unesco
Institute for Statistics

confemen
pasec
Programme d'analyse des systèmes éducatifs de la confemen

Laboratorio Latinoamericano Evaluación Calidad Educación

IEA
TIMSS & PIRLS International Study Center
Lynch School of Education
BOSTON COLLEGE

Rosetta Stone Policy Brief

Establishing a concordance between regional (ERCE and PASEC) and international (TIMSS/PIRLS) assessments

4 QUALITY EDUCATION

IEA
ROSETTA STONE
BOSTON COLLEGE

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 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 Inst tute for Stat st cs
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.

Published in 2022 by:
UNESCO Institute for Statistics
C.P 250 Succursale H
Montréal, Québec H3G 2K8
Canada

Email: uis.publications@unesco.org
http://www.uis.unesco.org
Ref: UIS/2022/LO/TD/10
ISBN: 978-92-9189-298-3

Layout by: FreireCreative
Cover design by: büro Svenja

Acknowledgements
The Rosetta Stone Analysis Report was a UNESCO Institute for Statistics (UIS) collaborative project. The International Association for the Evaluation of Educational Achievement (IEA) was the technical partner for this project along with the TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College. Technical and implementation support was provided by CONFEMEN and LLECE.

Andres Sandoval Hernandez is the author of this report. Luis Crouch (RTI), Silvia Montoya (UIS), Dirk Hastedt (IEA) and Oliver Neuschmidt (IEA) served as reviewers.

This report is based on research funded by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation.

This report contains excerpts from the Rosetta Stone technical reports provided by the TIMSS & PIRLS International Study Center on developing the TIMSS & PIRLS Concordances for ERCE and PASEC. The technical reports are available at https://timssandpirls.bc.edu/Rosetta-Stone-Reports/index.html and further questions on these should be addressed to TIMSS & PIRLS International Study Center: https://timssandpirls.bc.edu/

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION[1]

With less than ten years left to achieve the 17 Sustainable Development Goals (SDGs) established in the Agenda 2030, our global leaders have declared a Decade of Action and delivery for sustainable development. The Decade of Action calls for accelerating sustainable solutions to all the world's biggest challenges – ranging from poverty and gender to climate change, inequality and improving the quality of education for all.

This document focuses on SDG 4, in particular Indicator 4.1.1. SDG 4 establishes that by 2030 we must "ensure inclusive and equitable quality education and promote lifelong learning opportunities for all", and Indicator 4.1.1 operationalizes this goal as the demand to "ensure that all girls and boys complete free, equitable, and quality primary and secondary education leading to relevant and effective learning outcomes".[2] The international community agreed upon measuring the progress of this target as the percentage of children and youth achieving at least a minimal level of competency in literacy and numeracy at three points in the grade structure, and by sex: (a) in Grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary.

There is a wide variety of assessments at national, regional and international levels that produce information about the level of competency in literacy and numeracy for children and young people. These international large-scale assessments (ILSAs) in education have been considered by many the most suitable source of information to measure and monitor progress towards several of the SDG 4 indicators. ILSAs offer systematic information from more than 100 education systems, provide information about the level of literacy and numeracy of children and youth and have unrivalled data quality assurance mechanisms. Moreover, ILSAs produce data that is strictly comparable across countries and over time. However, each of these assessments has a different assessment framework, is measured on a different scale and is designed to inform decision-making in different educational contexts.

Therefore, while information is already available, it must be integrated into a global assessment approach. Such an approach would allow comparisons between education systems not only within a geographical region but globally and over time. This would make it possible to learn from the policy initiatives of countries with similar characteristics (e.g. level of economic development, structure of the education system) but located in different parts of the world. While similarities between education systems are important to enable fair and valid comparisons, contextual, historical and cultural differences allow the possibility of

1 This policy brief contains excerpts from the Rosetta Stone technical reports provided by the TIMSS & PIRLS International Study Center on developing the TIMSS & PIRLS Concordances for ERCE and PASEC. The reports are available at https://timssandpirls.bc.edu/Rosetta-Stone-Reports/index.html, and further questions should be addressed to timss@bc.edu.

2 See https://en.unesco.org/education2030-sdg4/targets

looking at the same problem from a variety of angles and, therefore, the development of novel approaches and solutions. At the same time, regional and national assessments, by focusing on a test created for a specific context, facilitate the development of more meticulous comparisons that, in turn, can inform policies in a more detailed manner. In this way, combining national, regional and global assessments has the potential to boost the depth and breadth of our efforts to measure and monitor SDG 4.1.1.

For these reasons, the UNESCO Institute for Statistics (UIS) has spearheaded a methodological programme that aims at producing a robust measurement strategy offering comparable information for as many countries as possible to measure and monitor progress toward SDG 4.1.1. This programme, a pioneer of its type, involves two regional and two global assessments, linked in a manner that does not require re-designing the regional assessments in any way, providing a potential retrofit without top-down implications. This is key to the spirit UNESCO and other agencies have tried to maintain throughout the SDG 4 process in not forcing the world to adopt one assessment.

This programme, led by the International Association for the Evaluation of Educational Achievement (IEA) and the TIMSS & PIRLS International Study Center, Lynch School of Education at Boston College, is named after the famous archaeological discovery that enabled translation between different written languages: the Rosetta Stone. The Rosetta Stone Study is designed to measure global progress towards SDG 4.1.1 by relating different national and regional assessment programmes to Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) international long-standing metrics and benchmarks of achievement3. The goal is to provide countries that participated in regional or national assessments but not in TIMSS and PIRLS with information about the proportions of primary school students who have achieved a minimal level of competency in literacy and numeracy (SDG 4.1.1) that allows international comparisons.

This document presents an executive summary of the first results of the Rosetta Stone Study, consisting of the establishment of a concordance table that projects the score distributions estimated from two regional assessments to distributions on TIMSS and PIRLS. These assessments are UNESCO's Regional Comparative and Explanatory Study (ERCE; Estudio Regional Comparativo y Explicativo) in Latin America and Caribbean countries and the Programme for the Analysis of Education Systems (PASEC; Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN) in francophone sub-Saharan African countries.

