unesco
Institute for Statistics

January 2023

The feasibility of harmonizing scores produced by Assessments for Minimum Proficiency Levels (AMPL) to the TIMSS and PIRLS test scores to measure and monitor SDG 4.1.1b

The feasibility of harmonizing scores produced by Assessments for Minimum Proficiency Levels (AMPL) to the TIMSS and PIRLS test scores to measure and monitor SDG 4.1.1b

Paper prepared for the UNESCO Institute for Statistics by Andrés Sandoval-Hernandez (University of Bath), Diego Carraso (MIDE UC) and David Torres Irribarra (Pontificia Universidad Católica).

Table of Contents

Executive Summary

The present document assesses the feasibility of harmonizing the scores generated by the Assessments for Minimum Proficiency Levels (AMPL) to the scores of Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS).

The rationale of the present inquiry is to produce a judgment call regarding the idea of using the information of the MILO study, where students in 2021 answered the AMPL[1] test, and the PASEC test[2]. And similarly, take advantage of the information generated by the IEA's Rosetta Stone study, here referred as "Rosetta-PASEC" study, where students answered the Rosetta Link test and the PASEC test, in 2020.

We argue that if all three tests (AMPL, PASEC, Rosetta Link) were using the same response model to generate scores (e.g., Rasch model), then it would be straightforward to make score conversions. However, even under this fictitious scenario, different assumptions and conditions should be met (e.g., model invariance, unidimensionality, assuring the three tests do measure the same attribute). However, the AMPL test, PASEC test and Rosetta Link test, do not use the same model to generate scores. As such, linear conversion between these test scores is not advisable, because such a conversion would distort the score meaning. The Rosetta Stone study recurs instead, to score projections between the source test and the reference test instead (i.e., model-based predictions). The Rosetta Stone design study, according to the Rosetta Study Report, is recommended only after a population of students have answered both tests simultaneously (source and target),

[1] Assessments for Minimum Proficiency Levels (AMPL)

[2] Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN (PASEC)

and do not encourage the use of the conversion tables for countries who have not participated in a Rosetta Study (UNESCO UIS, 2022a, p. 44), due to the prediction uncertainty.

The comparison of the MILO study and Rosetta-PASEC Study designs, lead us to believe both initiatives require field applications and cannot be implemented by secondary data use alone. Thus, If both initiatives, MILO and Rosetta Stone present the same costs for any new participating country willing to produce results for SDG 4.1.1b through these assessments, then one should judge each of these in terms of their trustworthiness. Because Rosetta Stone produces non-convergent results with two methods on the same sample, we believe the MILO design seems a more efficient method to produce results for the SDG 4.1.1b population estimates, without the need to use the TIMSS and PIRLS benchmarks.

Introduction

The present document assesses the feasibility of harmonizing the scores generated by the Assessments for Minimum Proficiency Levels (AMPL) to the scores of Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), using the Programme for the Analysis of Education Systems (PASEC; Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN) as a psychometric bridge.

The general idea is to evaluate if it is possible and sensible to transform the scores from the AMPL / Monitoring Impacts on Learning Outcomes (MILO) study to different scales generated using different methodologies. In particular, the generated scores using the TIMSS and PIRLS metric, given the MILO study already can express its scores in the PASEC metric. We will assess the feasibility of these score conversions from the point of view of **equating** and scale-linking practices from international large-scale studies.

This document is divided into the following sections. We first describe a summary of the main conceptual tools of equating studies used in large scales assessments, applicable to the present scenario. In the second section, we summarize the main features of the Monitoring Impacts on Learning Outcomes (MILO) and the Rosetta Stone study "Establishing a Concordance between PASEC and TIMSS/PIRLS" (Rosetta-PASEC) relevant to the feasibility assessment. We highlight the commonalities and contrasts of these two initiatives to inform the feasibility exercise. In the second section, we considered three different options for score harmonization. We describe their assumptions, requirements and suitability, to fulfil the aim of producing SG 4.1.1b population estimates. Finally, we close the document with a summary and conclusion.

