between the proportion of low performers according to each benchmark in order to test which countries may benefit more from both these benchmarks. If we focus on the bottom left side of Figure A.7. for primary education, we clearly distinguish sub-Saharan African countries for which results in the TIMSS benchmark are very low and hence include very few students. For instance, while the proportion of low-performing students is equal to 61% according to the basic numeracy benchmark (i.e. the SACMEQ benchmark), it decreases to 20% according to the standard mathematics benchmark (i.e. the TIMSS benchmark). We also compared the two benchmarks for secondary education between PISA and TIMSS. The relationship is quite close to 1, suggesting that we do not need to use an additional benchmark at the secondary level.[22] In Figure A.8., we directly compared results between the two benchmarks for sub-Saharan African countries. A significant difference exists between the two benchmarks and this divergence is not linear. It is very important to note that the relationship is not completely linear since our analysis is based on a proportion of pupils reaching a given benchmark. While the difference between the two benchmarks is very high for countries such as Mozambique and Swaziland, the contrary is observed for countries such as Mauritius and Chad. The same comparison was made for Latin American countries (Figure A.9). Almost all pupils in this region reach the basic literacy/numeracy benchmarks, indicating that the TIMSS/PIRLS benchmarks are more suited for these countries. This is particularly the case for reading.

c. **Equity ratios and the choice of two different benchmarks.** Proposing two different benchmarks for primary education raises the question of potential differences in the computation of equity ratios. As shown above, the use of only one benchmark has many drawbacks. The choice of the SACMEQ benchmark would give very optimistic results for all high-income countries, but in the meantime, it will be more reliable for SSA countries. On the contrary, focusing on only TIMSS/PIRLS benchmarks would give very poor results for SSA countries, whereas the education systems are currently changing dramatically since more and more pupils are enrolling in schools. Very few pupils from SSA reach the standard benchmark based on TIMSS/PIRLS studies. Proposing two alternative benchmarks raises concerns about the international comparability of the SDG indicator 4.1.1. In addition, the SDGs not only focus on the proportion of low performing students but also on differences between subsamples. For instance, we may be interested in looking at differences between gender (girls/boys) and type of residence (urban/rural). In Figure A.10, we compare the gender parity ratio and the location parity ratio for each benchmark. In secondary education, the choice of either TIMSS or PISA benchmarks does not give diverging results. In primary education, the choice of TIMSS/PIRLS benchmarks would give very diverging results for a significant number of countries. Interestingly, these countries are mostly from SSA. For instance, while the gender parity ratio would be equal to 1.4 in Mali according to the SACMEQ benchmark, its value would increase to 2 according to the TIMSS benchmark. Which value would be more appropriate? We suppose that given the fact that a very low proportion of pupils reach the TIMSS benchmark,

[22] In fact, two reasons can be given to explain the high correlation of benchmarks for the secondary level. First, there is no low-income country among participants in TIMSS/PISA assessments for secondary education. Second, even if a low-income country was taking part in these assessments, the results would be biased since the coverage ratio of the population tested would be low, due to a lower completion ratio in secondary schools, compared to middle-income and high-income countries.

it would be more realistic to focus on the SACMEQ benchmark for sub-Saharan African countries. For the remaining countries, the TIMSS/PIRLS benchmark can be chosen.

d. **Explaining the differences observed between each linking methodology.** Since our paper proposes different alternative linking methodologies, we can conduct a specific analysis of the difference of the proportion of students reaching the MPL for each methodology. As explained in Section 3, the pre-smoothed equipercentile linking is the most appropriate approach since our main aim is not to obtain anchored mean scores, but rather anchored proportions of students reaching a specific threshold. Since this threshold depends greatly on the distribution of score performance, using a methodology which only uses mean and standard deviation would clearly lead to different results and hence induce a greater estimation bias. Alternatively, if our purpose was only to provide mean scores for each country, the use of the pseudo-linear methodology would be more appropriate in order to avoid for abnormal distribution of scores for specific countries used as doubloon countries in the linking process. In Table 10, we compare results from different linking methodologies for four countries (Albania, Argentina, Kenya and Morocco). Each country corresponds to a special case. First, the case of Albania is shown to be representative of the results obtained when we anchor PISA 2015 with TIMSS 2015 achievement tests. Since the first benchmark deals with the PISA benchmark, we should expect approximately the same proportion of students to reach this MPL. This is indeed the case in all four linking strategies with values which are around 46%, while the official value reported by the OECD is equal to 46.7%. While the pseudo-linear approach provides results with a lower proportion of students reaching the MPL, its value is very close to the remaining linking approaches. Results for the second benchmark are also presented. As we explained in Section 3.2, this benchmark deals with the Low International Benchmark from the TIMSS achievement test. Results are very different between each linking methodology. The MPL from equipercentile and presmoothed equipercentile linking methodologies provide an intermediate value which is more accurate, given the fact that the MPL from the proportion of students reaching the MPL in TIMSS is somewhat higher than that of the same proportion from PISA assessment for countries which took part in both assessments at the same time. Since we have different estimations for the same country, we can thus compute a standard error for the linking estimation. In this case, the standard error for the estimation would be equal to 5.1 which is higher than the original standard error (1.9).[23] The second country is Argentina and deals with the linking between TERCE and TIMSS achievement tests. The first benchmark is the basic numeracy benchmark from the SACMEQ achievement test, while the second benchmark is the TIMSS Low International Benchmark. There is no specific reason for the original proportion of students reaching the MPL from TERCE to be the same with each standardized benchmark. While differences are very small between the original and the first benchmark, results are significantly diminished when we use the "standard mathematics" benchmark from the TIMSS study. As shown in Figures A.1. and A.2., these results are quite logical since the TERCE benchmarks have very low thresholds compared to those of TIMSS. Again, results from equipercentile linking strategies appear to be the most appropriate in our case since the values are always between the extreme cas-

[23] Since we can reasonably accept that the linking process is independent of the achievement test itself, standard errors (SE) of the estimations can be added. Indeed, SE are computed as follows:

es of linear and pseudo-linear linking methodologies. We proceed to the comparison of the SACMEQ and TIMSS achievement tests by looking at results from Kenya. As expected, the proportion of students reaching the MPL is very close to the original values from SACMEQ for benchmark 1 which corresponds to the "Basic Numeracy" benchmark, based on the SACMEQ achievement test. For this reason, standard errors are very close to the original results of the SACMEQ study. However, the rescaling to the TIMSS benchmark (i.e. "Standard Mathematics"') increases the value of the standard errors since very different coefficients are available for each linking methodology. More especially, the pseudo-linear approach, which does not take into account the variability of the distribution, provides a very low proportion of students reaching the MPL (29%), compared to the remaining linking approaches (more than 40%). Finally, the case of Morocco is used to obtain the proportion of students reaching the PISA benchmark (i.e. MPL Benchmark 1) to show a country for which we only have data from the TIMSS study. As we can see, the results are somewhat different for each methodology and this difference is converted into the standard error, which is multiplied by roughly 4. The use of equipercentile methodology permits us to account for the distribution of scores, not only the mean values and thus provides a more accurate value of the MPL for countries like Morocco where the overall performance is significantly lower than OECD countries. What can we learn from these results? First, we can say that results obtained from PISA sources are very close to the official benchmark provided by the OECD. Second, when we try to adjust one assessment to an external benchmark, it creates an estimation bias which is taken into account in our standard errors. Third, the equipercentile and the presmoothed equipercentile linking methodologies are more appropriate for the estimation of minimum level performers since these approaches are based on the distribution of scores instead of the first two moments (i.e. mean and standard deviation). Fourth, when we choose equipercentile linking, the proportion of children reaching the MPL is increased since it reflects more precisely children who are located around the threshold defined as the MPL.

5.3. Similarities and differences between international/regional and national assessments

In this section, we propose to compare results from different sources to see to what extent the results converge. The ideal case would be to obtain at least two datasets from national assessment for each regional/international student achievement test. Unfortunately, due to time constraints, we failed to obtain such data. Moreover, it should be noted that national assessments are difficult to obtain, due to specific rules within each country which are often very strict.[24] We were able to obtain two different datasets from two countries (Burkina Faso and Argentina). Therefore, we can make a comparison between the results obtained from national assessments and two different regional assessments (PASEC and TERCE).

We first compare results from a national assessment conducted in Argentina (*Operativo Nacional de Evaluación*, *ONE*) and the results from TERCE which is a regional assessment. Both assessments were conducted in 2013 and included pupils from grades 3 and 6. Since our dataset only includes

[24] For instance, it is currently impossible for a researcher outside the United States to obtain the dataset relative to the National Assessment of Educational Progress.

standardized data for grade 6 pupils, we will compare only children from this grade. The areas assessed in the *ONE* study are mathematics, language, social sciences and natural sciences, while in TERCE, only three skills are included (mathematics, reading and science). TERCE and *ONE* have many similarities, both being a test based on curriculum, although the *ONE* study may be more adapted to the curriculum of Argentina, since it is a national assessment. Both assessments use an IRT model for the evaluation of pupils, allowing over-time comparisons. Indeed, the ONE study allows a comparison between two waves (2010 and 2013), while the TERCE is also comparable with the SERCE study, conducted in 2006. Both assessments use questionnaires with multiple choice items and a number of open-ended questions. In *ONE*, three different proficiency levels are proposed for each skill: high (*alto*), medium (*medio*) and low (*bajo*), while the TERCE study includes five proficiency levels (below level I and level I to level IV – see Table 2 for more information). Since we consider in our study the MPL as being level II of TERCE, it is more appropriate to compare Level II of TERCE and the low level of *ONE* (*bajo*) in each skill and grade 6. As shown in Table 11, the proportion of pupils who reach the MPL is different when we compare the *ONE* assessment and the TERCE results. While in *ONE*, approximately 64% of pupils reached the MPL in maths in 2013, more than 92% of pupils reached this level according to TERCE results. Surprisingly, this proportion is very close to the *ONE* results in our anchored database (64.8% versus 64.4%). For reading, the difference is smaller among all three different datasets, and the results are quite similar between our anchored database and the national assessment (76.8% versus 72.0%, respectively). A trend analysis can also be made, but since the period of trends is different, comparison may be flawed. While the period only covers 3 years in the *ONE* assessment, it is extended to 7 years for the regional assessment. Trends are quite similar although they may appear to be different. In fact, we can conclude that the proportion of pupils reaching the MPL do not significantly change over time, regardless of the assessment used. In all trends, the variation is very low, which implies a global stagnation of minimum level performers in Argentina between 2006 and 2013.

We carried out the same analysis for Burkina Faso from which we received results from the national assessment for grade 6 pupils conducted in 2012. The assessment is called *Evaluation des Acquis Scolaires (EAS)* and includes approximately 9,000 pupils who are enrolled in grade 6 schools. In contrast to the *ONE*, the *EAS* assessment is not based on IRT methodology and thus does not provide any proficiency level. However, since the study is quite similar to the PASEC assessments before 2014, we can consider that the MPL is equal to 40 points (Michaelowa, 2001). It should also be noted that in our current version of the standardized database, we cannot include the PASEC 2014 assessment, due to the lack of SACMEQ IV results, which are expected for 2018. The comparison between the EAS and PASEC 2014 show that the proportion of minimum level performers is quite similar between the two assessments. While the national assessment tends to slightly overestimate this proportion, the gender ratio between the two tests is very close, suggesting that the PASEC 2014 assessment can be used as a tool to evaluate a country such as Burkina Faso.

6. Conclusion and recommendations

In this paper, we provided a new measure for tracking the performance of pupils at each education level and thus filling the gap regarding the SDGs for the education sector. We propose to use all possible results from international and regional student achievement tests to obtain comparable results for the proportion of students reaching the MPL in both primary and secondary education. By considering the fact that some countries took part in different assessments simultaneously, we proposed to link assessments with each other by using the results of these doubloon countries.

To obtain a global picture of performance around the world, our study focuses on a new international benchmark for tracking the students who reach the MPL. We define this new threshold by using different assessments which are more suited for developing countries. Similarly to recent initiatives which are focused on low-income countries, such as PISA for Development (Adams and Cresswell, 2016), we propose to use two different benchmarks for both mathematics and reading. Besides the Standard Skills Benchmark which is more appropriate to middle-income and high-income countries, we also provide a Basic Skills Benchmark for both reading and mathematics. Indeed, we show that focusing on countries with education systems that are still developing requires an additional benchmark. Although our dataset provides information for more than 160 countries/areas, the statistics are well suited for developing countries and more especially for SSA countries.