The remainder of this document is divided into three main sections. The first section describes the methods and analyses conducted to establish the concordance between the regional assessments and TIMSS and PIRLS. The resulting concordance tables and guidelines to use and interpret them are presented in the second section. The third section shows the proportion of students who achieve SDG 4.1.1 in each of the ERCE and PASEC countries according to the concordance tables produced by Rosetta Stone. The last section offers a set of key messages for policymakers, practitioners and academics.

3 See IEA, 2017, in Further Reading.

METHODS

Despite having different frameworks and approaches, ILSAs such as ERCE, PASEC, TIMSS and PIRLS share a largely similar set of procedures and methods. For example, ILSAs use complex sampling (stratified, multistage, clustered sample design) and a complex assessment (rotated block design). As a result, the process of producing a concordance between them is also a complex one. This section begins with a brief description of the procedures adopted by ILSAs, followed by an overview of the methods used to produce the concordance tables. Please note that this document offers a simplified version of both; it is strongly recommended to consult the technical reports of the corresponding ILSAs and the analysis reports of the Rosetta Stone Study for the full technical details (see list of readings at the end of this document).

A. ILSAs' complex design

The objective of most ILSAs is to produce nationally representative information about the educational achievement of a target student population in a given subject. To do this, ILSAs face two main practical challenges. First, a test cannot be administered to all the students in an education system and second, one single student cannot answer the full test. Testing all students would be very costly and would create many logistical complications, and those students who are tested do not answer all the items because it would take them between eight and ten hours to complete a test covering all the topics included in the study framework. Thus, only a sample of students answer the test, and each of the sampled students answers only a portion of the test. These challenges are overcome by using a complex sample and assessment design.

ILSAs' complex sampling design is typically characterised as multistage, stratified cluster sampling. Stratification consists of arranging the lists of schools in groups (i.e. strata) by some of the characteristics they share, like location in geographic regions or school types (e.g. public, private). The first sampling stage involves selecting schools within these strata. The list of schools is also ordered according to the schools' size, such that larger schools (i.e. those with more students) have a greater probability of being selected (probability proportional to size). The second stage involves sampling either students or classes (i.e. clusters) within the sampled schools. As the goal is to provide estimates for the entire target population, these unequal probabilities must be adjusted using sampling and replicate weights (other methods, such as multilevel models, can also be used in combination with the sampling weights).

A complex assessment design is necessary for ILSAs because a large number of test items is needed to assess student proficiency in a broad content domain with high reliability. To accommodate many items in a limited testing time, a rotated block design is used. In other words, items are grouped in blocks by the content and cognitive domains they measure. The blocks are then distributed across a number

of tests so that each test contains two or more blocks but no test contains all the items. The test is designed such that different tests are linked through common blocks of items. As such, no student takes all items and many students take different items. However, a comparable achievement score can still be estimated thanks to a scaling method known as item response theory (IRT). Through IRT, the ability of each respondent and the difficulty of each item are estimated. This allows for a score to be estimated that is comparable across students even when they answered different test items. Then, to estimate the achievement scores, a population model is used. This model takes the estimates from the IRT model and the students' available background information to estimate the score that each student would have received if he or she had taken all the items of the test. Because this score is expressed as a probability distribution, several random draws (usually five) are taken from it. Each random draw, also known as plausible value (PV), is considered a representative value from the distribution of potential scores for all students in the sample who have similar background characteristics and similar patterns of item responses (e.g. those who tend to respond correctly to items of the same difficulty).

Both the complex sampling and the complex assessment design result in some uncertainty or measurement error that has to be considered in all calculations. It is important to note that no assessment design completely eliminates measurement error. In practice, it is a matter of optimizing the uncertainty caused by the measurement error relative to cost and practical considerations. This uncertainty is accounted for in all analyses through the use of sampling and replicate weights (for the complex sample design) and through the use of IRT and PV methods (for the complex assessment design).

B. Rosetta Stone methods

This section describes the instruments and design of the Rosetta Stone linking study and the analytical procedure followed for the construction of the concordance tables. The Rosetta Stone Study, like most ILSAs, uses a complex assessment design in which each student is administered only a subset of the items in the pool. The Rosetta Stone Study is composed of two assessment parts. The first part is the source assessment (i.e. ERCE and PASEC), including achievement items and the context questionnaire. The second part is the centrepiece of the study, the Rosetta Stone (target) assessment, consisting of test booklets with item blocks from TIMSS and PIRLS. In total, eight rotating mathematics item blocks (TIMSS) and four rotating literacy passages (PIRLS) were used. Both assessment parts were administered as paper-based assessments to the same students. Each student was administered the booklets from the source assessment (PASEC or ERCE) on the first day and one Rosetta Stone booklet on the second day.

To establish concordance tables, the analysis of the data was carried out in four steps. First, data quality was evaluated based on descriptive statistics and nonresponse variability. Second, IRT models were used to construct comparable ERCE, PASEC and Rosetta Stone (i.e. TIMSS and PIRLS) scales across student

populations. Third, PV methodology was used separately for ERCE, PASEC and Rosetta Stone to produce mathematics and literacy scores. Fourth, concordance tables were established based on the average mathematics and literacy scores of the source (ERCE and PASEC) and target assessments (Rosetta Stone – TIMSS and PIRLS). The analysis was performed on data from three PASEC countries (Burundi, Guinea and Senegal) and two ERCE countries (Colombia and Guatemala) using sample weights provided by the ERCE and PASEC teams to the TIMSS & PIRLS International Study Center.

i. Data quality evaluation

Data quality was evaluated using classical item statistics. Item difficulty was evaluated by estimating the percentage of students who answered each item correctly, and the item discrimination by estimating the point-biserial correlation between success in the item and total score. The results of this evaluation confirmed that the data collected through the Rosetta Stone Study had adequate properties to proceed to the following steps of the analysis. Item difficulty was generally within an acceptable range (25% to 95% correct) and was similar for ERCE, PASEC, TIMSS and PIRLS items. The discrimination values were also similar across all studies and were within acceptable ranges (correlation coefficients no lower than .20), with typical values between .24 and .41 for PASEC and between .36 and .50 for ERCE.