Conceptual framework

Test scores and equating scores

Test scores properties

International large-scale assessment studies (ILSA) produce scores in different metrics. These different metrics are defined in an ad hoc manner for each study, each defining its own unit used to express results, yielding different score scales. These different score scales can vary in their means and standard deviations. Some of these studies use a mean of 500, and a standard deviation of 100 for the score of their tests, as it is in the case of Trends in International Mathematics and Science Study (TIMSS) (Martin et al., 2016), and in Programme for International Student Assessment (PISA) (Adams & Wu, 2000). In contrast, studies such as the Third Regional Explanatory and Comparative Study from 2013 (TERCE) (UNESCO-OREALC, 2016) use a mean of 700 and a standard deviation of 100. Even in the cases where the studies share common values for the scale means and standard deviations, the scores are not directly comparable, as these scores can represent different attributes and can be generated based on statistical models with different properties, in particular, as a linear transformation of the scores generated with different item response theory models.

Item response theory models are a family of statistical models that are used to analyze or predict the probability that different persons will present a given response to an item (like a “correct” answer or “agree” to a given statement) as a function of item properties, for instance, its difficulty, and a person’s level in a given attribute of interest, such as their level of reading proficiency. There are different kinds of item

response models, with different scaling properties depending, for instance, on the number of item characteristics that are considered in the model (Borsboom, 2005; Mellenbergh, 1994, De Boeck & Wilson, 2004). Despite their potential differences, these models are commonly based on the assumption that participants with higher levels in the attribute that the test was designed for, will provide more correct answers in the test than those with lower levels. And conversely, participants with lower levels of the attribute of interest will be less likely to provide correct answers in the test. Traditionally, these models generate raw estimates of these attributes of interest in a logit (log odds) scale, oftentimes expressed in a scale with both positive and negative values distributed around a mean of zero. These raw values are then transformed using a linear transformation in order to express them in an arbitrarily selected mean and standard deviation that is expected to be easier to communicate to the public, such as a mean of 500 and standard deviation of a 100, as we described earlier.

Setting test scores on a scale

Scale scores for the international testing programs discussed here are generated as transformations of two different but common item response theory models, namely the Rasch model and the 2 and 3 Parameter Logistic (2PL) model. The key difference between these models is that while the Rasch model characterizes a single feature of items, namely their difficulty, the 2PL and 3PL models consider respectively two and three item properties: (1) the item difficulty, (2) the strength of the relationship between the item and the attribute (often referred as item discrimination), and (3) the probability that a given item can be answered correctly purely by chance. A key consequence of this difference between the models is that under the Rasch Model, the person results have a one-to-one relation between the observed total scores and

the model estimates, this means in practice that within the same test, everyone that has the same number of correct responses will share the same estimate of their level in the attribute assessed by the test. This is not the case in the 2PL and 3PL models, where persons with the same total score may have different estimates due to the different weights that items have based on the relationship between the item and the attribute.

A common challenge when trying to compare results from a different test that aims to measure a given attribute is that they may be very different, each presenting different items and in different numbers. Within psychometrics, different techniques have been developed to deal with this challenge under the name of *equating procedures*. Based on the assumption that the different tests are indeed assessing the same attribute, these methods rely on different research designs to obtain common points of reference that would enable a comparison between the scales of the two different tests. This research design requires common persons in the study design, and or common items (González & Wiberg, 2017; Wu, 2010; Wu et al., 2016). When there are common items, in subsequent applications of the same test, the most common way to obtain common test scores is to use the fixed item calibration method (Kim, 2006; König et al., 2021; Zhao & Hambleton, 2017). This method consists of fixing parameters from a current test application, using known response model parameters from a previous test application. For example, if we use a sub-selection of items of an International Large Scale assessment study of reference, with known parameters, we can use these parameters and generate test scores in the metric of study of reference. Moreover, if a set of persons answer different tests in a single application, is possible to set the different tests assessing the same attribute in a common scale. This application is known as “horizontal equating” (Wu et al.,

2016). The Monitoring Impacts on Learning Outcomes (MILO) (ACER, 2022) and IEA's Rosetta Stone PASEC study (UNESCO UIS, 2022b), use these different methods to set the scale of their test scores in the metric of international large-scale study of reference[3]. However, it is important to keep in mind that these equating methods based on calibrated items assume (a) the use of a common item response model for both tests and (b) should be applied after examining if the properties of the items in the different tests remain empirically stable, as responses to items that are nominally identical in two tests could have prompted drastically different response behaviors in different populations.

