Analysis Report:

Establishing a Concordance between PASEC and TIMSS/PIRLS

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 11 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 fulfill 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.

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/RR/09

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 and the TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College is the author of this report. Technical and implementation support was provided by CONFEMEN and LLECE.

The UIS would like to thank the report writers, Lale Khorramdel, Liqun Yin, Pierre Foy, Ji Yoon Jung, Ummugul Bezirhan and Matthias von Davier. Silvia Montoya (UIS), Dirk Hastedt (IEA), and Oliver Neuschmidt (IEA) served as reviewers for this report.

For more information about TIMSS contact TIMSS & PIRLS International Study Center https://timssandpirls.bc.edu/

1. Summary 1
2. Introduction 2
3. Rosetta Stone Instruments and Test Design 3
4. Analysis Overview and Sample 5
5. Data Quality Evaluation 6
6. IRT Models 11
   6.1 IRT Scaling in Large-Scale Assessments 11
   6.2 IRT Models for Dichotomous Items: Rasch Model, 2PL Model and 3PL Model 12
   6.3 IRT Model for Polytomous Items: GPCM 13
   6.4 Unidimensionality 13
   6.5 Conditional Independence 14
   6.6 Monotonicity of Item-Proficiency Regressions 14
   6.7 Multidimensional IRT Models. 16
7. IRT Model Application to PASEC and Rosetta Stone Data 16
   7.1 Establishing Comparability through IRT Scaling 16
   7.2 Results for Unidimensional IRT Models 19
   7.3 Results for Multidimensional IRT Models 23
8. Population Models 23
   8.1 Integrating Achievement Data and Context Information 23
   8.2 Group-Level Proficiency Distributions and Plausible Values 24
9. Population Model Application to PASEC and Rosetta Stone Data 26
   9.1 Applied Population Models. 26
   9.2 Generating Plausible Values and PASEC Score Validation 27
   9.3 Transforming the Plausible Values to TIMSS and PIRLS Scales 29

10. Establishing an Enhanced Concordance between Scales 29
10.1 Predictive Mean Matching (PMM) 30
10.2 Technical Procedure for Establishing Concordance Tables. 31
10.3 Advantages of the Enhanced Concordance Method 33
11. Establishing an Enhanced Concordance between PASEC and TIMSS/PIRLS 33
11.1 Relationship between PASEC data and Rosetta Stone Linking data 34
11.2 Creating Preliminary Concordance Tables 36
11.3 Smoothing and Extrapolating the Concordance Tables. 37
12. How to Use and Interpret the Concordance Tables. 43
References 45
Appendix A: Example of Generated PVs based on the Concordance Table for PASEC Mathematics 49
Appendix B: Example of Generated PVs based on the Concordance Table for PASEC Reading. 55
Appendix C: Using the Rosetta Stone Concordance Tables – Analysis Steps 61
Analysis Steps. 62

ROSETTA STONE ANALYSIS REPORT:
Establishing a Concordance between PASEC and TIMSS/PIRLS

Lale Khorramdel, Liqun Yin, Pierre Foy, Ji Yoon Jung,
Ummugul Bezirhan and Matthias von Davier

timssandpirls@bc.edu

1. Summary

This report is concerned with establishing a concordance between the regional PASEC and the international TIMSS and PIRLS achievement scales in francophone Sub-Saharan countries.

The Rosetta Stone study consists of two assessment parts. The first part is the PASEC assessment including the PASEC context questionnaire. The second part is the Rosetta Stone assessment comprising test booklets with easier item blocks and passages from TIMSS and PIRLS. Both assessment parts were administered in three PASEC countries to the same students on two consecutive days.

Analyses were conducted using classical item statistics, item response theory (IRT) and population modeling. They comprise the evaluation of the data quality, evaluation of the psychometric quality of the instruments, establishing common scales across countries and assessments, and constructing concordance tables which account for the uncertainty of the measurement (measurement error).

The key findings can be summarized as follows:

- The difficulty of the selected TIMSS and PIRLS item blocks and passages are appropriate for the Rosetta Stone analysis and the goals of the study.
- Comparable PASEC and Rosetta Stone IRT scales could be established across countries.
- Comparable IRT scales could be established across Rosetta Stone and TIMSS/PIRLS.
- Latent correlations in multidimensional IRT models between PASEC mathematics and TIMSS (r = .81-.86) and PASEC reading and PIRLS (r = .78-.83) suggest that constructs are not identical but similar enough to enable a concordance.
- Population models were able to be estimated providing proficiency distributions for PASEC and Rosetta Stone scales.