ii. Constructing comparable scales for ERCE, PASEC and Rosetta Stone (i.e. TIMSS and PIRLS)

Unidimensional IRT models were used to scale the Rosetta Stone linking items (i.e. TIMSS and PIRLS items) as well as the PASEC and ERCE items (separately for numeracy and literacy). The unidimensional IRT models showed high levels of comparability (i.e. high proportions of items with good fit) across ERCE countries and the Rosetta Stone scales (with values ranging from 91.2% to 95.1%), as well as across PASEC countries and the Rosetta Stone scales (with values ranging from 81.2% to 86.3%). The models also showed high levels of comparability across countries for ERCE and PASEC scales, providing a solid basis for establishing a concordance (with values ranging from 91.5% to 98.7%).

After PASEC, ERCE and Rosetta Stone items were scaled with separate unidimensional IRT models, multidimensional IRT models were utilized to examine how similar or different the measured constructs of the different assessments are. More precisely, these multidimensional models were used to investigate the relationship between the ERCE and PASEC mathematics and TIMSS numeracy scales, and between the ERCE and PASEC reading and PIRLS literacy scales. The item parameters in these multidimensional models were fixed to the item parameter values obtained from the set of previously described unidimensional models.

The results of the multidimensional IRT models indicated that the corresponding Rosetta Stone and PASEC, and Rosetta Stone and ERCE scales measure constructs that are not identi-

cal (as expected since they were developed based on different frameworks) but that are similar enough (with latent correlations between dimensions higher than .80 and lower than .90) to enable a meaningful concordance for a projection of score distributions. Latent correlation ranged from .82 and .90 for ERCE and from .81 to .83 for PASEC.

iii. Producing mathematics and reading scores

Once the IRT scores were estimated, a population model was used to estimate the distribution of proficiencies or PVs. This population model uses the estimates from the IRT model and the students' available background information to estimate a distribution of possible scores that each student would have received if he or she had taken the whole test. The final goal of population modelling is to produce distributions of proficiencies from which PVs can be drawn. Following the procedures used in TIMSS and PIRLS, five PVs were drawn from the proficiency distribution for each domain and each student.

iv. Establishing concordance

The final step was to construct the Rosetta Stone concordance tables, which establish a relationship between scores on the source and target assessments. In Rosetta Stone, a range of plausible TIMSS and PIRLS scores is projected from ERCE and PASEC mathematics and reading scores, respectively. That is, ERCE mathematics, ERCE reading, PASEC mathematics and PASEC reading represent the source assessments and TIMSS and PIRLS represent the target assessments.

The first step was to evaluate the relationship between the data from source and target assessments. The correlation analyses between the average scores of ERCE mathematics and TIMSS (r = .80 to .82), PASEC mathematics and TIMSS (r = .70 to .80), ERCE reading and PIRLS (r = .78 to .82) and PASEC reading and PIRLS (r = .73 to .81) indicated that the ERCE and Rosetta Stone scales and the PASEC and Rosetta Stone scales measure different but similar constructs; that is, correlations were reasonably high for constructing a concordance.

The concordance scores and levels were then identified based on estimated PASEC and ERCE average scores using the combined data of Burundi, Guinea and Senegal for PASEC and Colombia and Guatemala for ERCE. The score ranges of the average scores of the PASEC and ERCE mathematics and reading scales were rounded either up or down to cover almost all the data of the participating countries, and to be as symmetric as possible around the overall mean (500 for PASEC, 700 for ERCE). For both PASEC scales, mathematics and reading, scores ranged from about 200 to 800. For both ERCE scales, mathematics and reading, scores ranged from about 440 to 940. In both cases, these ranges cover almost 99.5% of the data.

For both the ERCE and PASEC scales, 20 points were specified as the score interval to include enough score or proficiency levels and to retain as much information as possible. As a result, there are 26 score levels within the score range of 440 to 940 and 200 and 800, respectively.

For each identified concordance score level, a predictive mean matching model was used to select five donors (i.e. test-takers) for each test and from each country so that each country contributes equally to each of the concordance tables. Each of the donors donated five PVs on the target tests. The mean and standard deviation of the donors' PVs from the Rosetta Stone linking data were calculated based on the total donated PVs at each concordance level. Note that these steps were implemented separately for ERCE mathematics and reading and for PASEC mathematics and reading.

Preliminary concordance tables for ERCE mathematics, ERCE reading, PASEC mathematics and PASEC reading were created by assigning the estimated mean and standard deviation of each set of PVs based on the Rosetta Stone (TIMSS and PIRLS) linking data, respectively, to each concordance score level in the specified range of ERCE mathematics, ERCE reading, PASEC mathematics and PASEC reading. Finally, for each concordance score point, the mean of the donated PVs was smoothed by applying a simple moving average, using a window of seven score points. The standard deviation of the PVs of each score point was smoothed in a similar way as the means of PVs, using a moving geometric mean of the variances of each set of the five donated PV means clustered at the corresponding score level in the table. The square root of this smoothed variance becomes the smoothed conditional standard deviation.

THE CONCORDANCE TABLES

Tables 1 and 2 show the final concordance tables for ERCE/PASEC mathematics and ERCE/PASEC reading, respectively. The first and fifth columns of each table show the source assessment (ERCE or PASEC) concordance score levels, either mathematics or reading. The second and sixth columns show the predictive distribution, projected means and conditional standard deviations of the source assessment score on the TIMSS or PIRLS scale. The third, fourth, seventh and eighth columns show the scale values for the percentiles of this conditional distribution given the source assessment (ERCE or PASEC) on the TIMSS or PIRLS scales. The lower and upper values at the 68th and 95th percentiles are provided.