In the following section, we discuss the main features of these study designs, to later on, assess the feasibility of representing the scores of the MILO study, into the TIMSS and PIRLS metric.

[3] In the following section each study will be discussed in more detail.

MILO and Rosetta-PASEC studies

In the following section, we summarize the research design of the Monitoring Impacts on Learning Outcomes (MILO) (ACER, 2022) and IEA's Rosetta Stone (UNESCO UIS, 2022b), here referred as "Rosetta-PASEC" for short. In this section we highlight the commonalities and the differences between studies, to bring in the relevant features of each study design, and to make a judgment call regarding the feasibility of the score harmonization.

Target Population

Both studies include participating countries from the sub-Saharan Africa. In Table 1, we include the countries participating in each study.

Table 1. Participating countries in the MILO and Rosetta-PASEC studies

	MILO	Rosetta-PASEC
Burkina Faso	yes	no
Burundi	**yes**	**yes**
Côte d'Ivoire	yes	no
Guinea	no	yes
Kenya	yes	no
Senegal	**yes**	**yes**
Zambia	yes	no

Note: countries participating in the study indicated in the column heading are flagged with a "yes" in the present table, while countries not participating in the respective study, are flagged with a "no".

The MILO study, assess representative samples of sixth graders, targeting students at the end of the primary school grade, during 2021. Similarly, The Rosetta-PASEC

study, assesses representative samples of students from grade six, in 2020, and after a year of PASEC 2019 (Neuschmidt, 2022). Thus, both studies assess the population of students in sixth grade, from sub-Saharan countries, and have produce data for two common countries at different moments: Burundi and Senegal[4].

Study purposes and instruments

The MILO study was design to fulfill different purposes, stated as (see ACER, 2022, p. 15):

- evaluate the impact learning of outcomes COVID-19 by on reporting reading and against SDG indicator 4.1.1b[5].
- identify the impact put in of place different to remediate distance the learning learning disruption generated by COVID-19
- education expand the UIS bank of items for primary
- to generate international a toolkit to benchmarks, scale assessment reporting results against SDG 4.1.1b.

For these purposes, the MILO study uses a collection of instruments to assess math and reading abilities of sixth graders. This collection of instrument includes the Assessments for Minimum Proficiency Levels (AMPL), a shorter version of PASEC 2019 here referred as PASEC 2021 link[6], and national assessment test in the case of Kenya and Zambia.

[4] We need to highlight that although Burundi and Senegal have participated in both initiative, is unlikely these two studies have a collection of responses from the same personas, to the three test of interest: MILO, PASEC, and TIMSS & PIRLS.

[5] The SDG indicator 4.1.1b measures the proportion of at the end of primary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.

[6] Although the PASEC 2021 Link is not exhaustively describe in the MILO report, regarding how many items it is stated that is shorter version and with reduced time of application (see ACER, 2022, p. 111).

In comparison, the Rosetta-PASEC study was design to develop concordance tables between PASEC scores and the TIMSS and PIRLS achievement scales in in francophone Sub-Saharan countries (UNESCO UIS, 2022a, p. 1). The aim of this concordance tables 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.*" (UNESCO UIS, 2022b, p. 6). For this purpose, the Rosetta-PASEC study included the application two sets of instruments. The first, is the 2019 PASEC assessment[7], and the Rosetta Stone linking booklets. These latter booklets are a sublecction of items from TIMSS and PIRLS assessment instruments, that are originally targeting students at grade 4th. Both sets of instruments allow the assessment of reading and math ability among sixth grade students. In table 2, we provide a summary of the instruments present in both studies.

Table 2. Instruments used in the MILO and Rosetta-PASEC studies

	MILO	Rosetta-PASEC
Burkina Faso	AMPL & 2021 PASEC LINK	no
Burundi	AMPL & 2021 PASEC LINK	Rosetta-Link & 2019 PASEC
Côte d'Ivoire	AMPL & 2021 PASEC LINK	no
Guinea	no	Rosetta-Link & 2019 PASEC
Kenya	AMPL & NASMLA 2019 (Grade 7)	no
Senegal	AMPL & 2021 PASEC LINK	Rosetta-Link & 2019 PASEC
Zambia	AMPL & NAS 2016 (Grade 5)	no

[7] For more details on 2019 PASEC assessment see PASEC

Note: AMPL = Assessments for Minimum Proficiency Levels; 2021 PASEC LINK = shorter version of 2019 PASEC assessment, and applied with a reduced time; Rosetta-Link = Item booklets of TIMSS and PIRLS assessment; 2019 PASEC = instruments develop to measure reading and math ability of students.