Since our methodology of anchoring is based on assumptions which are not completely valid, we conduct the standardization with a set of four different estimations for each combination between countries, education levels and skills. While in previous research papers, only a single methodology was used, we are able to provide alternative estimations of the results of our anchoring process. By using standard errors of the anchoring process, we are able to provide confidence intervals for the estimation of students reaching MPLs around the world.

A significant contribution of this work is to be able to track the proportion of minimum performing students over time and to distinguish between different subsamples for equity purposes. Since we provide comparable scores both across time (between 1995 and 2015) and between different groups within each country, our international anchored dataset includes more than 16,000 combinations of results for students reaching the MPLs. Subsamples included in our dataset are mainly gender-based, or make a distinction between location of schools, socio-economic levels of families, different languages spoken at home and immigration status. Our anchored database includes comparable data for more than 100 countries/areas around the world between 1995 and 2015.

The database obtained by using the methodology presented in this paper may include severe estimation bias, since some assumptions may be not valid. This is the reason why we provided standard errors to highlight the uncertainty surrounding the findings. For purposes of clarity, our paper presents these assumptions and provides some discussion about the validity of our approach.

To what extent can our standardized database fill in the gap of comparable data for tracking SDGs for the education sector? Obviously, it would take several years to conduct a comparable achievement test which includes most countries. We have to use existing data and try to track the progress of countries over time. Although the dataset provided in our paper cannot be considered to be perfect, it proposes a first overview of countries' performance in education systems and hence a global picture of SDG progress for the education sector.

A next step in this work would probably be the coordination of different actors who organize achievement tests in order to share items across tests. This coordination will facilitate the comparability of student achievement tests and reduce the estimation bias. However, it may take several years and maybe several decades to obtain such results.

References

Adams, R. and Cresswell, J. (2016). *PISA For Development Technical Strand A. Enhancement of PISA Cognitive Instruments* (1993-9019). doi: 10.1787/5jm5fb3f85s0-en

American Educational Research Association (AERA), National Council on Measurement in Education (NCME) and American Psychological Association (APA) (1999). *The Standards for Educational and Psychological Testing*". Washington, D.C.: American Psychological Association.

Albano, A. D. (2016). "equate: An R package for observed-score linking and equating", *Journal of Statistical Software, 74*(8), 1-36.

Altinok, N. and Diebolt, C. et al. (2014). "A new international database on education quality: 1965–2010", *Applied Economics, 46*(11), 1212-1247.

Altinok, N. (2017). « The contribution of learning achievement tests to the monitoring of SDG 4 ». *UIS Discussion Paper*.

Altinok, N., Angrist, N., & Patrinos, H. (2017). A Global Data Set on Educational Quality (1965-2015). Education Global Practice, World Bank Group (draft).

Angrist, N., Patrinos, H. A. and Schlotter, M. (2013). "An Expansion of a Global Data Set on Educational Quality: A Focus on Achievement in Developing Countries", *The World Bank: Policy Research Working Papers*. doi:10.1596/1813-9450-6536

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

Casassus, J., Froemel, J. et al. (2002). *First International Comparative Study of Language, Mathematics, and Associated Factors in Third and Fourth Grade. Second Report*. Santiago, Chile: UNESCO, Regional Office of Education for Latin America and the Caribbean.

Casassus, J., Froemel et al. (1998). *First International Comparative Study of Language, Mathematics, and Associated Factors in Third and Fourth Grade.* Santiago, Chile: UNESCO, Regional Office of Education for Latin America and the Caribbean.

Hanushek, E. A., and Woessmann, L. (2015). *The Knowledge Capital of Nations: Education and the Economics of Growth*: The MIT Press. doi: 10.7551/mitpress/9780262029179.001.0001

Hanushek, E. A., and Woessmann, L. (2012). "Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation". In *Journal of Economic Growth, 17*(4), 267-321. doi:10.1007/s10887-012-9081-x

Harmon, M., Smith, T. A. et al. (1997). *Performance Assessment: IEA's Third International Mathematics and Science Study (TIMSS)*: International Association for the Evaluation of Educational Achievement (IEA): Chestnut Hill, MA.

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

Holland, P. W., and Thayer, D. T. (1998). "Univariate and bivariate loglinear models for discrete test score distributions". In *ETS Research Report Series, 1998*(2).

Hungi, N., Makuwa, D. et al. (2010). SACMEQ III project results: Pupil achievement levels in reading and mathematics. *Working document, 1*. Available at: https://learningportal.iiep.unesco.org/en/notice/030689

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

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

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

Martin, M. O., Mullis, I. V. et al. (2000). *Effective schools in science and mathematics: IEA's Third International Mathematics and Science Study*. TIMSS International Study Center: Boston College.

Martin, M. O., Mullis, I. V. et al. (2007). *Progress in International Reading Literacy Study (PIRLS): PIRLS 2006 Technical Report*: Education Research Information Center (ERIC), Institute of Education Sciences (IES): Washington, D.C.

Martin, M. O., Mullis, I. V. et al. (2012). *TIMSS 2011 International Results in Science*: TIMSS International Study Center: Boston College.

Michaelowa, K. (2001). "Scolarisation et acquis des élèves: les indicateurs de résultats dans l'analyse des politiques de l'enseignement en Afrique francophone". *Politiques d'Education et de Formation, 1*(3), 77-94.

Mislevy, R. J. (1992). *Linking Educational Assessments: Concepts, Issues, Methods, and Prospects*, ETS Policy Information Center Report, ETS: Washington, D.C.

Mullis, I. V., Martin, M. O. et al. (2012). *TIMSS 2011 international results in mathematics*: ERIC, IES: Washington, D.C.

Mullis, I. V., Martin, M. O. et al. (2012). *PIRLS 2011 International Results in Reading*: ERIC, IES: Washington, D.C.

Mullis, I. V., Martin, M. O. et al. (2016). *TIMSS 2015 international results in mathematics*: ERIC, IES: Washington, D.C.

Mullis, I. V., Martin, M. O. et al. (2004). *TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*: ERIC IES: Washington, D.C.

Mullis, I. V., Martin, M. O. et al. (2000). *TIMSS 1999 international mathematics report* International Study Center, Lynch School of Education: Boston College.

Mullis, I. V., Martin, M. O. et al. (2003). *PIRLS 2001 international report*: IEA: Amsterdam.

Mullis, I. V. and Martin, M. O. et al. (2008). « Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades ». International Mathematics Report. TIMSS&PIRLS International Study Center, Boston College.

Mullis, I. V., Martin, M. O. et al. (2009). *TIMSS advanced 2008 international report: Findings from IEA's study of achievement in advanced mathematics and physics in the final year of secondary school*. TIMSS and PIRLS International Study Center, Lynch School of Education: Boston College.

Mullis, I. V. Martin, M.O. et al. (2016). *TIMSS 2015 International Results in Mathematics*. TIMSS and PIRLS International Study Center, Lynch School of Education: Boston College.

Mullis, I. V. Martin, M.O. et al. (2016). *TIMSS 2015 International Results in Science*.

Organisation for Economic Co-operation and Development (OECD) (2010). *PISA 2009 results: learning trends: changes in student performance since 2000 (vol. v)*. OECD Publishing: Paris. Doi: 10.1787/9789264091580-en

OECD (2016). *Low-Performing Students: Why They Fall Behind and How to Help Them Succeed*. OECD Publishing: Paris. Doi:10.1787/9789264250246-en

OECD. (2016). *PISA 2015 Results (Volume I)*- Excellence and Equity in Education. OECD Publishing: Paris. Doi:10.1787/9789264266490-en

Ross, K., and Postlethwaite, T. (1991). *Indicators of the quality of education: A National Study of Primary Schools in Zimbabwe* (Volume I). International Institute for Educational Planning (IIEP) (UNESCO). Zimbabwean Ministry of Education and Culture: Harare.

UNESCO Institute for Statistics (UIS) (2016). "Laying the Foundation to Measure Sustainable Development Goal 4". In *Sustainable Development Data Digest.* UIS: Montreal. Retrieved from http://uis.unesco.org/sites/default/files/documents/laying-the-foundation-to-measure-sdg4-sustainable-development-data-digest-2016-en.pdf

Table 1. Review of main characteristics of large-scale student achievement tests

General information					Estimation methodology				
Nb.	**Assessment**	**Year(s)**	**Countries / areas**	**Grade**	**Methodology**	**Plausible values**	**Mean score**	**Standard deviation**	**Availability of SES index**
1	PISA	2000 – 2003 – 2006 – 2009 – 2012 – 2015	43, 41, 57, 74, 65, 71	15 years old	IRT	Yes	500	100	Yes
2	TIMSS	1995 – 1999 (Grade 8 only) –2003 – 2007 – 2011 – 2015	45, 38, 26, 48, 66, 65, 65	Grades 4, 8, 12	IRT	Yes	500	100	No
3	PIRLS	2001 – 2006 – 2011	35, 41, 55	Grade 4	IRT	Yes	500	100	No
4	PASEC	Different years between 1996 & 2010 – 2014	22 (before 2014), 10 (2014)	Grades 2, 5 (before 2014) & 6 (after 2014)	IRT	Yes	500	100	Yes
5	LLECE	1997, 2006, 2013	13, 16, 16	Grades 3, 4 (for LLECE I), 6 (for LLECE II & III)	IRT	Yes	250 for 1997 700 for 2006/2°13	50 for 1997 100 for 2006/2013	Yes
6	SACMEQ	1995, 2000, 2010	7, 15, 16	Grade 6	IRT	No	500	100	Yes

Table 2. Overview of proficiency levels in international and regional assessments

Nb.	Assessments	Skill	Original names	Level L1*	Level L2	Level L3	Level L4	Level L5	Level L6	Level L7	Level L8
1	PISA	Maths	1-6	<358	359	420	483	546	608	669	
		Science	1-6	<335	336	410	485	560	634	708	
		Reading	1a,1b-6	<262	263	336	408	481	554	626	698
2	TIMSS	Maths	LIB-AIB	<400	400	475	550	625			
		Science	LIB-AIB	<400	400	475	550	625			
3	PIRLS 2011	Reading	LIB-AIB	<400	400	475	550	625			
4	PASEC G2 (a)	Maths	1-3	<66.9	66.9	400	489	577			
		Reading	1-4	<126	126	399	469	540	610		
5	PASEC G6 (a)	Maths	1-3	<68.1	68.1	433	521	609			
		Reading	1-4	<72.1	72.1	365	441	518	595		
6	LLECE G3	Maths	I-IV	<687	688	751	842				
		Reading	I-IV	<675	676	728	812				
7	LLECE G6	Maths	I-IV	<685	686	789	877				
		Science	I-IV	<668	669	782	862				
		Reading	I-IV	<612	613	755	809				
8	SACMEQ III	Maths	1-8	<369	370	466	533	591	648	723	806
		Reading	1-8	<372	373	414	462	514	563	619	704

(a) No specific benchmark was defined in PASEC prior to 2014. Hence, proficiency levels are only valid for 2014.* The number in each case represents the lower boundary. For instance, if a student achieves at least 400 score points, he/she will reach Level L2 in TIMSS 2001 G4. The number of proficiency levels presented in this table may differ from official reports, since we considered that the sum of the proportion of pupils within each proficiency level may be equal to 100%. In some assessments like TIMSS, Level L1 does not exist. Instead, only Levels L2-L5 are defined.

Table 3. Description of the Minimum International Benchmark

Skill	Assessment used for the definition	Lower score limit	What students can typically do
Primary education (Grades 4-6)			
Basic Level – SACMEQ Benchmark			
Mathematics	SACMEQ	466	Translates verbal information (presented in a sentence, simple graph or table using one arithmetic operation) in several repeated steps. Translates graphical information into fractions. Interprets place value of whole numbers up to thousands. Interprets simple common everyday units of measurement.
Reading	SACMEQ	414	Interprets meaning (by matching words and phrases completing a sentence, matching adjacent words) in a short and simple text by reading forwards or backwards.
Standard Level – IEA Benchmark			
Mathematics	TIMSS	400	Students have some basic mathematical knowledge. Students can add and subtract whole numbers. They have some recognition of parallel and perpendicular lines, familiar geometric shapes, and coordinate maps. They can read and complete simple bar graphs and tables.
Reading	PIRLS	400	When reading Literary Texts, students can locate and retrieve an explicitly stated detail. When reading Informational Texts, students can locate and reproduce explicitly stated information that is at the beginning of the text.
Secondary education (Grades 8-11)			
Mathematics	PISA	420	At this level, students can interpret and recognize situations in contexts that require no more than direct inference. They can extract relevant information from a single source and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions to solve problems involving whole numbers. They are capable of making literal interpretations of the results.
Reading	PISA	410	Some tasks at this level require the reader to locate one or more pieces of information, which may need to be inferred and may need to meet several conditions. Others require recognizing the main idea in a text, understanding relationships, or construing meaning within a limited part of the text when the information is not prominent and the reader must make low level inferences. Tasks at this level may involve comparisons or contrasts based on a single feature in the text. Typical reflective tasks at this level require readers to make a comparison or several connections between the text and outside knowledge, by drawing on personal experience and attitudes.