TABLE 1. Concordance table for ERCE and PASEC in mathematics

ERCE Mathematics Score	Projected Score on TIMSS Scale		Lower Bound		Upper Bound	
	Mean	SD	95%	68%	68%	95%
400	290	64	162	226	354	417
420	304	63	178	241	367	430
440	319	62	194	256	381	443
460	318	63	192	255	381	444
480	322	62	198	260	384	446
500	326	62	201	264	389	451
520	334	62	211	273	396	458
540	342	64	214	278	406	470
560	357	63	231	294	419	482
580	371	62	247	309	433	495
600	389	61	266	327	450	511
620	403	61	282	342	463	524
640	420	58	303	361	478	537
660	432	57	317	375	489	546
680	449	53	344	397	502	555
700	465	52	362	414	517	569
720	481	51	379	430	532	583
740	497	49	399	448	547	596
760	515	50	415	465	565	616
780	531	50	431	481	581	631
800	548	48	453	500	596	643
820	563	46	471	517	609	655
840	576	46	484	530	622	668
860	590	45	500	545	635	680
880	599	46	508	554	645	691
900	608	46	516	562	654	699
920	617	48	520	568	665	713
940	624	51	522	573	675	726
960	638	50	538	588	688	739
980	653	49	554	603	702	751

PASEC Mathematics Score	Projected Score on TIMSS Scale		Lower Bound		Upper Bound	
	Mean	SD	95%	68%	68%	95%
220	197	76	44	121	274	350
240	210	76	58	134	286	362
260	223	75	72	148	298	374
280	224	75	74	149	300	375
300	229	77	75	152	306	382
320	231	76	78	155	307	383
340	235	75	85	160	310	385
360	241	71	99	170	312	383
380	254	69	115	184	323	392
400	265	68	130	197	333	401
420	273	68	137	205	341	410
440	284	65	154	219	349	414
460	297	62	172	234	359	421
480	315	61	194	254	376	437
500	336	63	209	273	399	462
520	344	65	215	280	409	473
540	355	64	227	291	419	483
560	371	63	245	308	434	497
580	382	66	251	317	448	514
600	395	70	256	326	465	535
620	403	71	260	332	475	546
640	417	71	274	345	488	559
660	437	69	299	368	506	575
680	453	69	316	385	522	591
700	469	67	335	402	536	602
720	484	64	357	421	548	612
740	500	57	386	443	556	613
760	513	52	408	461	566	618
780	526	52	422	474	578	630
800	539	51	436	487	590	641

SD = standard deviation

TABLE 2. Concordance table for ERC and PASEC in reading

ERCE Reading Score	Projected Score on PIRLS Scale		Lower Bound		Upper Bound	
	Mean	SD	95%	68%	68%	95%
400	284	84	116	200	369	453
420	296	83	130	213	380	463
440	309	82	144	226	391	473
460	321	81	158	240	402	483
480	329	80	169	249	409	489
500	342	77	187	265	419	497
520	347	76	195	271	424	500
540	373	74	224	298	447	521
560	378	74	229	303	452	526
580	395	68	258	326	463	532
600	410	65	279	344	475	540
620	421	62	297	359	483	545
640	436	62	311	373	498	560
660	455	63	328	392	518	581
680	467	64	338	402	531	595
700	482	60	362	422	542	602
720	488	63	361	425	551	615
740	501	65	372	437	566	631
760	522	64	393	457	586	650
780	526	63	399	463	590	653
800	536	61	414	475	597	658
820	546	60	425	486	607	667
840	555	60	436	495	614	674
860	565	59	447	506	623	682
880	568	56	456	512	624	680
900	576	56	463	519	632	688
920	588	58	473	530	646	703
940	596	59	477	537	655	715
960	608	58	492	550	666	725
980	620	57	506	563	678	735

PASEC Reading Score	Projected Score on PIRLS Scale		Lower Bound		Upper Bound	
	Mean	SD	95%	68%	68%	95%
220	146	72	2	74	218	290
240	161	72	17	89	233	304
260	175	72	31	103	247	319
280	178	72	34	106	249	321
300	181	72	38	110	253	325
320	190	71	47	118	261	332
340	196	72	52	124	267	339
360	205	71	63	134	276	347
380	216	72	73	145	288	359
400	228	72	84	156	300	372
420	238	76	87	163	314	390
440	253	74	104	179	327	401
460	265	73	120	193	338	411
480	280	71	139	209	351	422
500	297	71	155	226	369	440
520	317	73	172	244	390	462
540	330	72	186	258	402	474
560	351	66	219	285	417	482
580	364	66	232	298	430	496
600	377	68	241	309	446	514
620	392	69	255	323	461	529
640	405	67	271	338	471	538
660	420	63	295	357	483	545
680	444	66	312	378	511	577
700	456	69	319	388	525	593
720	473	71	332	402	544	615
740	486	70	346	416	555	625
760	492	72	347	420	565	637
780	507	72	362	434	579	651
800	521	72	377	449	593	665

SD = standard deviation

Concordance tables must be interpreted with caution as they are not perfect predictions of how a student would perform on a target test (e.g. TIMSS or PIRLS). They do not provide a direct link between assessments and are dependent on the characteristics of the sample. Therefore, the uncertainty of the prediction has to be taken into consideration when using and interpreting concordance tables. For example, a PASEC mathematics score of 500 does not result in a TIMSS score of 336 but in a distribution of TIMSS scores with a mean of 336. In other words, assuming approximately normal conditional score distributions, 68% of the generated PVs on the TIMSS scale would likely fall in the score range of 273 to 399 (if a student with similar ability took the TIMSS assessment) and 95% of generated PVs on the TIMSS scale would likely fall in the score range of 209 to 462, as shown in Table 1.

PROPORTION OF STUDENTS ACHIEVING SDG 4.1.1

Besides making inferences about the likely score range on TIMSS or PIRLS scales given an ERCE or PASEC score, it is also possible to generate the likely PVs for individual students (who participated in ERCE or PASEC but not in TIMSS or PIRLS) on the TIMSS and PIRLS scales by using the projected means and standard deviations from the concordance tables. The Rosetta Stone analysis reports provide detailed instructions on how to generate random PVs using the concordance tables. These projected PVs can then be used to estimate the percentage of students reaching a given benchmark, for example, the global minimum proficiency level (MPL) for the end of primary, that indicates the achievement of SDG 4.1.1.