Comparing the available instruments of each study, it seems feasible to examine the stability of the item properties between samples and estimate item parameters for a given item response model from MILO and Rosetta-PASEC study for Burundi and Senegal, if the raw response data is made available from each study. As we discussed earlier, having common items on the different test applications would offer in this case a minimum base for generating scores of different test under a common metric. However, when using an equating procedure based on calibrated items it is necessary to assume a common item response model before a score transformation is undertaken.

Scoring Methods used in each study

The MILO study uses Rasch model, to generate scores for the AMPL instrument (ACER, 2022, p. 107), and for the 2021 PASEC LINK instrument. Moreover, the PASEC studies also uses a Rasch model (PASEC, 2015, 2020), thus equating scores between different applications with common items is easier. In contrast, the Rosetta-PASEC study uses a collection of different IRT models. The Rosetta-Link instruments, rely on the original IRT model that TIMSS and PIRLS uses, which uses IRT 3PL for multiple choice items, IRT 2PL for construct response items of 1 score point, and the GPCM from constructed response items of more than one point (UNESCO UIS, 2022a, p. 12). Whereas, in the case of the calibration of the item parameters of the 2019 PASEC assessment, it relies on the Rasch model, according to the PASEC study procedures (UNESCO UIS, 2022a, p. 16). In Table 3, we summarize the response model used in each study to generate the scores for each instrument.

Table 3. IRT models use to generate scores in the instruments presed in the MILO and Rosetta-PASEC studies

Study	Instrument	Rasch	IRT 2PL	IRT 3PL	GPCM
MILO	AMPL	yes	no	no	no
MILO	2021 PASEC LINK	yes	no	no	no
Rosetta PASEC	Rosetta-Link	no	yes	yes	yes
Rosetta PASEC	2019 PASEC	yes	no	no	no

Note: AMPL = Assessments for Minimum Proficiency Levels; 2021 PASEC LINK = shorter version of 2019 PASEC assessment, and applied with a reduced time; Rosetta-Link = Item booklets of TIMSS and PIRLS assessment; 2019 PASEC = instruments develop to measure reading and math ability of students. Rasch = refers to the Rasch model, or one parameter logistic model. In this model, item location (diifculties) and person parameters are orthogonal; IRT 2PL = item response theory model, where item location (difficulties), slopes (discrimination) and person locations (ability) are modeld; IRT 3PL = item response theory model, where item location, slopes, intercepts (guessing parameter) and person locations are modeled; GPCM =

generalized partial credit model, where item location and item slopes are modeled for items with more than two ordered categories.

How the SDG 4.1.1b was obtained and reported in each study

The MILO study uses a standard-setting exercise to derive cut scores to report minimum proficiency levels at the end of primary schooling (ACER, 2022, p. 112). As such, the ability of the MILO study to report SDG 4.11.1b doesn't need to rely on transforming the scores generated with the AMPL instrument to the metric of the PASEC study. However, because the purpose of the study was to compare previous results with current SDG 4.1.1b results for similar cohorts of students, it reports the minimum proficiency level (MPL) using the historical assessment, using the PASEC scores, and/or national assessment correspondingly, converted to the AMPL scale (ACER, 2022, p. 112). In essence, the MPL cut scores generated for the AMPL test to report the SDG 4.1.1b, are also calculated for the historical assessment (PASEC or National Study) using horizontal equating (ACER, 2022, p. 111), thus using a "common person, different tests" design (González & Wiberg, 2017; Wu et al., 2016). The following figures, reproduced the tables presenting the MPL in the MILO study report.