Note: * Lower bounds for each benchmark are original values. Adjusted values may differ according to the linking methodology used.

Table 4. List of countries used for the linking between assessments

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

Table 5. Parameters of the linking methodology

Linking number	Anchored Assessment	Reference Assessment	Number of countries	Mean linking	Linear linking	Linear linking	Pseudo-linear
1	LLECE I, grades 3-4, maths	TIMSS 1995, grade 8	1	-143.72	-261.81	1.24	0.71
2	LLECE I, grades 3-4, reading	PIRLS 2001, Grade 4, reading	2	-84.26	-98.52	1.03	0.83
3	LLECE III, grade 6, maths	TIMSS 2011, grade 4, maths	2	-100.19	-85.21	0.97	0.81
4	LLECE III, grade 6, reading	PIRLS 2011, grade 4, reading	2	-51.31	-64.16	1.03	0.90
5	SACMEQ II, grade 6, maths	TIMSS 2003, Grade 8, maths	2	-184.35	-150.92	0.93	0.63
6	SACMEQ III, grade 6, reading	PIRLS 2006, grade 4, reading	1	-191.93	-277.20	1.17	0.61
7	PISA 2000, 15 years old pupils, maths	TIMSS 1999, grade 8, maths	21	23.48	84.95	0.87	1.05
8	PISA 2003, 15 years old pupils, maths	TIMSS 2003, grade 8, maths	16	17.62	103.68	0.82	1.04
9	PISA 2006, 15 years old pupils, maths	TIMSS 2007, grade 8, maths	25	25.33	67.63	0.91	1.05
10	PISA 2012, 15 years old pupils, maths	TIMSS 2011, grade 8, maths	28	21.09	45.25	0.95	1.04
11	PISA 2015, 15 years old pupils, maths	TIMSS 2015, grade 8, maths	30	28.44	56.42	0.94	1.06
12	PASEC I & II, grade 5, maths	SACMEQ III, maths	1	141.05	346.46	0.57	1.29
13	PASEC I & II, grade 5, reading	SACMEQ III, reading	1	72.98	332.87	0.48	1.15
81	PISA 2003, 15 years old pupils, maths	TIMSS 2003, grade 8, maths	10	27.58	108.79	0.83	1.06
82	PISA 2003, 15 years old pupils, maths	TIMSS 2003, grade 8, maths	8	5.98	104.93	0.80	1.01
91	PISA 2015, 15 years old pupils, maths	TIMSS 2015, grade 8, maths	16	29.83	52.98	0.95	1.06
92	PISA 2015, 15 years old pupils, maths	TIMSS 2015, grade 8, maths	16	22.26	50.44	0.94	1.05

Table 6. Descriptive statistics for the proportion of students reaching the MPL, standardized database

	Primary					Secondary				
Countries	**Nb. of countries**	**Mean**	**SD**	**Minimum**	**Maximum**	**Nb. of countries**	**Mean**	**SD**	**Minimum**	**Maximum**
Mathematics										
All	125	62.85	32.97	0.91	99.79	120	63.91	24.05	9.40	97.75
By Income level										
High Income	56	88.63	14.00	32.92	99.79	69	77.42	15.72	25.72	97.75
Upper-Middle Income	29	62.49	24.50	7.38	98.31	34	49.79	19.69	9.40	96.21
Lower-Middle Income	24	34.40	24.08	0.91	87.59	17	37.30	22.15	12.70	80.90
Low-Income	16	15.95	11.54	3.76	43.41	0	n.a.	n.a.	n.a.	n.a.
By Region										
Arab States	14	47.81	23.54	2.89	86.08	16	53.64	17.37	28.16	88.98
Central & Eastern Europe	15	90.77	6.46	78.77	98.31	21	76.63	14.90	38.16	94.97
Central Asia	5	76.91	11.70	66.74	96.47	6	66.01	23.87	20.53	90.83
East Asia & the Pacific	10	83.21	23.63	34.04	99.79	15	83.41	20.28	38.84	98.93
Latin America & Caribbean	19	53.06	22.38	7.38	82.68	16	42.93	14.58	15.18	63.17
N. America & Western Europe	31	93.37	5.65	75.93	99.03	40	88.17	5.90	72.40	98.09
South & West Asia	1	65.17	n.a.	65.17	65.17	2	43.96	27.37	24.60	63.31
Sub-Saharan Africa	30	21.36	17.71	0.91	66.23	4	41.06	16.99	21.50	61.01
Reading										
All	117	66.24	35.21	0.20	99.55	87	68.24	19.69	16.70	97.09
By Income level										
High Income	57	90.90	12.40	47.42	99.55	51	79.14	9.94	48.43	90.72
Upper-Middle Income	26	68.07	23.11	12.66	99.05	30	55.15	17.85	21.03	97.09
Lower-Middle Income	18	37.61	30.14	0.20	91.26	6	41.15	26.69	16.70	86.16
Low-Income	16	7.58	11.02	0.81	44.76	0	n.a.	n.a.	n.a.	n.a.
By Region										
Arab States	9	50.13	24.15	0.20	75.67	6	42.83	15.82	23.89	62.26
Central & Eastern Europe	14	91.42	10.18	65.76	99.04	19	71.57	14.88	91.93	90.64
Central Asia	2	84.17	3.27	81.86	86.49	4	40.73	19.77	18.60	62.43
East Asia & the Pacific	6	90.77	12.37	66.23	99.20	14	81.00	14.52	48.66	97.58
Latin America & Caribbean	21	72.06	15.91	25.22	92.05	14	57.69	12.98	30.51	80.30
N. America & Western Europe	34	95.58	4.54	74.45	99.55	28	83.32	5.48	66.43	91.16
South & West Asia	1	75.74	n.a.	75.74	75.74	1	19.11	n.a.	19.11	19.11
Sub-Saharan Africa	30	15.56	17.29	0.81	56.70	1	56.37	n.a.	56.37	56.37

Table 7. Robustness check: Comparison of main statistics between assessments for the restricted doubloon countries samples

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

Note: For more information about the list of countries considered as doubloon countries, please consult Table 4.

Table 8. Robustness check: Effect of PISA results on TIMSS scores for doubloon countries

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Mean score		**Proportion of low performing students**		**Mean score**		**Proportion of low performing students**	
	All grades	**All grades**	**All grades**	**All grades**	**1 grade difference**	**>1 grade difference**	**1 grade difference**	**>1 grade difference**
PISA results	0.906	0.929	1.037	1.060	0.784	1.005	0.959	1.139
	(0.052)***	(0.048)***	(0.066)***	(0.061)***	(0.082)***	(0.056)***	(0.058)***	(0.081)***
Difference in grades								
2 years		-7.368		-2.990				
		(6.020)		(1.766)*				
3 years		-35.390		-6.771				
		(10.100)***		(1.992)***				
Dummies for skills	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Dummies for years	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	240	240	240	240	82	158	82	158
Countries	42	42	42	42	19	27	19	27
R squared	0.806	0.820	0.774	0.786	0.731	0.858	0.694	0.849

Note: Cluster-robust standard errors provided in brackets. Clusters are countries. Difference in grades is calculated by the rounded mean grade in PISA and the actual grade in TIMSS. Hence, a difference of two years means that the mean grade tested in PISA is grade 10 while the grade tested in TIMSS is always grade 8. Both mathematics and science scores are included. All years are taken into account (2000, 2003, 2006, 2011 and 2015). Since TIMSS was undertaken in 1999 instead of 2000, we directly compared TIMSS 1999 and PISA 2000 results. A similar comparison was made between TIMSS 2007 and PISA 2006. Since no comparable assessments were done for TIMSS, results from PISA 2009 were not included.

Table 9. Robustness check: Results for anchored value of USA mean score with alternative sub-samples of doubloon countries

	Linear linking	Pseudo-linear linking	Equipercentile linking	Pre-smoothed equipercentile linking	Mean
Anchoring between PISA 2003 and TIMSS 2003 assessments					
1. All doubloon countries	501.58	500.28	481.20	481.10	491.04
2. Only first panel of doubloon countries	510.28	510.53	481.20	481.10	495.76
3. Only second panel of doubloon countries	489.13	488.84	481.20	481.10	485.07
Anchoring between PISA 2015 and TIMSS 2015 assessments					
2. All doubloon countries	497.93	497.40	501.71	501.67	499.68
3. Only first panel of doubloon countries	499.46	498.52	499.61	499.60	499.29
4. Only second panel of doubloon countries	491.27	491.66	506.03	506.00	498.74

Note: Results are based on mathematics for secondary level by comparing the anchored results of PISA 2003 aand 2015 achievement scores for the USA using different samples of countries in the anchoring process. Four methods of linking are presented: linear, pseudo-linear, equipercentile and pre-smoothed equipercentile linking. See text for more information about these linking techniques. Our methodology is first based on score anchoring, then we compute the proportion of students reaching the MPL by using anchored benchmarks from PISA benchmarks.

Table 10. Robustness check: Comparison between different linking strategies, sample of 4 countries

Country	Level	Year	Original MPL		MPL Benchmark 1		MPL Benchmark 2		Assessment adjusted	Linking
			Value	SE	Value	SE	Value	SE		
Albania	Secondary	2015	46.72	1.9	46.40	1.9	72.78	5.1	PISA 2015	Linear
Albania	Secondary	2015	46.72	1.9	46.03	2.0	61.85	5.2	PISA 2015	Pseudo-Linear
Albania	Secondary	2015	46.72	1.9	46.73	2.0	56.85	5.2	PISA 2015	Equipercentile
Albania	Secondary	2015	46.72	1.9	46.92	1.9	56.76	5.2	PISA 2015	PS Equipercentile
Argentina	Primary	2013	92.33	0.6	95.83	0.6	61.42	1.9	TERCE 2013	Linear
Argentina	Primary	2013	92.33	0.6	97.10	0.5	67.68	1.8	TERCE 2013	Pseudo-Linear
Argentina	Primary	2013	92.33	0.6	95.83	0.6	64.71	1.9	TERCE 2013	Equipercentile
Argentina	Primary	2013	92.33	0.6	96.09	0.6	64.76	1.9	TERCE 2013	PS Equipercentile
Kenya	Primary	2007	88.77	1.0	88.77	1.0	41.60	6.8	SACMEQ III	Linear
Kenya	Primary	2007	88.77	1.0	88.77	1.0	29.60	6.7	SACMEQ III	Pseudo-Linear
Kenya	Primary	2007	88.77	1.0	88.77	1.0	47.97	6.8	SACMEQ III	Equipercentile
Kenya	Primary	2007	88.77	1.0	88.77	1.0	47.97	6.8	SACMEQ III	PS Equipercentile
Morocco	Secondary	2015	40.73	1.1	20.96	4.0	40.73	1.1	TIMSS 2015	Linear
Morocco	Secondary	2015	40.73	1.1	25.86	4.0	40.73	1.1	TIMSS 2015	Pseudo-Linear
Morocco	Secondary	2015	40.73	1.1	30.71	4.0	40.73	1.1	TIMSS 2015	Equipercentile
Morocco	Secondary	2015	40.73	1.1	30.96	4.0	40.73	1.1	TIMSS 2015	PS Equipercentile

Note: In all estimations, we use results for mathematics. Original MPL means the original value of the proportion of students reaching the MPL based on the original definition of the given assessment. MPL Benchmark 1 is the standardized benchmark based on either SACMEQ (for primary education) or PISA (for secondary education). MPL Benchmark 2 is the standardized benchmark based on TIMSS benchmark (Low International Benchmark). See text for more details about the linking strategies. PS = presmoothed.