It should be noted that the ERCE and PASEC benchmarks serving as MPLs are meant to map into the global MPL and, therefore, to indicate the achievement of SDG 4.1.1. However, they were set independently by the technical teams of each assessment. Consequently, they are qualitatively different to each other. The descriptions of what students know and can do when they reach the MPLs set for ERCE and PASEC, as well as for the global MPL (set by international expert consensus and aligned with TIMSS/PIRLS benchmarks[4]) can be found in Table 3.

4 See ACER, 2019 and UIS, 2022 in *Further Reading*.

TABLE 3. Definition of the ERCE, PASEC and TIMSS/PIRLS benchmarks serving as MPLs for Grade 6 (ERCE and PASEC) and Grade 4 (TIMSS/PIRLS)

Study	Mathematics		Reading	
ERCE	**Level 3 (789 points)**	Students at this level showed evidence of being able to: • Solve more complex problems that require interpreting information and involve two or more operations including multiplication or division. • Interpret the meaning of proportional variations in contextualized situations. • Identify equivalent fractions (with a denominator other than 10) and calculate additions and subtractions of fractions with the same denominator. • Relate decimal numbers to simple proper fractions or simple mixed numbers (e.g with denominator 2) and calculate or estimate additions and subtractions of decimal numbers. • Determine missing intermediate terms of a sequence presented in a contextualized situation, interpreting its pattern of formation. • Identify relations of perpendicularity and parallelism in the plane. • Solve complex problems that involve calculation or estimation of areas and perimeters of geometric figures. • Solve problems involving measurements (mass volume and time measurements) and convert units of measurement. • Solve problems that require reading and interpreting information from tables and graphs or identifying graphs that represent information delivered in different formats.	**Level 3 (754 points)**	Students at this level showed evidence of being able to: • Make inferences from connections between specific or secondary ideas and located in different parts of one or more texts. • Infer the central theme of a paragraph or part of the text, establishing a relationship with the text as a whole. • Make inferences (e.g. inferring the central theme, the characteristics or feelings of the characters, the conflict and the outcome) that require understanding the text globally and integrating implicit ideas present in it. • Interpret expressions in figurative language based on clues that are implicit in the text or that challenge the student's knowledge of the world around him or her. • Relate visual and verbal information in a text.
PASEC	**Level 3 (609 points)**	Students at this level master an oral number sequence (counting up to 60 in two minutes) and are able to compare numbers, complete logical series and perform operations (sums and subtractions) with numbers over 50. They can solve problems with numbers under 20 using reasoning skills.	**Level 4 (595 points)**	Intermediate reader: enhanced reading autonomy is bolstering students' understanding of sentences and texts. Students have acquired written language decoding and listening comprehension competencies which enable them to understand explicit information in words, sentences and short passages. They can combine their decoding skills and their mastery of the oral language to grasp the literal meaning of a short passage.
TIMSS/ PIRLS	**TIMSS Intermediate (475 points)**	Students can apply basic mathematical knowledge in simple situations. They can compute with three- and four-digit whole numbers in a variety of situations. They have some understanding of decimals and fractions. Students can identify and draw shapes with simple properties. They can read, label and interpret information in graphs and tables.	**PIRLS Low (400 points)**	When reading predominantly simpler Literary Texts, students can: • Locate and retrieve explicitly stated information, actions or ideas. • Make straightforward inferences about events and reasons for actions. • Begin to interpret story events and central ideas. When reading predominantly simpler Informational Texts, students can: • Locate and reproduce explicitly stated information from text and other formats (e.g. charts, diagrams). • Begin to make straightforward inferences about explanations, actions and descriptions.

The proportions of students in ERCE and PASEC countries who reach SDG 4.1.1 for Mathematics (intermediate benchmark in TIMSS) according to ERCE (Performance Level 3) and PASEC (Performance Level 3), as well as according to the concordance tables produced by the Rosetta Stone Study, are shown in Table 4. The proportions of students in ERCE and PASEC countries who reach SDG 4.1.1 for Reading (low benchmark in PIRLS) according to ERCE (Performance Level 3) and PASEC (Performance Level 4), as well as according to the concordance tables produced by the Rosetta Stone Study, are shown in Table 5. The standard errors of these percentages are calculated using proper weights and PV formulas to include assessment and sampling errors (see the Rosetta Stone analysis report for details).

It is important to note that the percentages estimated based on Rosetta Stone are in many cases considerably different from those reported based on PASEC and ERCE scores. In most cases, the percentages are higher when the estimations are based on Rosetta Stone for ERCE and lower for PASEC. These discrepancies could be due to differences in the assessment frameworks and the constructs measured in each. As mentioned above, the MPLs set by each assessment are substantially different. For example, while ERCE considers that the MPL has been reached when students can "interpret expressions in *figurative* language based on clues that are *implicit* in the text", PASEC considers that the MPL has been reached when students can "[…] combine their decoding skills and their mastery of the oral language to grasp the *literal* meaning of a short passage" (see Table 3).

Furthermore, PASEC was developed within and for the context of francophone sub-Saharan African countries, while ERCE was developed within and for the context of Latin American countries. On the other hand, TIMSS and PIRLS were designed to be global initiatives. While these assessments measure both mathematics and reading, the very definition of the constructs varies among them; they have different starting points and set different aspirations for their education systems.

Finally, the target populations of the assessments are different: Grade 6 for PASEC and ERCE and Grade 4 for TIMSS and PIRLS. Other factors could also play an important role in explaining these differences, like the stage when the transition from local languages to the testing language (e.g. French, Spanish) is introduced in the school system. In many countries, this transition occurs in Grade 1, while in others it is as late as Grade 4. At this point, these are all unverified hypotheses, and more research (and larger samples of countries and students participating in Rosetta Stone) is needed to test them and provide clear answers.