Figure 1. Estimated proportion of students reaching the SDG 4.1.1b (reading) using the AMPL and historical assessment in the MILO study

TABLE C.3 Proportions of students who met or exceeded SDG-aligned MPLs for reading with standard errors

	STUDENTS WHO REACHED OR EXCEEDED MPL IN 2021 AMPL: READING (%)			STUDENTS WHO REACHED OR EXCEEDED MPL IN HISTORICAL ASSESSMENT: READING (%)		
Country	**All**	**Boys**	**Girls**	**All**	**Boys**	**Girls**
Burkina Faso	9.0 (1.50)	9.3 (1.85)	8.8 (1.50)	5.8 (0.91)	5.6 (1.00)	5.9 (0.98)
Burundi	**0.1 (0.09)**	**0.1 (0.15)**	**0.1 (0.09)**	**0.3 (0.21)**	**0.3 (0.18)**	**0.4 (0.28)**
Côte d'Ivoire	10.8 (1.33)	9.9 (1.38)	11.7 (1.76)	10.4 (1.59)	9.9 (1.70)	10.9 (1.71)
Kenya	46.7 (2.33)	44.9 (2.45)	48.4 (2.56)			
Senegal	13.3 (1.81)	11.6 (1.75)	14.6 (2.22)	14.7 (2.31)	14.1 (2.31)	15.2 (2.65)
Zambia	2.3 (0.81)	2.4 (1.01)	2.2 (0.72)	1.8 (0.41)	**1.5 (0.42)**	2.1 (0.53)

Standard errors (SE) are reported in brackets.
Statistics in bold are from fewer than 30 students and/or 5 schools.

Figure 2. Estimated proportion of students reaching the SDG 4.1.1b (mathematics) using the AMPL and historical assessment in the MILO study

TABLE C.4 Proportions of students who met or exceeded SDG-aligned MPLs for mathematics with standard errors

	STUDENTS WHO REACHED OR EXCEEDED MPL IN 2021 AMPL: MATHEMATICS (%)			STUDENTS WHO REACHED OR EXCEEDED MPL IN HISTORICAL ASSESSMENT: MATHEMATICS (%)		
Country	**All**	**Boys**	**Girls**	**All**	**Boys**	**Girls**
Burkina Faso	23.7 (1.83)	25.8 (2.23)	22.1 (1.95)	17.9 (1.52)	18.8 (1.70)	17.1 (1.58)
Burundi	13.5 (1.83)	16.5 (2.23)	11.1 (1.91)	17.0 (1.73)	22.0 (2.06)	12.9 (1.66)
Côte d'Ivoire	8.9 (1.24)	8.8 (1.29)	9.1 (1.64)	7.6 (1.27)	8.2 (1.38)	6.9 (1.37)
Kenya	74.1 (1.90)	73.5 (2.08)	74.6 (2.06)	79.7 (3.18)	82.8 (4.06)	78.4 (3.26)
Senegal	34.0 (2.33)	34.1 (2.64)	33.9 (2.55)	34.6 (2.87)	34.6 (3.10)	34.7 (3.07)
Zambia	2.1 (0.78)	**2.0 (0.90)**	2.1 (0.77)	3.5 (0.56)	3.7 (0.61)	3.4 (0.68)

Standard errors (SE) are reported in brackets.
Statistics in bold are from fewer than 30 students and/or 5 schools.

In comparison, the Rosetta-PASEC study uses the TIMSS and PIRLS metric to report on the SDG 4.1.1b for math and reading. It uses two different methods for this purpose. The first, is the Rosetta Link booklets. These are a subselection of items from TIMSS and PIRLS, selected to assess the SDG 4.1.1b. By using the fixed item calibration method (Kim, 2006; König et al., 2021; Zhao & Hambleton, 2017) is possible to generate test scores using the responses to the Rosetta-Link booklets, in the scale of the TIMSS and PIRLS metric. The second method, is to project the PASEC scores onto the TIMSS and PIRLS metric, using the concordance method. This is not a linear transformation of scores, is not a direct link between test, but a projection of a source test, onto a target test using a prediction method (UNESCO UIS, 2022a, p. 30).