Table 11. Robustness check: Comparability of results between national and regional assessments

Argentina						
	National assessment		Regional assessment			
	ONE		SERCE/TERCE			
			Original		Anchored	
Full sample	Maths	Reading	Maths	Reading	Maths	Reading
2006			86.58	80.85	57.52	74.10
2010	64.3	72.3				
2013	64.4	72.0	92.33	82.92	64.76	76.82
Annual variation	0.033	-0.100	0.821	0.296	1.034	0.389
Gender - 2013						
Male	65.4	70.1	91.78	80.22	67.78	74.29
Female	62.0	73.8	92.92	85.65	61.57	79.39
Ratio F/M	0.948	1.053	1.012	1.068	0.908	1.069
Burkina Faso						
	National assessment		Regional assessment			
	EAS		PASEC 2014			
			Original		Anchored	
Full sample	Maths	Reading	Maths	Reading	Maths	Reading
2012	68.49	68.87				
2014			58.82	56.94	n.a.	n.a.
Gender - 2013						
Male	71.66	70.02	61.92	58.01	n.a.	n.a.
Female	65.19	67.74	55.83	55.91	n.a.	n.a.
Ratio F/M	0.910	0.967	0.902	0.964	n.a.	n.a.

Figure 1. Proportion of students reaching the MPL, mathematics, primary education

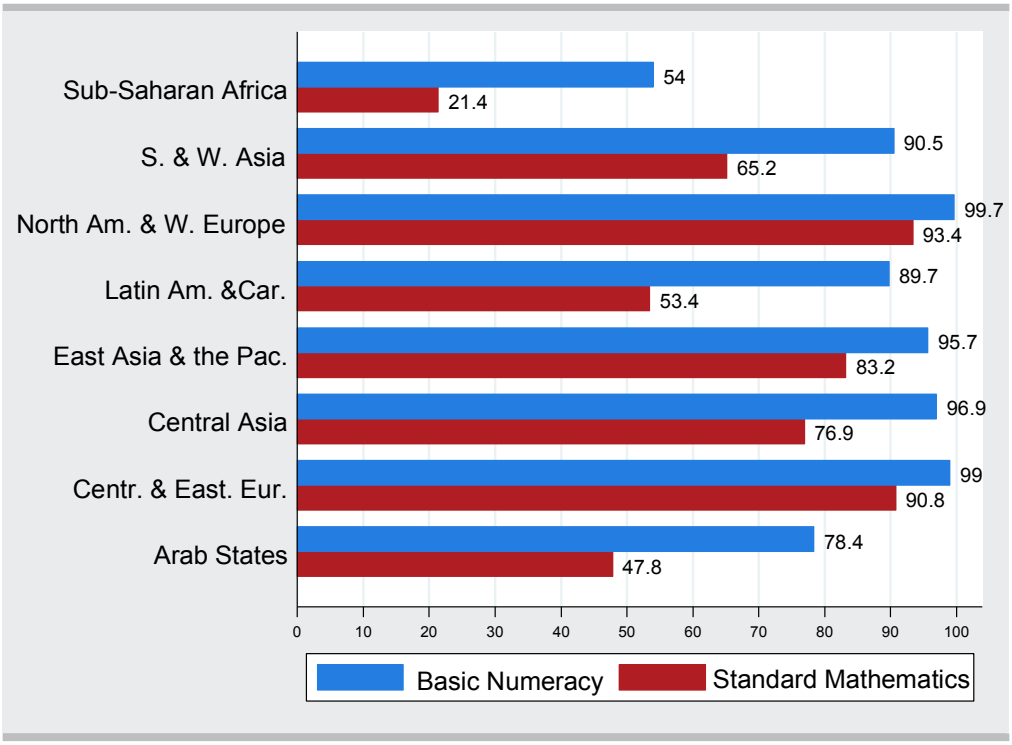


Figure 2. Proportion of students reaching the MPL, reading, primary education

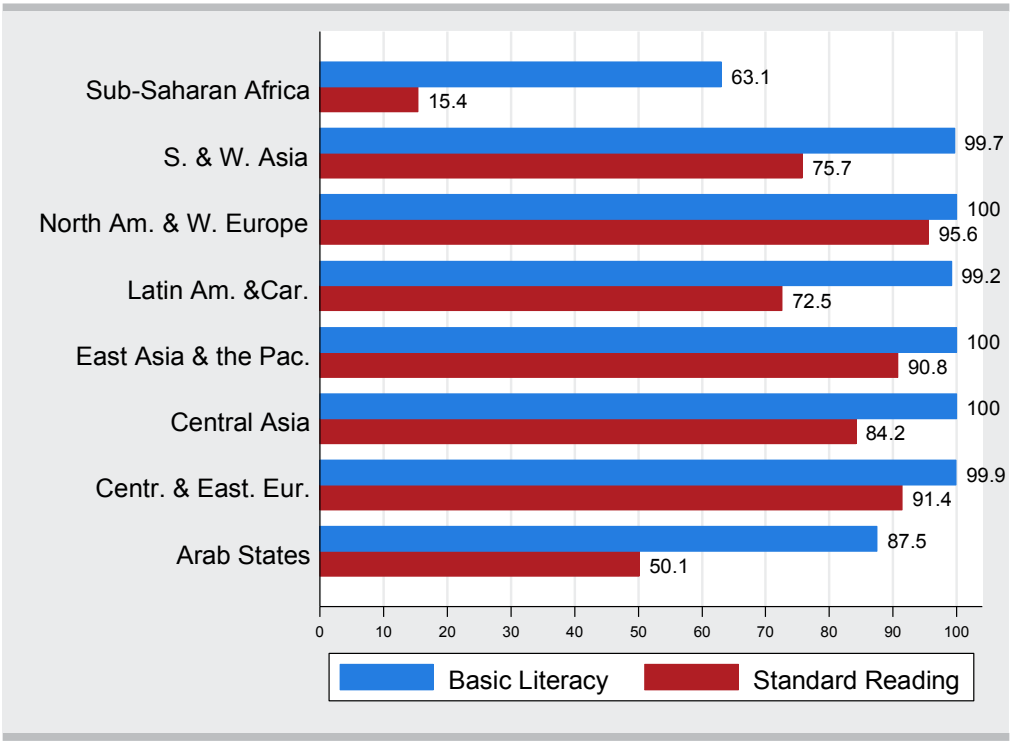


Figure 3. Proportion of students reaching the MPL, mathematics, primary education

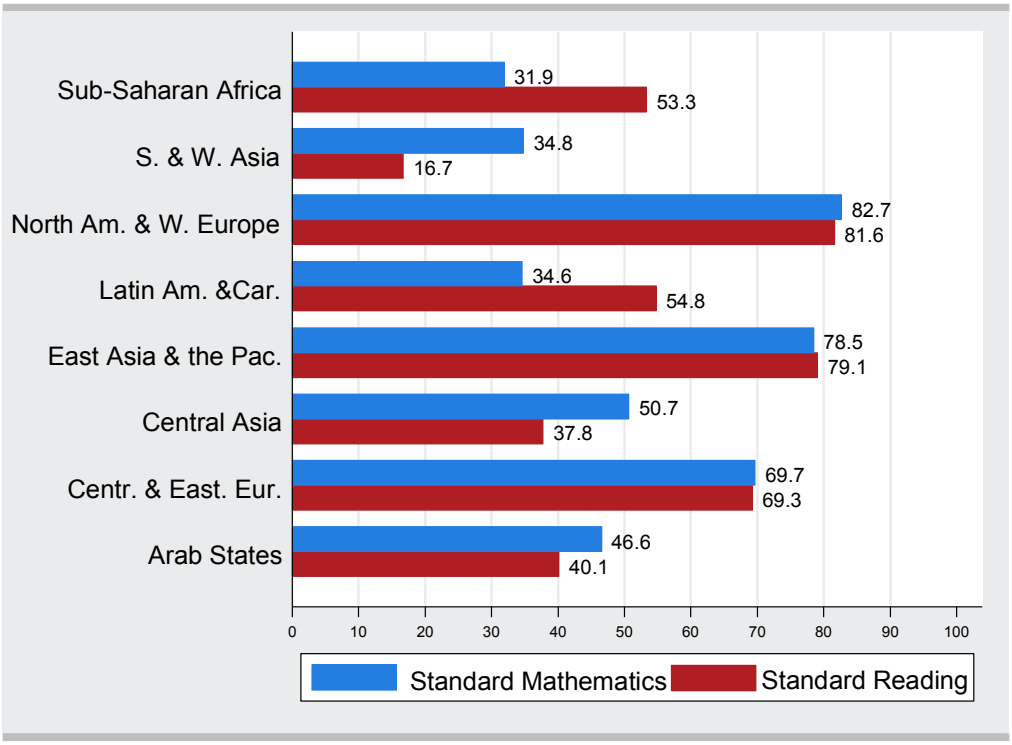


Figure 4. Gender Parity Ratio for the proportion of students reaching the MPL, primary education

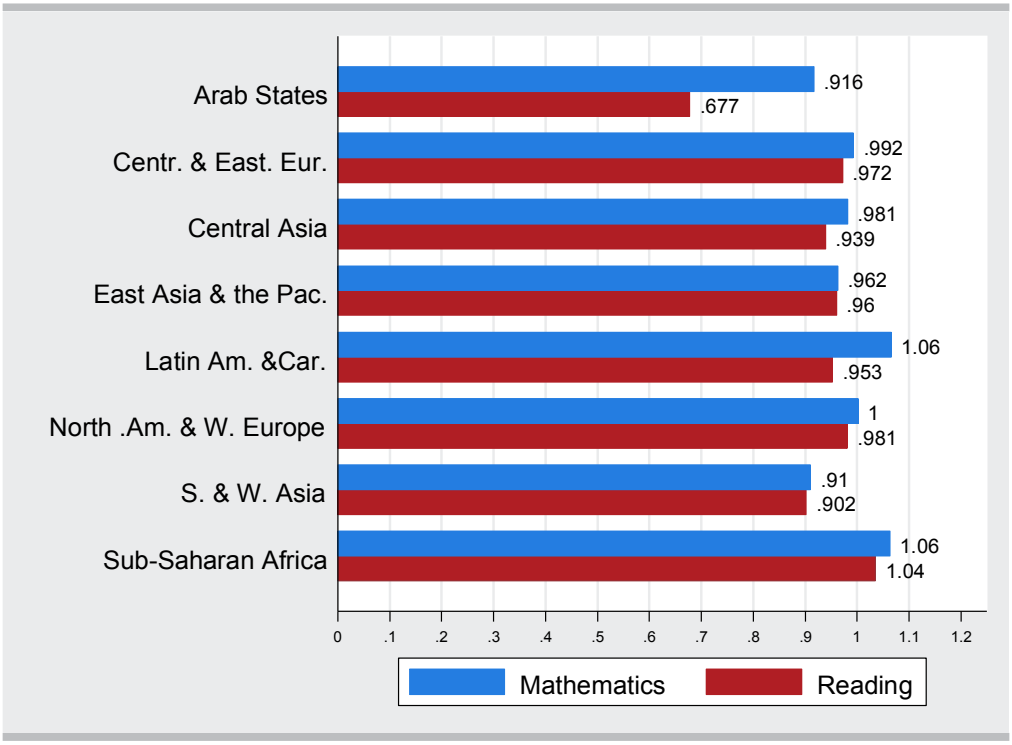


Figure 5. Proportion of students reaching the MPL, secondary education

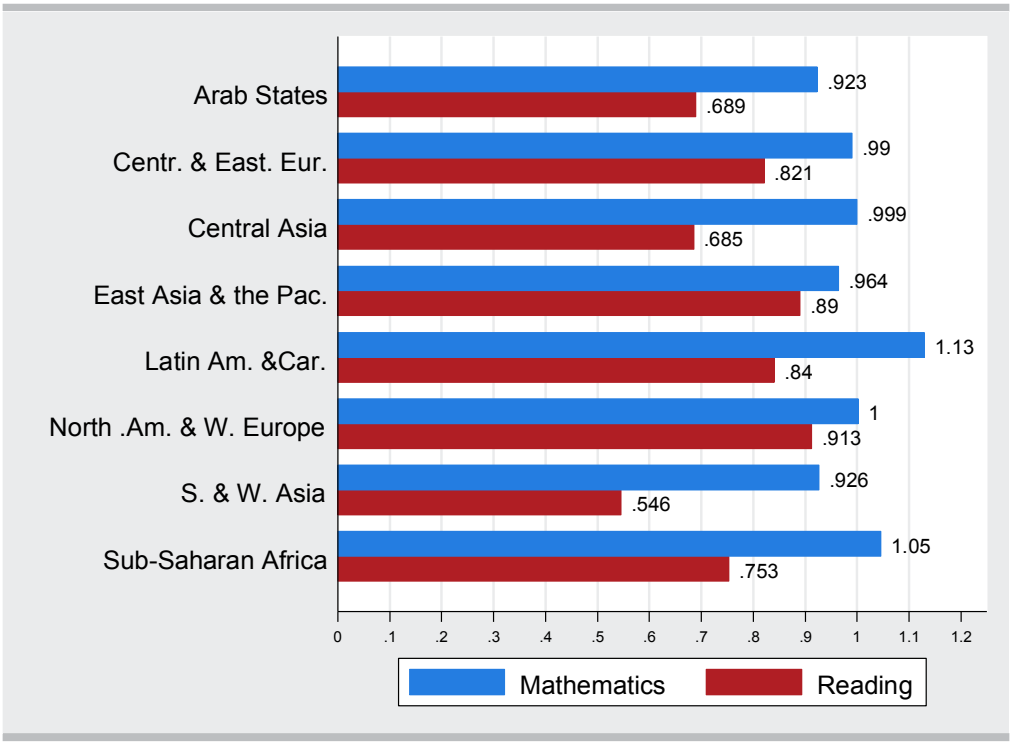


Figure 6. Residence Parity Ratio for the proportion of students reaching the MPL, primary education