TABLE 4. The proportion of students in ERCE and PASEC countries achieving SDG 4.1.1 for Mathematics

Country	Students above MPL Overall		Students above MPL Boys		Students above MPL Girls		Students above MPL Overall		Students above MPL Boys		Students above MPL Girls	
	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE
	ERCE (6th grade) MPL						Rosetta Stone (TIMSS 4th grade) MPL					
Argentina	13.1	(0.9)	14.2	(1.1)	12.0	(0.9)	42.7	(1.4)	44.2	(1.6)	41.2	(1.6)
Brazil	28.7	(1.5)	30.5	(1.7)	26.7	(1.7)	56.9	(1.5)	58.2	(1.9)	55.6	(1.8)
Colombia	16.5	(1.2)	17.6	(1.4)	15.3	(1.7)	48.2	(1.8)	49.4	(2.1)	46.9	(2.1)
Costa Rica	20.8	(1.7)	22.6	(1.9)	19.0	(1.9)	56.4	(1.6)	57.9	(2.2)	54.9	(1.9)
Dominican Republic	2.1	(0.5)	2.1	(0.5)	2.1	(0.6)	22.2	(1.4)	20.9	(1.4)	23.5	(2.0)
Cuba	20.6	(1.5)	20.5	(1.6)	20.8	(1.7)	42.2	(1.9)	42.0	(2.0)	42.7	(2.1)
Ecuador	22.8	(1.4)	23.0	(1.5)	22.4	(1.6)	53.1	(1.5)	53.1	(1.7)	53.0	(1.7)
El Salvador	7.4	(0.6)	8.0	(0.7)	6.8	(0.8)	35.7	(1.0)	36.5	(1.7)	34.9	(1.6)
Guatemala	6.6	(0.9)	7.6	(1.2)	5.6	(0.8)	30.1	(1.2)	32.4	(1.5)	27.8	(1.5)
Honduras	11.1	(1.9)	11.9	(2.0)	10.4	(1.9)	38.5	(2.2)	40.4	(2.3)	36.7	(2.6)
Mexico	37.9	(1.3)	36.9	(1.6)	38.9	(1.7)	66.1	(1.5)	64.9	(1.7)	67.3	(1.6)
Nicaragua	3.0	(0.5)	3.8	(0.7)	2.3	(0.5)	30.2	(1.2)	33.0	(1.8)	27.5	(1.5)
Panama	3.3	(0.4)	3.4	(0.5)	3.2	(0.5)	25.2	(1.0)	25.4	(1.3)	25.0	(1.2)
Paraguay	5.6	(0.7)	5.8	(0.8)	5.3	(0.9)	27.1	(1.3)	27.8	(1.6)	26.4	(1.6)
Peru	38.7	(1.5)	38.3	(1.8)	39.2	(1.8)	66.3	(1.3)	66.1	(1.6)	66.6	(1.7)
Uruguay	37.8	(1.5)	38.1	(1.8)	37.5	(1.7)	65.8	(1.4)	66.6	(2.0)	65.0	(1.9)
	PASEC (6th grade) MPL						Rosetta Stone (TIMSS 4th grade) MPL					
Benin	19.3	(2.6)	18.9	(2.4)	19.8	(3.0)	8.9	(1.4)	8.8	(1.5)	9.0	(1.7)
Burkina Faso	25.2	(1.2)	26.5	(1.5)	24.1	(1.4)	10.2	(0.8)	10.9	(0.9)	9.5	(1.1)
Burundi	18.2	(1.5)	23.7	(1.8)	13.8	(1.5)	7.8	(0.7)	9.6	(1.0)	6.4	(0.8)
Cameroon	11.2	(1.0)	10.5	(1.0)	12.1	(1.4)	5.1	(0.8)	5.0	(0.9)	5.1	(1.0)
Chad	1.9	(0.6)	1.8	(0.6)	2.0	(0.7)	1.2	(0.3)	1.3	(0.4)	1.0	(0.5)
Congo	7.8	(0.9)	6.9	(1.0)	8.6	(1.3)	3.7	(0.5)	3.5	(0.7)	3.9	(0.7)
Côte d'Ivoire	2.6	(0.6)	2.9	(0.7)	2.3	(0.7)	1.5	(0.4)	1.6	(0.4)	1.4	(0.5)
DRC	3.2	(1.1)	3.4	(1.3)	3.0	(1.0)	1.9	(0.7)	2.1	(0.9)	1.6	(0.6)
Gabon	23.1	(1.9)	26.5	(2.4)	19.9	(2.0)	8.8	(1.1)	9.8	(1.8)	7.9	(1.1)
Guinea	6.9	(1.2)	7.5	(1.4)	6.2	(1.4)	3.1	(0.7)	3.6	(0.9)	2.5	(0.6)
Madagascar	6.3	(1.7)	5.7	(1.7)	7.0	(1.9)	3.0	(0.7)	2.9	(0.8)	3.0	(0.8)
Niger	8.0	(1.3)	8.1	(1.5)	7.8	(1.4)	3.5	(0.7)	3.6	(0.8)	3.4	(0.8)
Senegal	27.6	(1.9)	27.8	(2.2)	27.4	(2.2)	12.0	(1.3)	12.6	(1.5)	11.5	(1.3)
Togo	16.0	(1.2)	16.1	(1.4)	15.8	(1.3)	7.0	(0.6)	7.0	(0.7)	6.9	(0.7)

MPL = minimum proficiency level; SE = standard error

TABLE 5. **The proportion of students in ERCE and PASEC countries achieving SDG 4.1.1 for Reading**