Figure 3. Estimated proportion of students reaching the SDG 4.1.1b for mathematics and reading, using the Rosetta Link booklets, and using the concordance method

Exhibit 11.8: Estimated Percentages of Students Reaching the TIMSS and PIRLS Low (400) International Benchmarks

Estimated Percentages based on Rosetta Stone		
Country	**TIMSS (400)**	**PIRLS (400)**
Burundi	8.9 (1.1)	4.1 (0.7)
Guinea	16.6 (2.0)	19.7 (2.1)
Senegal	47.5 (3.7)	41.0 (3.9)
Average	**24.3 (1.5)**	**21.6 (1.5)**
Estimated Percentages based on Concordance		
Country	**TIMSS (400)**	**PIRLS (400)**
Burundi	29.3 (1.8)	10.5 (1.0)
Guinea	10.1 (1.7)	15.6 (2.2)
Senegal	34.6 (3.1)	36.3 (3.4)
Average	**24.7 (1.3)**	**20.8 (1.4)**

Note: Standard errors appear in parentheses.

The estimated proportions of students reaching the SDG 4.1.1b target in the MILO study from each method are similar. Thus, regardless of using the AMPL direct method; or the historical data from PASEC assessment, the MILO leads to similar substantive conclusions. In contrast, the substantive conclusions made with the Rosetta-PASEC study are very different. The estimated proportion for Guinea and Senegal are smaller using the concordance table, in contrast to using the Rosetta-Link in Math and Reading; while the opposite is true for Burundi. Moreover, the expected average of the three countries are more similar between the two methods.

Feasibility of score harmonization

In the following section, we assess the feasibility to harmonize the scores from the MILO study to the TIMSS and PIRLS metric, and its implications for reporting the SDG 4.1.1b. We assess three different options: extrapolation, projecting scores after a Rosetta Stone Study, and an equating score study.

Extrapolation or Direct projection using the concordance tables

One alternative is to generate projections to the TIMSS and PIRLS scores, using the concordance tables, relying on available PASEC scores. In the MILO study, we have at least two countries with PASEC scores, scaled to the 2019 PASEC scores. Using the concordance tables, one could generate plausible values into the TIMSS and PIRLS metric. Although this is possible, this operation is not a transformation of scores, but a projection of scores using a predictive model. As such, the expected scores will include prediction uncertainty which be larger the less correlated are the source test (PASEC scores), to the target test (TIMSS and PIRLS). This alternative is not recommended by the IEA team due to the uncertainty added to the estimates and recommended to conduct a Rosetta Stone study first (UNESCO UIS, 2022a, p. 46). Moreover, the observed difference of the SDG 4.1.1b estimated proportions per country, conditional to the method use, prevents to trust the concordance and Rosetta-Link as interchangeable methods. The comparison of methods does not lead to the same substantive conclusion at the country level.

It seems the methods used by the MILO study are more straightforward, to produce SDG 4.1.1b results. It reaches similar estimated proportions, regardless of the test scores used to estimate the SDG 4.1.1b results., at different moments (pre and post-COVID-19 pandemic). Yet, it should be noted the minimum proficiency levels (MPL) from the MILO Study to estimate SDG 4.1.1b, is not equivalent to SDG 4.1.1b from the Rosetta-PASEC studies. Although both studies aim to report the same indicator, both are using different rationales to retrieve the same estimation, and are not interchangeable.

Projecting scores after a Rosetta Stone Study

A second option is to project scores after a common person design. That is, to have a representative sample of students answering the AMPL booklets, and simultaneously the Rosetta-Link booklets. With this latter design, is possible to create concordance tables between the source test (AMPL) and the target test (TIMSS and PIRLS), while accounting for common person variance.

In this alternative, the most sensible expectation to have is to find similar results from those found in the Rosetta-PASEC study. That is, to have two methods allowing to report SDG 4.1.1b, yet with the risk of reaching non-convergent results.

Equating using a non-equivalent groups and common items design

Finally, a third possible option to consider to generate harmonization of scores between the AMPL instrument and the TIMSS and PIRLS metric, is to use a fixed item parameter calibration method (Kim, 2006; König et al., 2021; Zhao & Hambleton, 2017). This method consists of estimating the item parameters of a portion of a test, while fixing the item parameters of a different portion of the test, to a known reference group. The application of the current method needs to be done in different steps. First, it needs to obtain the item parameters of the PASEC assessment, while fixing the item parameters of the TIMSS and PIRLS in the Rosetta-PASEC study, for the samples of Burundi and Senegal. Then, those PASEC assessment items parameters would be fixed in the TIMSS and PIRLS scale. In the second step, one would need to estimate freely the parameters of AMPL estimates while fixing the PASEC assessment items parameters, using the parameters used in the first step. This application, however, has certain assumptions. First, the application assumes that all item parameters come from the same model. This is assumption is violated, due to the fact that the TIMSS and PIRLS scores are generated by a set of different response models, the IRT 2pl, IRT 3pl and the GPCM. A second assumption of this