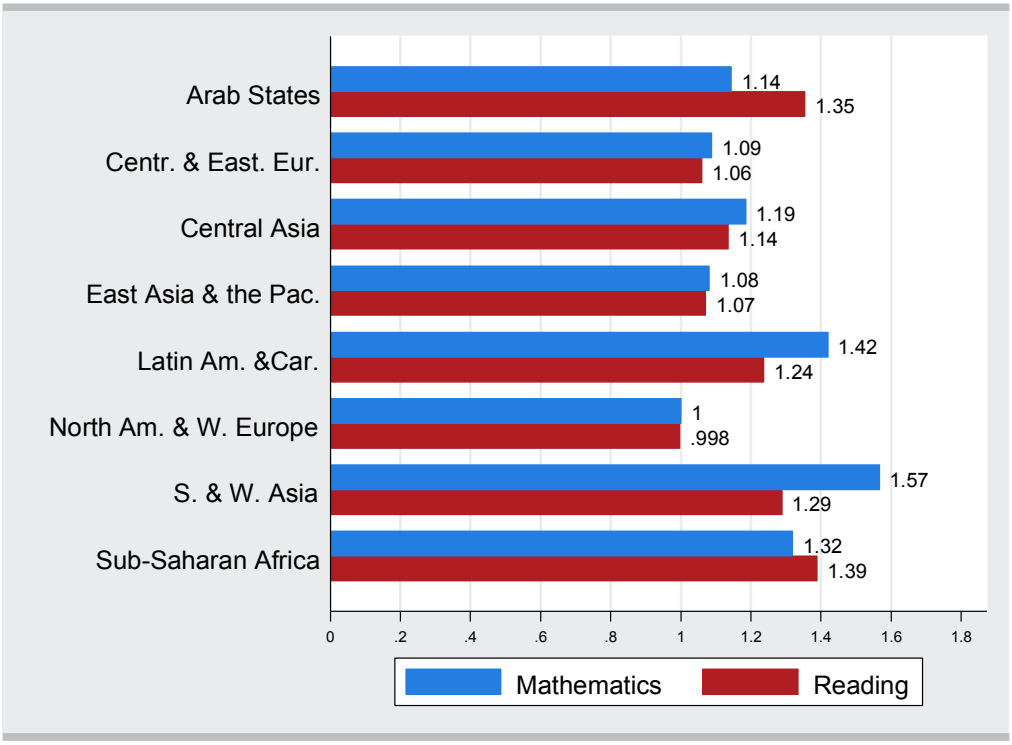


Figure 7. Residence Parity Ratio for the proportion of students reaching the MPL, secondary education

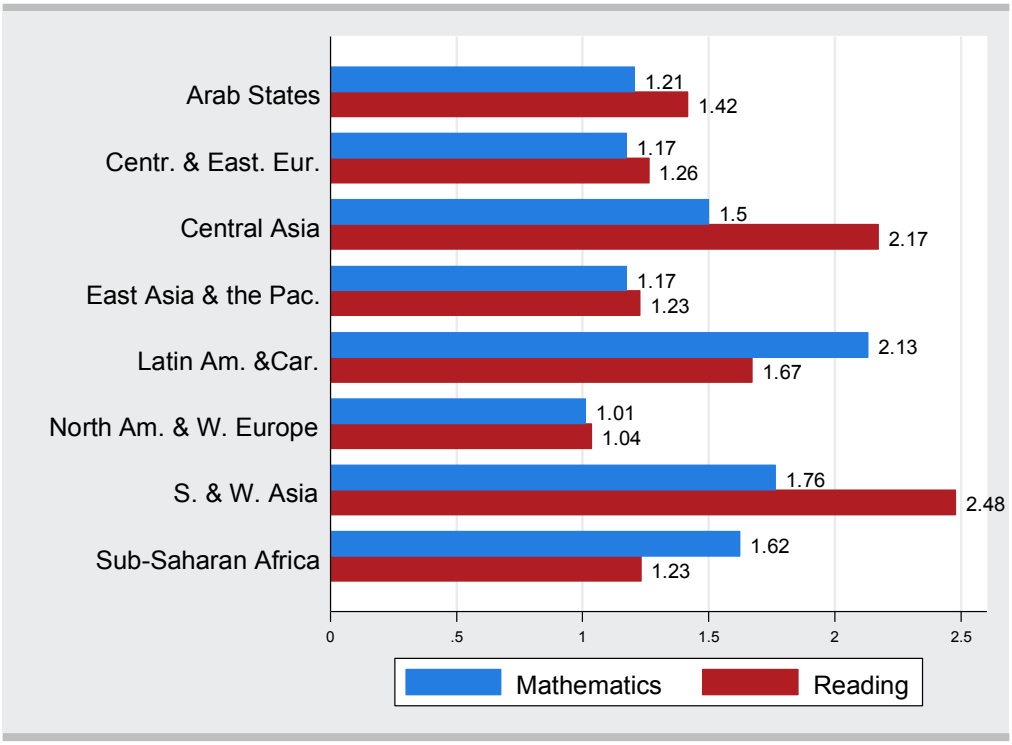


Figure 8. Socio-Economic Parity Ratio for the proportion of students reaching the MPL, primary education

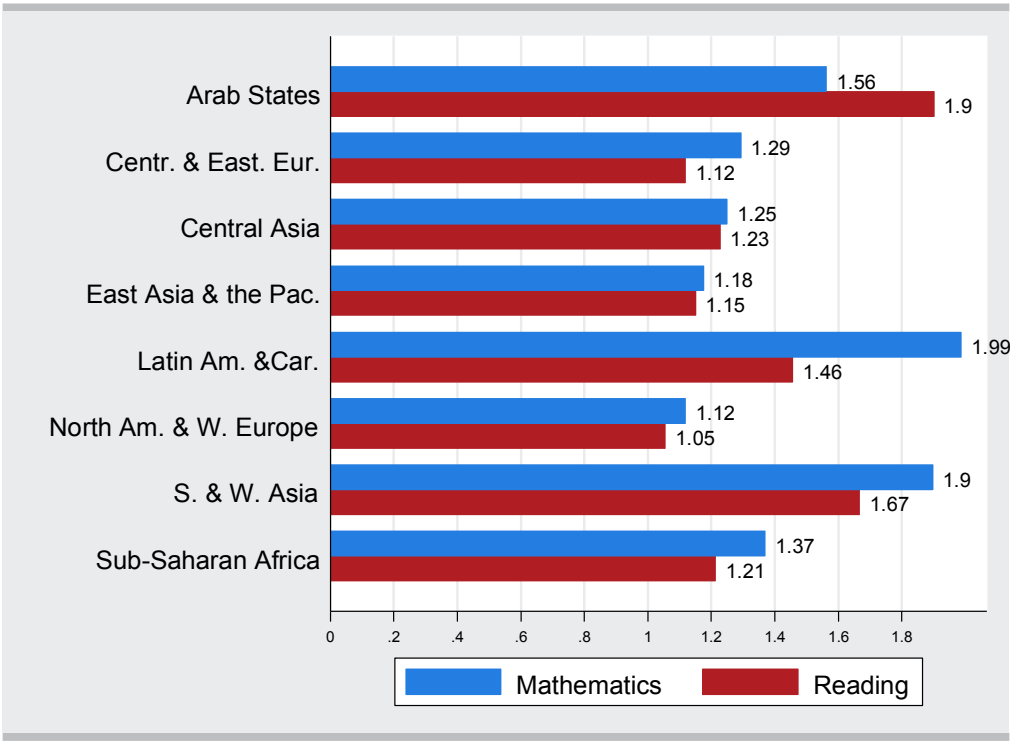


Figure 9. Socio-Economic Parity Ratio for the proportion of students reaching the MPL, lower secondary education

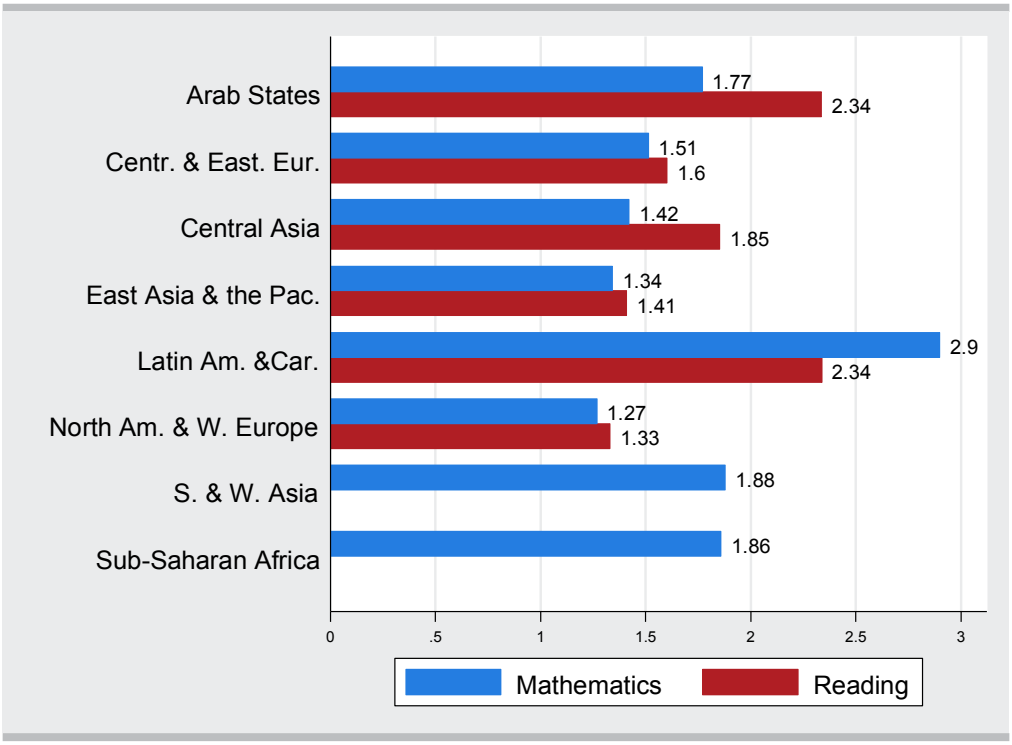


Figure 10. Trends in the proportion of students reaching the MPL "standard mathematics", primary education, selected countries