Country	Students above MPL Overall		Students above MPL Boys		Students above MPL Girls		Students above MPL Overall		Students above MPL Boys		Students above MPL Girls	
	%	SE	%	SE	%	SE	%	SE	%	SE	%	SE
ERCE (6th grade) MPL							**Rosetta Stone (PIRLS 4th grade) MPL**					
Argentina	31.7	(1.2)	28.8	(1.3)	34.5	(1.4)	78.9	(1.1)	77.6	(1.3)	80.2	(1.3)
Brazil	43.3	(1.5)	40.2	(1.7)	46.7	(1.8)	86.2	(0.8)	84.4	(1.1)	88.1	(1.0)
Colombia	37.3	(1.7)	35.2	(2.2)	39.5	(2.0)	83.7	(1.1)	82.3	(1.5)	85.1	(1.5)
Costa Rica	53.8	(1.7)	51.1	(2.0)	56.4	(1.9)	90.6	(0.9)	89.8	(0.9)	91.5	(1.3)
Cuba	44.4	(1.3)	38.6	(1.6)	50.4	(1.6)	86.6	(1.1)	83.6	(1.6)	89.6	(1.1)
Dominican Republic	16.3	(1.1)	12.1	(1.0)	20.6	(1.5)	65.7	(1.2)	61.1	(1.7)	70.4	(1.6)
Ecuador	26.0	(1.0)	24.1	(1.2)	27.9	(1.3)	76.5	(1.1)	75.9	(1.5)	77.1	(1.2)
El Salvador	29.2	(1.2)	26.3	(1.3)	32.3	(1.6)	80.2	(1.1)	78.3	(1.3)	82.1	(1.3)
Guatemala	15.9	(1.0)	15.5	(1.4)	16.3	(1.1)	67.0	(1.4)	67.0	(2.0)	66.9	(2.3)
Honduras	16.1	(1.5)	14.0	(1.5)	18.1	(1.9)	72.1	(1.5)	71.7	(1.8)	72.5	(1.6)
Mexico	41.4	(1.4)	37.8	(1.8)	45.0	(1.6)	84.1	(0.9)	83.0	(1.1)	85.1	(1.1)
Nicaragua	12.8	(0.8)	11.7	(0.8)	14.0	(1.1)	71.4	(1.2)	70.4	(1.7)	72.5	(1.5)
Panama	17.4	(1.1)	15.3	(1.1)	19.4	(1.2)	68.7	(1.4)	67.2	(1.7)	70.0	(1.5)
Paraguay	18.7	(1.1)	15.8	(1.2)	21.5	(1.4)	69.7	(1.3)	66.9	(1.6)	72.4	(1.6)
Peru	48.8	(1.2)	44.6	(1.5)	53.4	(1.5)	85.7	(1.0)	84.5	(1.4)	87.1	(1.4)
Uruguay	43.5	(1.3)	40.3	(1.5)	46.7	(1.5)	85.3	(0.9)	83.7	(1.0)	87.0	(1.3)
PASEC (6th grade) MPL							**Rosetta Stone (PIRLS 4th grade) MPL**					
Benin	45.5	(2.7)	44.0	(2.6)	47.3	(3.1)	37.4	(2.2)	36.2	(2.5)	38.7	(2.8)
Burkina Faso	33.0	(1.4)	32.6	(1.5)	33.3	(1.6)	27.8	(1.0)	27.8	(1.5)	27.7	(1.5)
Burundi	4.5	(1.0)	4.7	(1.0)	4.4	(1.1)	9.9	(0.8)	10.5	(0.9)	9.4	(1.0)
Cameroon	30.3	(1.7)	28.2	(1.7)	32.7	(2.2)	25.0	(1.3)	24.0	(1.5)	26.1	(1.8)
Chad	7.6	(1.2)	7.8	(1.2)	7.4	(1.4)	8.8	(1.1)	9.2	(1.3)	8.2	(1.5)
Congo	33.6	(1.8)	30.2	(1.9)	37.1	(2.2)	28.0	(1.5)	25.6	(1.9)	30.5	(1.9)
Côte d'Ivoire	22.1	(1.9)	21.0	(2.2)	23.3	(1.9)	19.3	(1.4)	19.2	(1.7)	19.5	(1.6)
DRC	9.2	(1.5)	9.9	(1.4)	8.5	(1.8)	11.0	(1.2)	11.9	(1.4)	10.0	(1.4)
Gabon	76.3	(1.8)	73.7	(2.1)	78.9	(1.9)	56.7	(1.8)	55.2	(2.3)	58.3	(2.2)
Guinea	22.2	(1.7)	21.9	(1.9)	22.6	(2.1)	19.5	(1.6)	19.3	(1.9)	19.7	(1.9)
Madagascar	6.3	(2.0)	5.4	(1.8)	7.1	(2.4)	8.1	(1.2)	7.1	(1.3)	9.1	(1.4)
Niger	14.4	(1.6)	13.1	(1.8)	15.9	(1.7)	13.4	(1.2)	12.6	(1.5)	14.4	(1.5)
Senegal	41.1	(2.2)	38.2	(2.3)	43.5	(2.7)	34.2	(1.9)	32.9	(2.3)	35.3	(2.3)
Togo	19.4	(1.1)	18.5	(1.3)	20.2	(1.3)	17.5	(0.9)	16.7	(1.2)	18.2	(1.3)

MPL = minimum proficiency level; SE = standard error

KEY MESSAGES

- The SDG framework explicitly calls for national systems to track their goal achievement using comparable indicators – that is, indicators that can be compared across countries and over time. However, United Nations agencies have been reluctant to impose a specific measurement framework in a top-down manner. Aside from the humanistic and political issues involved, such a task would require excessive time and resources from all stakeholders. Furthermore, many theories of economic development and public finance suggest minimizing solutions that are overly standardized. An optimal compromise was considered to be developing scales for existing measurement tools and rendering them comparable ex-post, via analytical and empirical means. This is what Rosetta Stone has done keeping the perspective that national contexts remain different (both sub-Saharan African and Latin American countries) and that global comparisons are sometimes limited. To our knowledge, this is a pioneering effort, perhaps the first of its kind in the field of learning measurement. Similar, but somewhat different, efforts by the UIS and other partners are also under way.

- Comparable scales were established across countries and across assessments. The Rosetta Stone results indicate that, from a statistical perspective, the reading and mathematics constructs measured by PASEC and ERCE are similar enough to the constructs measured by TIMSS and PIRLS to enable the desired concordance. These concordance tables are the first steps that need to be part of a broader exercise that allows more robust inferences. The concordance tables included in this document can be used to make inferences about the likely score range on TIMSS or PIRLS scales given a PASEC or ERCE score, allowing countries to compare their students' achievement globally and providing data that help them to measure global progress towards SDG 4.1.1.