application, it requires item parameter invariance between groups (Robitzsch, 2021), that is between the MILO study and the Rosetta-PASEC samples (Burundi and Senegal). This assumption at least can be tested. A third assumption is that this method assumes unidimensionality between the AMPL, Rosetta-Link, and PASEC items. This later assumption cannot be tested with the current available studies (MILO study, and Rosetta-PASEC).

This third option, although attractive is not feasible. Its feasibility gets trumped by the fact that TIMSS and PIRLS scores are generated with a different item response model. Fixed parameter calibration are sensitive to model misfit, that is deviations of items to the expected model (Zhao & Hambleton, 2017). The items in the Rosetta-Link booklets were selected using IRT 2PL, IRT 3PL and GPCM, while the items present in the AMPL and PASEC instruments are retained using a Rasch Model. Thus, there is no guarantee of item model fit, given that these different test were built with different response models. The unidimensionality between the three tests cannot be assessed, because this requires a “common person different test design”. In other words, we would beed a study where the same group of students answered MILO, Rosetta Link and PASEC test, to assess if the responses to these three test present unidimensioanality. In summary, the current scenario doesn’t fit the expected conditions to bridge item parameters between the three tests (AMPL, PASEC, Rosetta-Link).

Summary and conclusions

Summary

Although is possible to create score transformation between test forms, generating comparable results for tests that are scored using markedly different statistical models presents additional challenges. In the MILO, and in the Rosetta-PASEC study, both studies do equate scores from a form to an international study of reference. The MILO study uses a shorter form of the PASEC assessment, the 2021 PASEC LINK, and set test scores into the PASEC scale. Similarly, the Rosetta-PASEC study uses the Rosetta-Link booklets and can set the test scores of these instruments into the TIMSS and PIRLS scales. These two exercises rely on a common design: using common items onto non-equivalent groups while fixing item parameters in the response model before generating the test score logits, with the same response model of the test.

The second operation is more complicated. Expressing test scores in the metric of a different test is only possible if the two tests are actually similar forms of the same test. If the necessary conditions are not present to incure in equating procedures, score projections seems attractive as it is the case of the Rosetta-PASEC study. Moreover, test score projections do not seem to retrieve the same estimates for SDG 4.1.1b for single countries; in contrast to the fixed item calibration method.

Fixed item calibration seems an attractive choice to bridge the test scores of the AMPL, PASEC, and Rosetta-Link booklets. Yet, the currently available information from the MILO study (ACER, 2022), and the Rosetta-PASEC study (UNESCO UIS, 2022a) appears to indicate that the necessary conditions are not met for this method to be directly applicable. In addition to the necessary checks for model fit and item stability

that would need to be empirically examined, the response models used to generate the test score of each of these studies are not the same, and therefore the simple application of shared calibrated items would not resolve the problem. In addition to the previous technical challenges it is also important to consider that an underlying assumption in any and all of these procedures is that all the test are unidimensional and that they are all measuring the same attribute of interest.

Conclusions

After reviewing the available information of the MILO study in its report (ACER, 2022), and the Rosetta-PASEC study report (UNESCO UIS, 2022a) we believe is not feasible to harmonize the AMPL scores in a sensible manner, into the TIMSS and PIRLS scores. Even if this harmonization is done after a full Rosetta Stone Study, there is a risk of producing estimates of SDG 4.1.1b using the concordance table method, not convergent to the estimates generated using the Rosetta-Link scale test scores.

We believe is not sensible to transform the AMPL scores into the TIMSS and PIRLS metric to report the SDG 4.1.1b indicator. If the main aim is to generate SDG 4.1.1b estimates, the AMPL and its standard settings can provide them without the need to transform the scores to other international test scales of reference. The Rosetta Stone application generate estimates of the SDG 4.1.1b indicator via two methods. The first, is using the Rosetta Link test application, and thorugh fixed item calibration it reacheas SDG 4.1.1b results using TIMSS & PIRLS benchmarks. The second method, is to generate SDG 4.1.1b via a projection of scores from a source test (e.g., PASEC), onto the test of reference (TIMSS & PIRLS) via concordance tables. Both methods

requires the application of the source test, and the Rosetta Link tests to a country where these estimates are of interest. As such, MILO and Rosetta initiatives requires field studies for countries interested in generating results for the SDG 4.1.1b indicator.