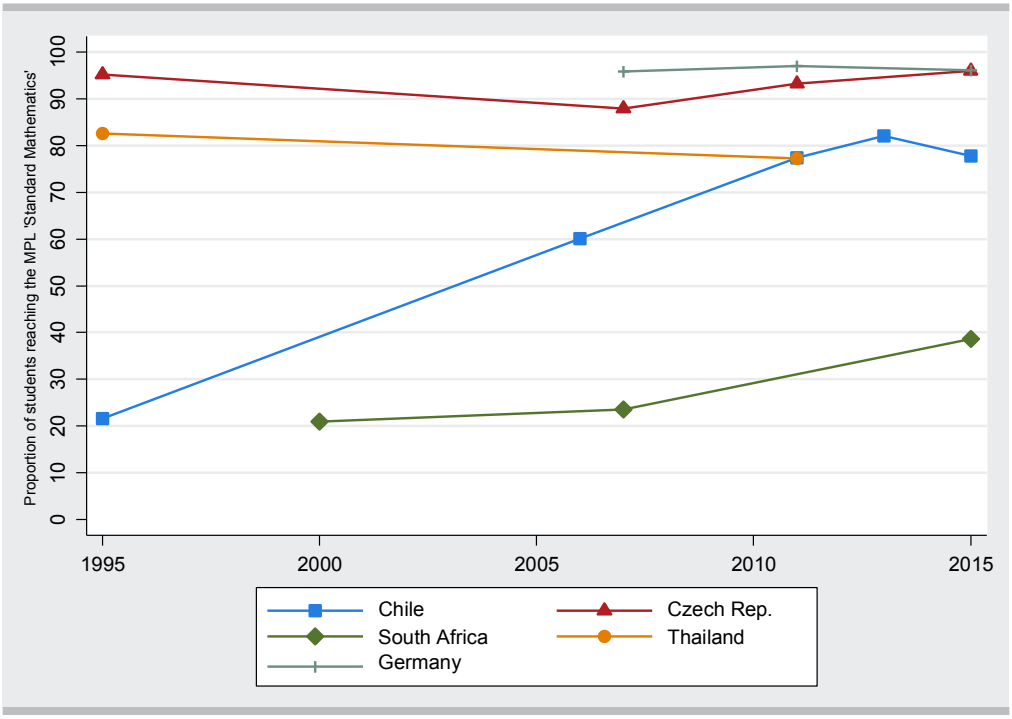
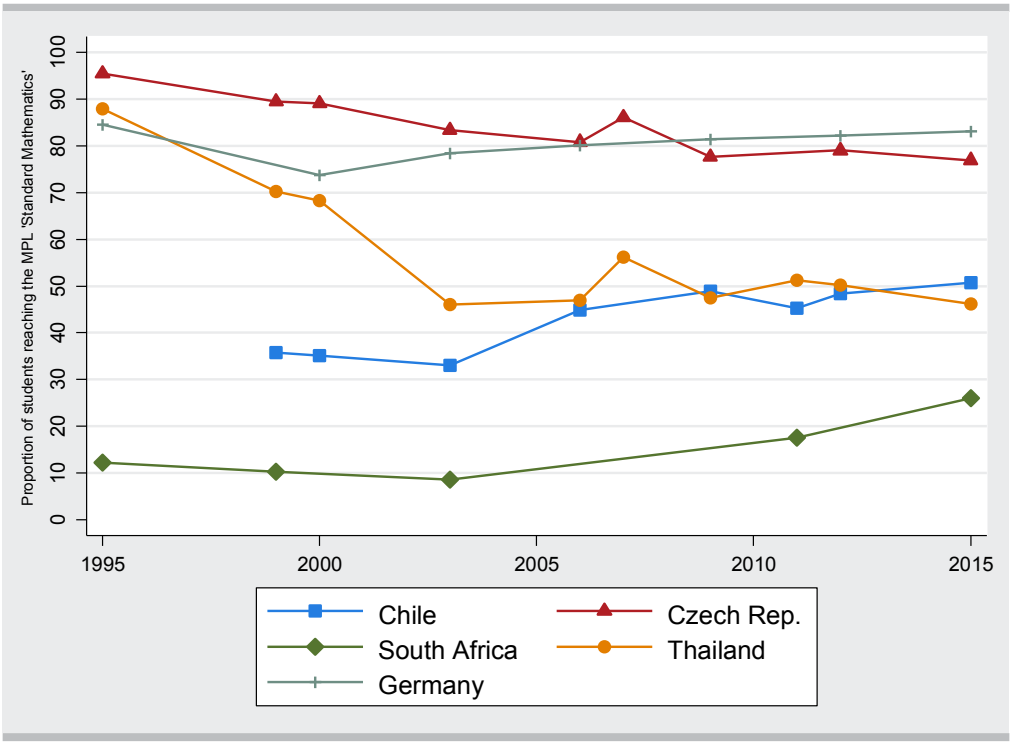
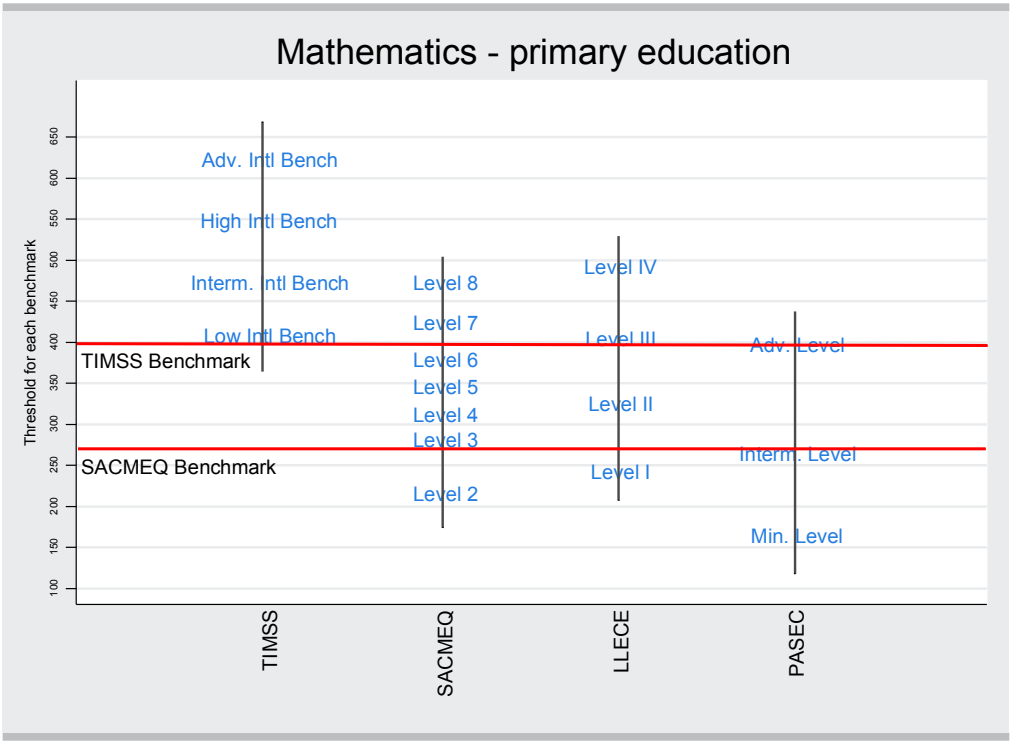


Figure 11. Trends in the proportion of students reaching the MPL "Standard Mathematics", secondary education, selected countries



Appendix A. (Not for publication)

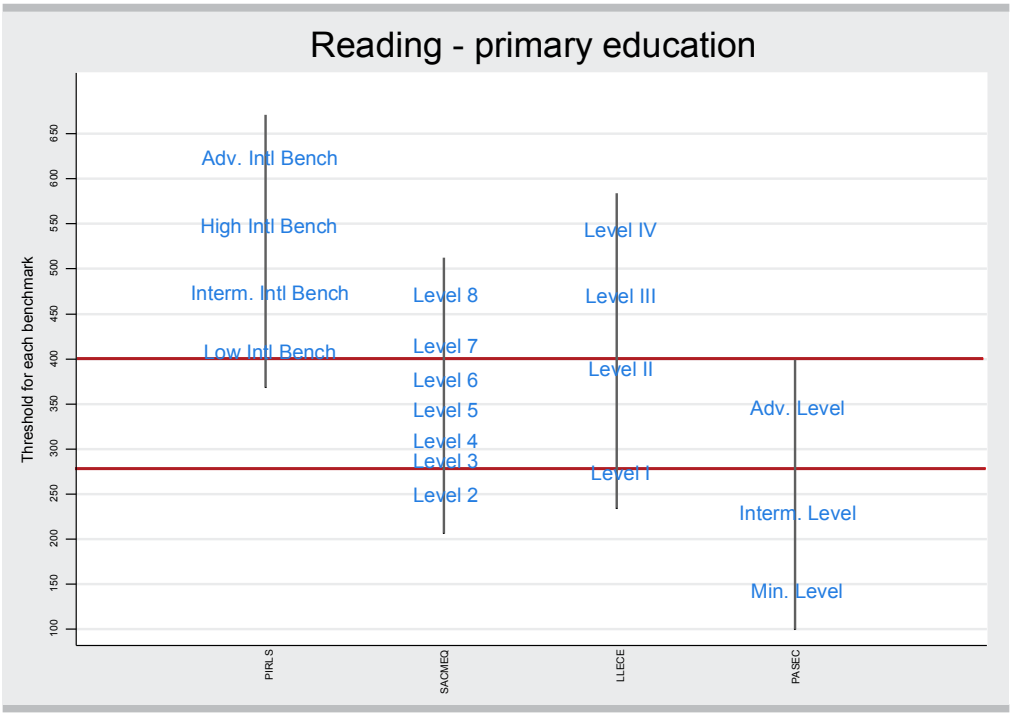
Figure A.1. Anchored benchmarks in primary education, mathematics



Note: This figure presents the adjusted scores for each benchmark from all assessments which provide results for mathematics in primary education (between grades 4 & 6). Results for PASEC are based on data prior to 2014. Red lines show the two options used for the choice of the minimum proficiency level.

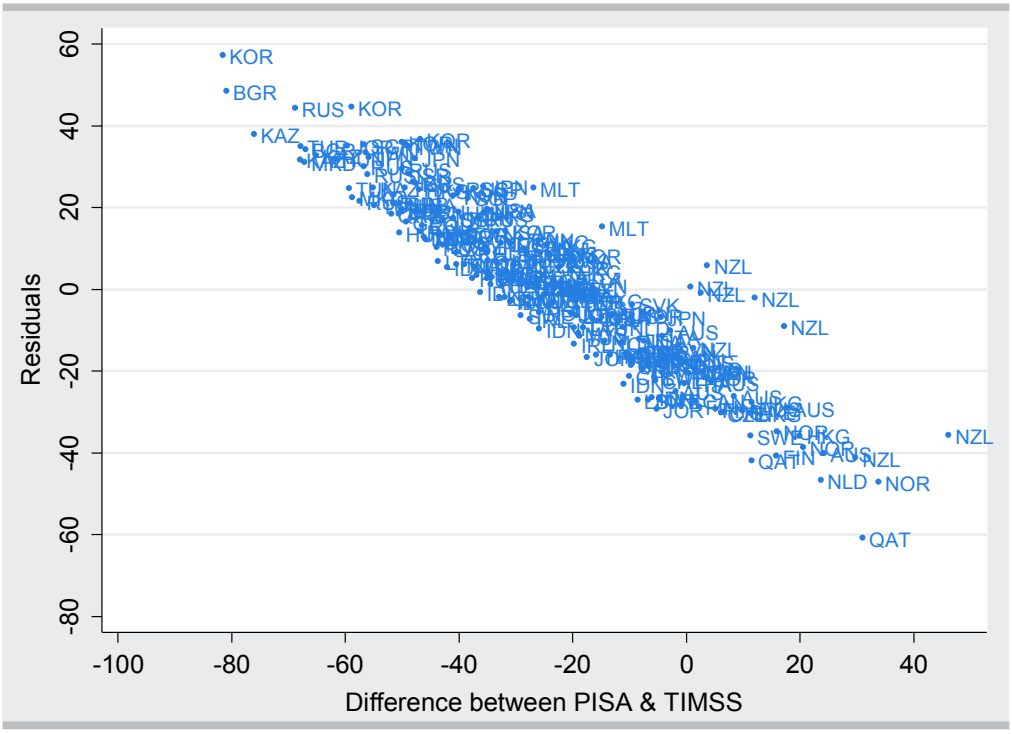
Source: All figures are from the author, based on original micro data on student achievement tests.