- The Rosetta Stone Study results for ERCE and PASEC suggest that concordance tables can be established for other regional assessments (e.g. SAQMEC, SEA-PLM, PILNA). The resulting information and further analyses would be extremely useful and valuable as it would allow all regional assessments to compare not only to TIMSS and PIRLS but also to each other.

- The information provided by Rosetta Stone and regional assessments like PASEC and ERCE must be considered complementary. Both regional assessments and Rosetta Stone play an irreplaceable role in the global strategy for measuring and monitoring progress towards SDG 4.1.1. Together they enhance the possibilities for depth and breadth of the analysis that can be carried out and, in consequence, improve the quality and relevance of the information available to policymakers and other stakeholders.

- Rosetta Stone opens up endless possibilities for secondary analyses that can help improve global reporting and facilitate comparative analyses of education systems around the globe. Each assessment reflects the curriculum of the region/country, while the proficiency level distribution is adequate for global reporting.

- The concordance tables presented here are based on only two countries for ERCE and three countries for PASEC. For countries that did not participate in the Rosetta Stone Study (i.e. did not administer TIMSS/PIRLS linking booklets), the use of the concordance tables provided in this document is an extrapolation and comes with added uncertainty that cannot be fully accounted for without also conducting a Rosetta Stone data collection. However – and this point cannot be overemphasized – this approach is better than any existing alternative.

- Increasing national sample sizes and adding more countries per regional assessment would further improve the estimated concordance and would allow research to be conducted to explain the observed differences in the percentage of students reaching SDG 4.1.1 when estimated with Rosetta Stone versus ERCE or PASEC.

- Further reflection about the establishment of the minimum proficiency levels for each regional study that best map into the agreed global proficiency level is needed to ensure fairer comparisons of the percentages of students that reach SDG 4.1.1 in each education system.

- Should countries be interested in establishing a concordance between their national assessment and TIMSS and PIRLS, the Rosetta Stone approach could also be conducted. However, the costs per country would be relatively high because the results would only be applicable to one country and the linking instruments would need to be adapted to each national assessment.

- While adding national assessments to Rosetta Stone might be an option for countries that are not yet included in any regional assessment, because of the complementarity of regional assessments and Rosetta Stone discussed above, it is clear that having more countries participating in both regional (e.g. PASEC and ERCE) and international (e.g. TIMSS and PIRLS) assessments would be a much more efficient strategy.

FURTHER READING

Australian Council for Educational Research (ACER) (2019). *Minimum Proficiency Levels Described, Unpacked and Illustrated*. GAML/REF/2.
https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2019/05/GAML6-REF-2-MLP-recommendations-ACER.pdf

International Association for the Evaluation of Educational Achievement (IEA) (2017). *Measuring SDG 4: How PIRLS Can Help; How the Progress in International Reading Literacy Study (PIRLS) Helps Monitor Sustainable Development Goal 4 Targets*. Paris: UNESCO. (ED/IPS/IGE/2017/05).
https://unesdoc.unesco.org/ark:/48223/pf0000260607

International Association for the Evaluation of Educational Achievement (IEA) (2017). M. O. Martin, I. V. S. Mullis and M. Hooper (eds), *Methods and Procedures in PIRLS 2016*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and IEA.
https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

International Association for the Evaluation of Educational Achievement (IEA) (2020). M. O. Martin, M. von Davier and I. V. S. Mullis (eds), *Methods and Procedures: TIMSS 2019 Technical Report*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and IEA.
https://timssandpirls.bc.edu/timss2019/methods

Khorramdel, L., L. Yin, P. Foy, J. Yoon Jung, U. Bezirhan and M. von Davier (2022). *Establishing a Concordance between ERCE and TIMSS/PIRLS.* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
https://timssandpirls.bc.edu/Rosetta-Stone-Reports/index.html

Khorramdel, L., L. Yin, P. Foy, J. Yoon Jung, U. Bezirhan and M. von Davier (2022). *Establishing a Concordance between PASEC and TIMSS/PIRLS.* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
https://timssandpirls.bc.edu/Rosetta-Stone-Reports/index.html

Programme d'Analyse des Systèmes Éducatifs (PASEC) (2017). *Rapport technique de l'évaluation internationale PASEC 2014.* Dakar: PASEC, Conférence des Ministres de l'Education des États et Gouvernements de la Francophonie (CONFEMEN).
https://www.pasec.confemen.org/wp-content/uploads/2017/12/RAPPORT-TECHNIQUE-PASEC2014.pdf

UNESCO (2016). *Tercer Estudio Regional Comparativo y Explicativo: Reporte Técnico*. Santiago: UNESCO.
https://unesdoc.unesco.org/ark:/48223/pf0000247123

UNESCO Institute for Statistics (UIS) (2022), *Metadata for Global Indicator 4.1.1*
https://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/Minimum-Proficiency-Levels-MPLs.pdf.

LIST OF ACRONYMS

ERCE	Estudio Regional Comparativo y Explicativo.
IEA	International Association for the Evaluation of Educational Achievement.
ILSAs	International large-scale assessments.
IRT	Item response theory.
MPL	Minimum Proficiency Level.
PASEC	Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN.
PIRLS	Progress in International Reading Literacy Study.
PV	Plausible value.
SAQMEC	The Southern and Eastern Africa Consortium for Monitoring Educational Quality.
SD	Standard Deviation is a measure of how dispersed the data is in relation to the mean.
SE	Standard Error is the standard deviation of the mean in repeated samples from a population.
SDGs	Sustainable Development Goals.
SEA-PLM	South-East Asia Primary Learning Metrics.
TIMSS	Trends in International Mathematics and Science Study.
UIS	UNESCO Institute for Statistics.


unesco
Institute for Statistics


IEA
TIMSS & PIRLS
International Study Center
Lynch School of Education
BOSTON COLLEGE


BILL & MELINDA
GATES foundation