If both initiatives, MILO and Rosetta Stone presents the same costs for any new participating country willing to produce results for the SDG 4.1.1b indicator, then one should judge the trustworthiness of each process. Because Rossetta Stone produce non-convergent results with two methods onto the same sample, we believe the MILO design seems a more straightforward method to produce results for the SDG 4.1.1b population estimates, without the need to use the TIMSS and PIRLS benchmarks.

References

ACER. (2022). *COVID-19 in Sub-Saharan Africa: Monitoring Impacts on Learning Outcomes*. UNESCO Institute for Statistics. https://learningportal.iiep.unesco.org/en/library/covid-19-in-sub-saharan-africa-monitoring-impacts-on-learning-outcomes-main-report

Adams, R. J., & Wu, M. L. (2000). PISA 2000 Technical Report. In R. J. Adams & M. L. Wu (Eds.), *PISA 2000 Technical Report*. OECD Organisation for economic co-operation and development. https://www.oecd.org/pisa/data/33688233.pdf

González, J., & Wiberg, M. (2017). *Applying Test Equating Methods*. Springer International Publishing. https://doi.org/10.1007/978-3-319-51824-4

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381. https://doi.org/10.1111/j.1745-3984.2006.00021.x

König, C., Khorramdel, L., Yamamoto, K., & Frey, A. (2021). The Benefits of Fixed Item Parameter Calibration for Parameter Accuracy in Small Sample Situations in Large-Scale Assessments. *Educational Measurement: Issues and Practice*, *40*(1), 17–27. https://doi.org/10.1111/emip.12381

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). Methods and procedures in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center* (Vol. 21). TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Neuschmidt, O. (2022). *IEA's Rosetta Stone: Implementation and Results* (Issue November). UNESCO Institute for Statistics. https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/11/WG_GAML_7_2211_IEA_Rosetta_GAML.pdf

PASEC. (2015). *PASEC 2014. Education system perfomance in Francophone Sub-Saharan Africa. Competencies and Learning Factors in Primary Education*. Programme d'Analyse des Systèmes Educatifs de la CONFEMEN. https://www.pasec.confemen.org/wp-content/uploads/2015/12/Rapport_Pasec2014_GB_webv2.pdf

PASEC. (2020). *PASEC 2019: Quality of Education systems in Frecnh-Speacking Sub-Sharan Africa*. Programme for the Analysis of Education Systems of CONFEMEN,.

Robitzsch, A. (2021). A Comparison of Linking Methods for Two Groups for the Two-Parameter Logistic Item Response Model in the Presence and Absence of Random Differential Item Functioning. *Foundations*, *1*(1), 116–144. https://doi.org/10.3390/foundations1010009

UNESCO UIS. (2022a). *Rosetta Stone Analysis Report: Establishing a Concordance between PASEC and TIMSS/PIRLS*. UNESCO Institute for Statistics.

https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/07/Rosetta-Stone_PASEC_Analysis-Report_2022.pdf
UNESCO UIS. (2022b). *Rossetta Stone Policy Brief. Establishing a concordance between regional (ERCE and PASEC) and international assessments*. UNESCO Institute for Statistics. https://tcg.uis.unesco.org/wp-content/uploads/sites/4/2022/07/Rosseta-Stone-Policy-Brief-EN-WEB.pdf
UNESCO-OREALC. (2016). *Reporte Técnico. Tercer Estudio Regional Comparativo y Explicativo, TERCE.*
Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27. https://doi.org/10.1111/j.1745-3992.2010.00190.x
Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for Applied Researchers*. Springer Singapore. https://doi.org/10.1007/978-981-10-3302-5
Zhao, Y., & Hambleton, R. K. (2017). Practical consequences of item response theory model misfit in the context of test equating with mixed-format test data. *Frontiers in Psychology*, *8*(MAR), 1–11. https://doi.org/10.3389/fpsyg.2017.00484