Figure A.2. Anchored benchmarks in primary education, reading



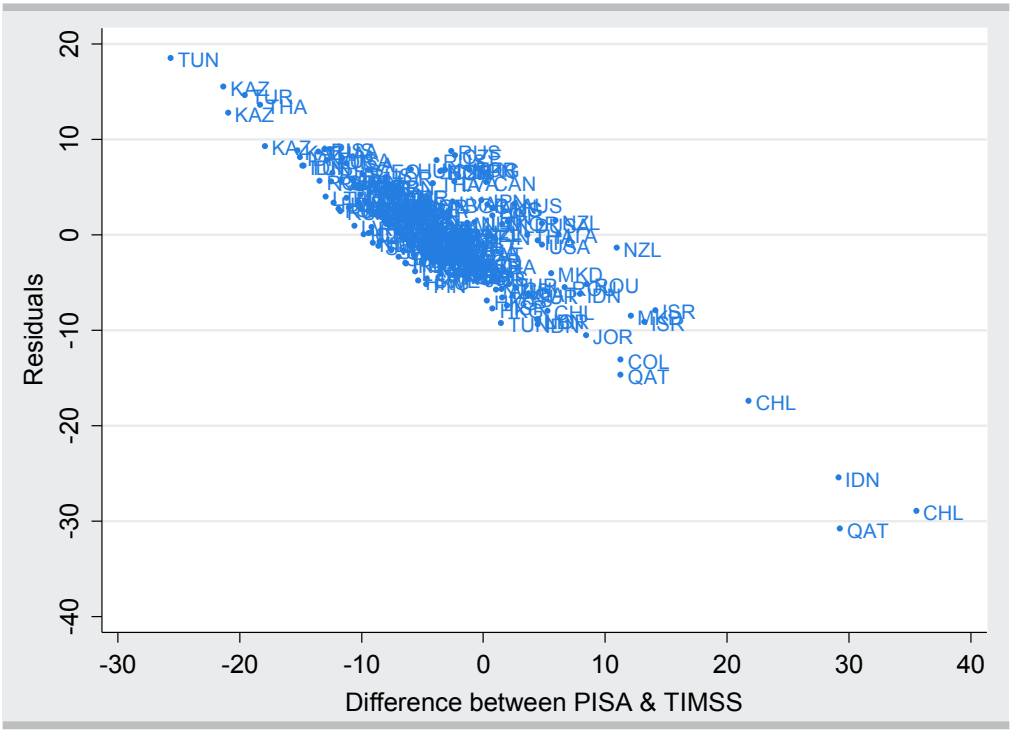
Note: This figure presents the adjusted scores for each benchmark from all assessments which provide results for reading in primary education (between grades 4 & 6). Results for PASEC are based on data prior to 2014. Red lines show the two options used for the choice of the minimum proficiency level.

Figure A.3. Comparison of original scores between PISA and TIMSS assessments



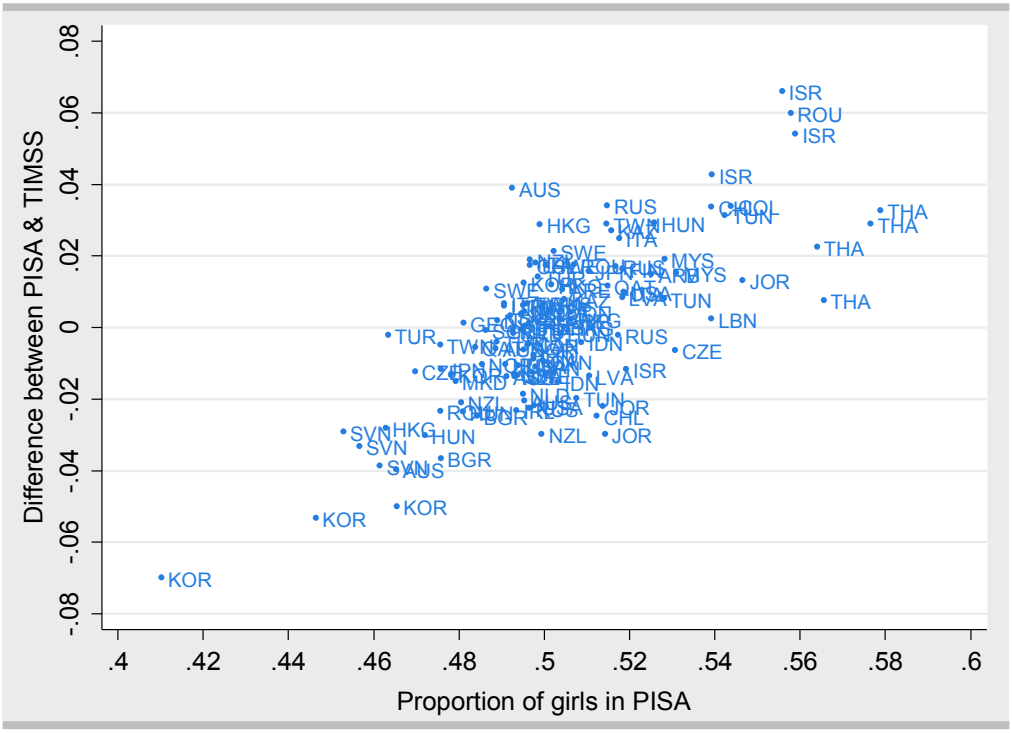
Note: Both mathematics and science skills are included. All years are taken into account (2000, 2003, 2006, 2011 and 2015). Since TIMSS was undertaken in 1999 instead of 2000, we directly compared TIMSS 1999 and PISA 2000 results. A similar comparison was made between TIMSS 2007 and PISA 2006. Results from PISA 2009 were not included. Residuals are obtained from Table 6, column (2).

Figure A.4. Comparison of original value of the proportion of students reaching the MPL between PISA and TIMSS assessments



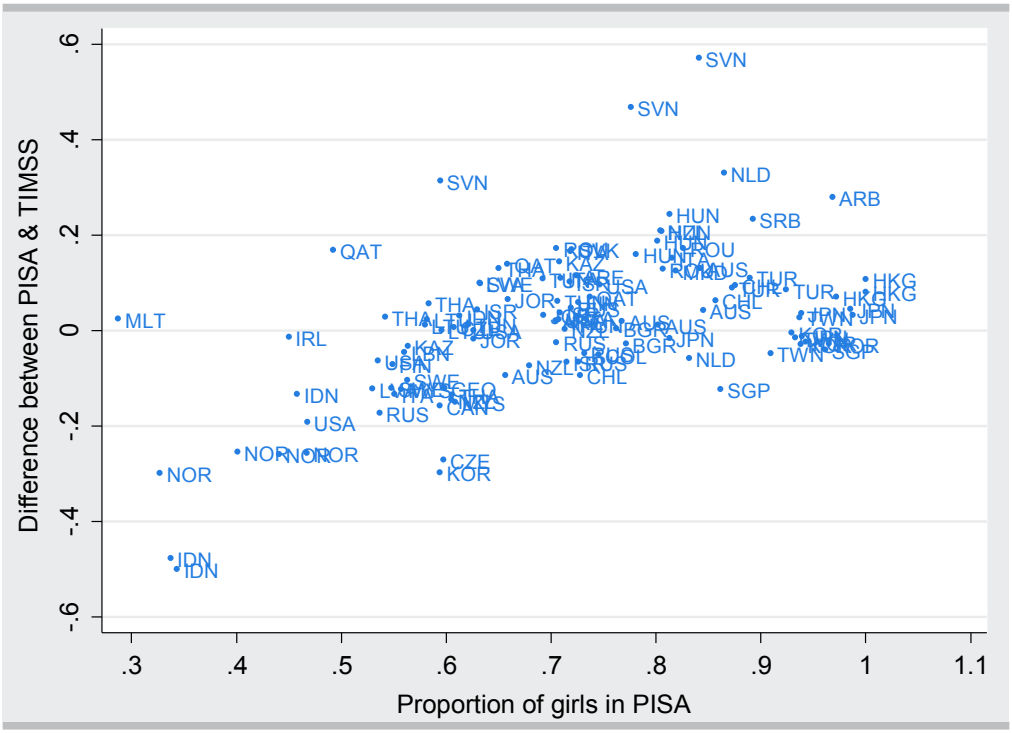
Note: A similar process of selection as Figure A.1 was made. The main difference is that the comparison is made between the proportion of low-performing students between TIMSS and PISA. Residuals are obtained from Table 6, column (4).

Figure A.5. Proportion of girls in PISA and TIMSS assessments



Note: Country abbreviations may appear several times due to the fact that we used five different rounds of PISA (2000, 2003, 2006, 2012 and 2015) and TIMSS (1999, 2003, 2007, 2011 and 2015).

Figure A.6. Proportion of students who live in urban areas in PISA and TIMSS assessments



Note: Country abbreviations may appear several times due to the fact that we used five different rounds of PISA (2000, 2003, 2006, 2012 and 2015) and TIMSS (1999, 2003, 2007, 2011 and 2015).

Figure A.7. Comparison of anchored value of proportion of students reaching the MPL for the two benchmarks

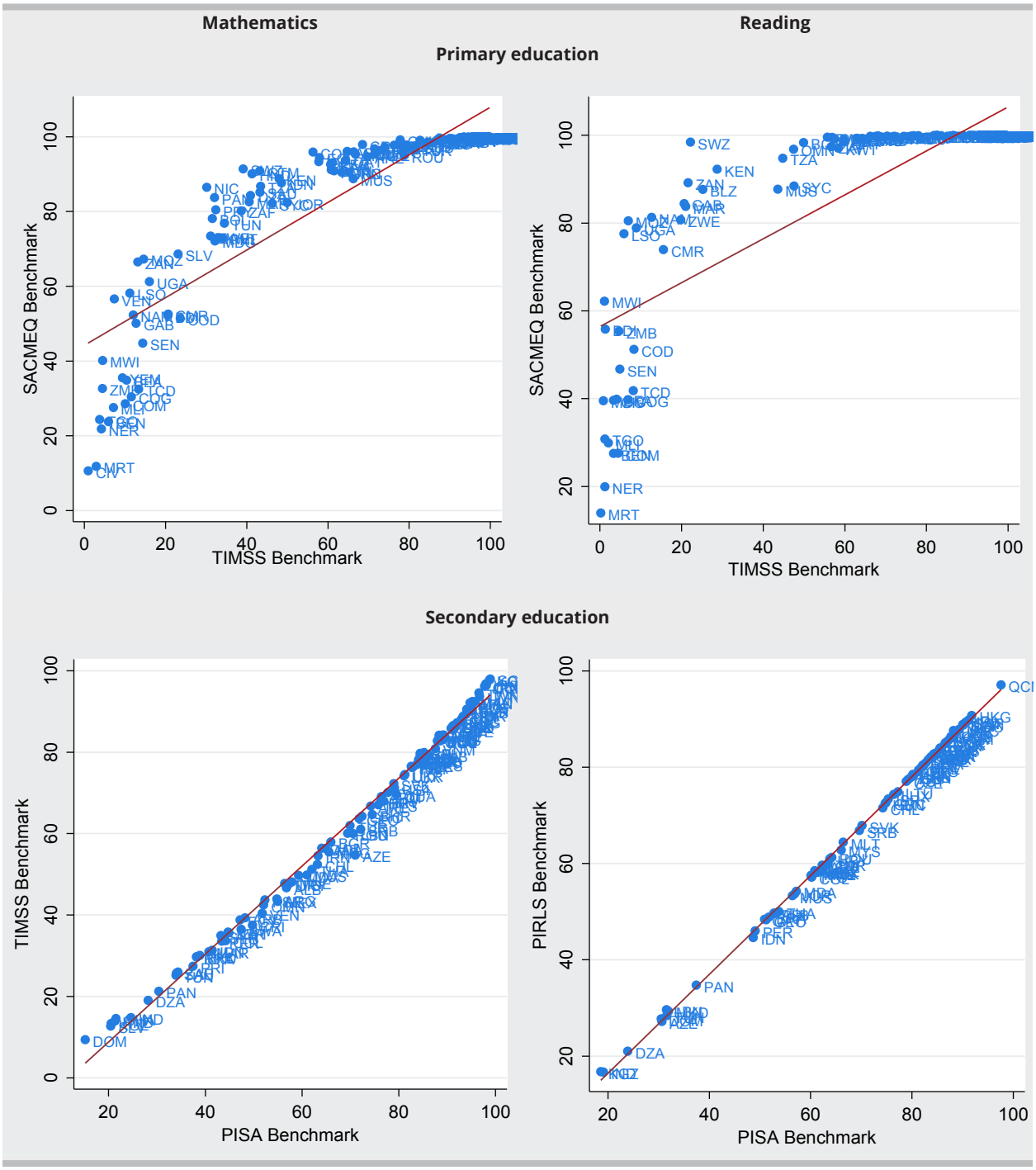


Figure A.8. Comparison of anchored value of the proportion of students reaching the MPL for the two benchmarks, sub-Saharan Africa

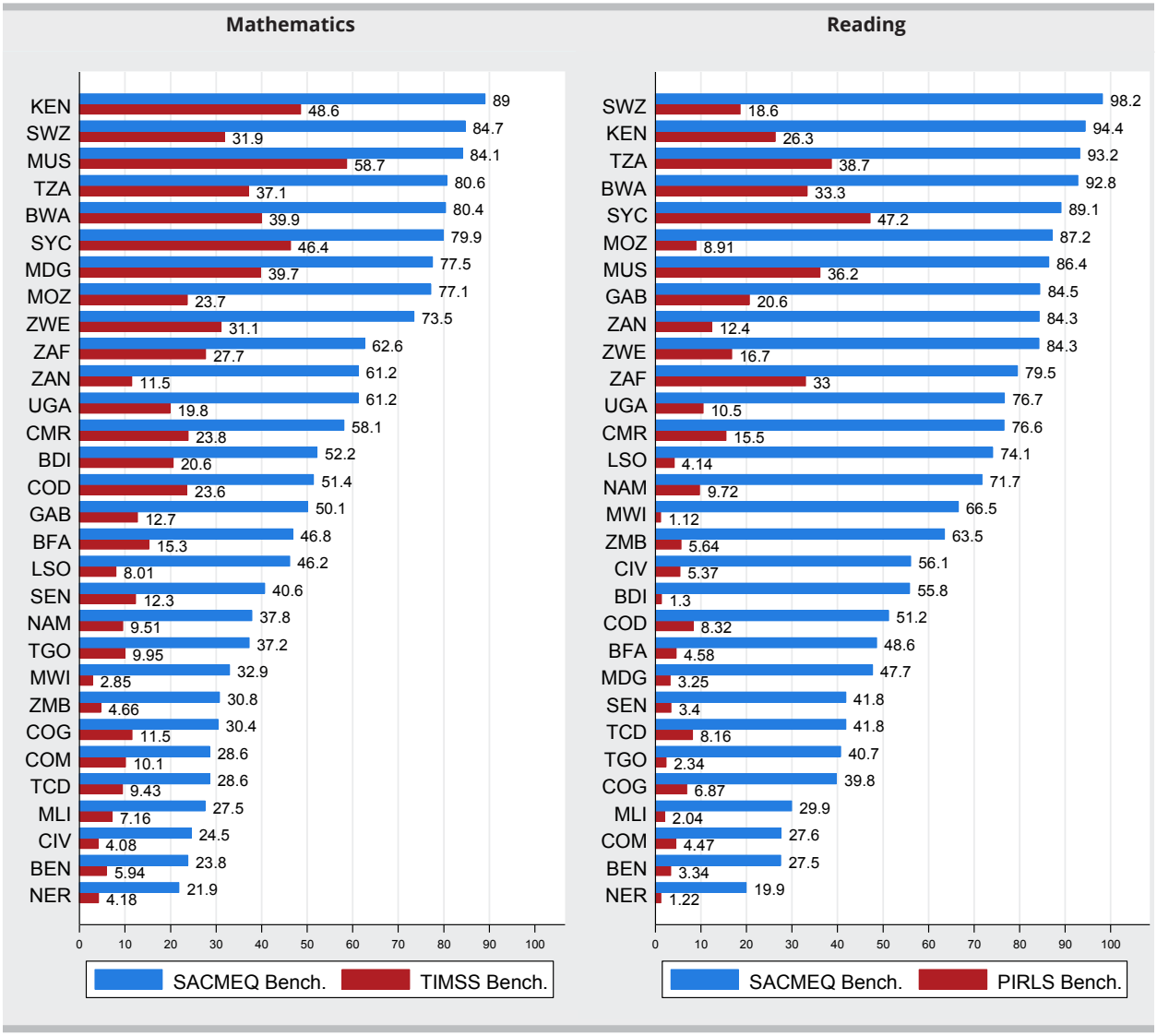


Figure A.9. Comparison of anchored value of proportions of students reaching the MPL for the two benchmarks, mathematics, primary education, Latin America and the Caribbean

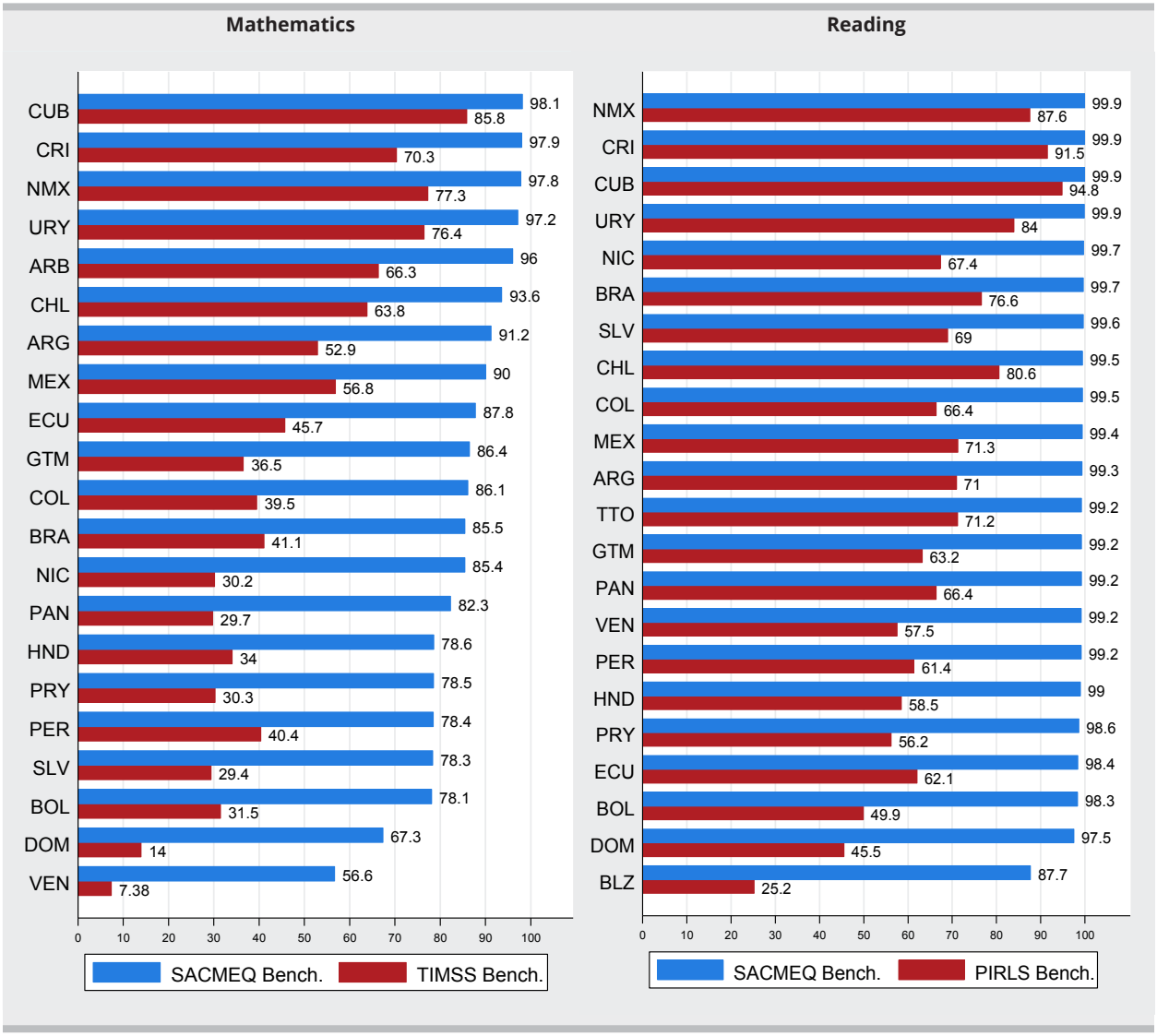
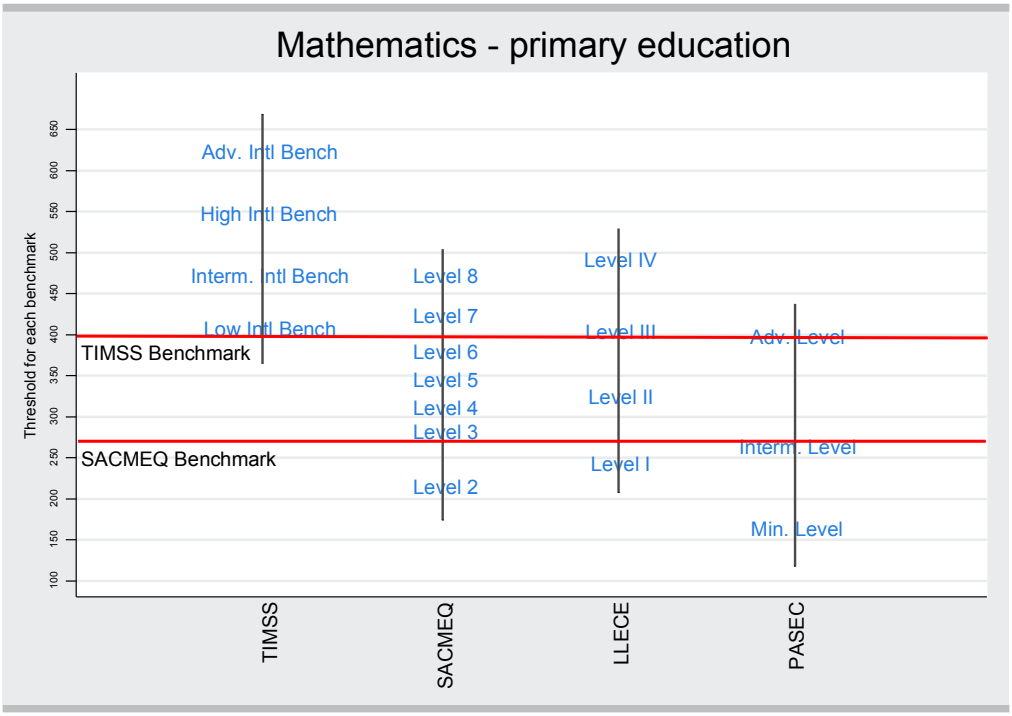
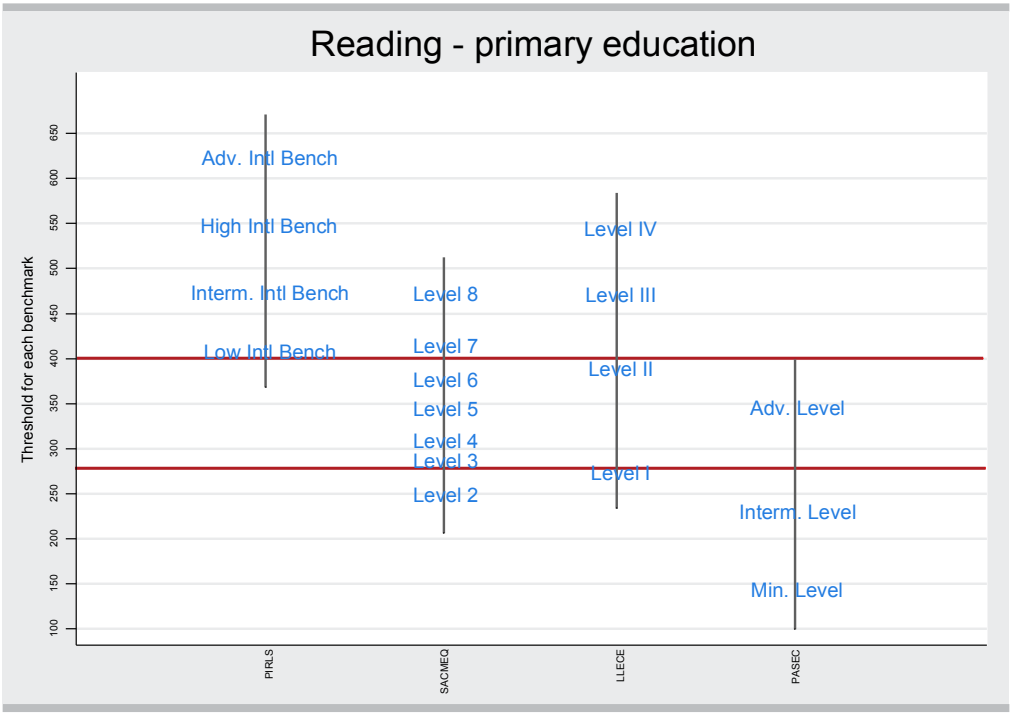


Figure A.10. Anchored benchmarks in primary education, mathematics



Note: This figure presents the adjusted scores for each benchmark from all assessments which provide results for mathematics in primary education (between grades 4 and 6). Results for PASEC are based on data prior to 2014. Red lines show the two options used for the choice of the MPL.

Figure A.11. Anchored benchmarks in primary education, reading



Note: This figure presents the adjusted scores for each benchmark from all assessments which provide results for reading in primary education (between grades 4 and 6). Results for PASEC are based on data prior to 2014. Red lines show the two options used for the choice of the MPL.

Figure A.12. Gender parity index for the two benchmarks (SACMEQ and TIMSS), mathematics, primary education