- Plausible values (PVs) for PASEC scales were imputed independently by the TIMSS & PIRLS International Study Center based on the Rosetta Stone study data for validation purposes. They were found to be highly correlated to the PVs provided by the PASEC team (PASEC mathematics: r = .96-.97; PASEC reading: r = .96-.98) indicating very good agreement of analytic processes.
- Moreover, country means based on PVs for PASEC scales provided by the PASEC team were compared to country means based on the published PASEC 2019 PVs and were found to be very similar. This is indicating that the Rosetta Stone student sample is comparable to the PASEC 2019 student sample.
- Population models were applied to the Rosetta Stone data to obtain posterior means and PVs for TIMSS numeracy and PIRLS literacy.
- Estimates from both assessments, PASEC and Rosetta Stone, were used to establish concordance tables that provide a conditional distribution on the TIMSS and PIRLS scales for a range of PASEC score levels.
- The concordance should be used with care, being aware of the limitations of country participation and sample sizes, and differences between assessments.
- The concordance provides a projection and not a direct linking of scales. However, when used and interpreted properly, concordance tables can provide useful and valuable information by comparing regional assessment results with international benchmarks.
- New countries seeking a concordance between PASEC and TIMSS and PIRLS are encouraged to participate in a Rosetta Stone study first.

The following sections in this report describe the instruments and design of the Rosetta Stone linking study, the psychometric analyses, and the construction of the concordance tables as well as their limitations and appropriate use and interpretation.

2. Introduction

IEA's Rosetta Stone study is designed to measure global progress toward the United Nations (UN) Sustainable Development Goal 4 for quality in education (SDG 4, Target 4.1) by relating different regional assessment programs to TIMSS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study) international long-standing metrics and benchmarks of achievement . The goal is to provide participating countries, who participated in regional assessments but not in TIMSS and PIRLS, with information about the proportions of primary school students that have achieved established international proficiency levels in literacy and numeracy for allowing international comparisons.

This analysis report describes the study, methods, and analysis conducted to establish a concordance between the Programme for the Analysis of Education Systems (PASEC; Programme d'Analyse des Systèmes Éducatifs) in francophone Sub-Saharan countries and TIMSS and PIRLS. PASEC assesses student achievement in mathematics, reading, and listening comprehension at grades two and six (i.e., at the beginning and end of primary) and is conducted by the Conference of Ministers of Education of French-Speaking Countries (CONFEMEN).

To construct the concordance, the 2019 PASEC assessment was administered to students at the sixth grade together with the Rosetta Stone linking booklets that contained items from TIMSS and PIRLS. The content of PASEC's mathematics assessment was expected to align well with the TIMSS fourth grade assessments in numeracy and mathematics. Similarly, the content of PASEC's reading assessments was expected to align with the PIRLS fourth grade assessment in literacy and reading comprehension. The TIMSS & PIRLS International Study Center at Boston College was responsible for the development of the Rosetta Stone assessment, the psychometric analysis, and the establishment of the concordance tables.

The overarching goal is to construct a concordance table that projects the score distributions estimated from the PASEC mathematics and reading assessments to distributions on TIMSS and PIRLS, respectively. The concordance table would therefore represent the "Rosetta Stone", analogous to the original Rosetta Stone which provided a link between Greek and Egyptian hieroglyphics, that enables a translation between the countries' regional assessment results and the TIMSS and PIRLS achievement scales. Countries participating in the regional assessments can then use the translations to estimate percentages of their students that could be expected to reach the TIMSS and PIRLS international benchmarks.

The Rosetta Stone study for PASEC is a collaborative project between the UNESCO Institute for Statistics (UIS), the PASEC study center (CONFEMEN), IEA, and the TIMSS & PIRLS International Study Center at Boston College, as well as the national teams of the participating countries Burundi, Guinea, and Senegal. Questions about linking design, the data analyses, and the report for the Rosetta Stone study for PASEC should be directed to the TIMSS & PIRLS International Study Center at Boston College (timssandpirls@bc.edu).

3. Rosetta Stone Instruments and Test Design

One of the major goals and design principles of large-scale surveys of student achievement is to provide valid comparisons across student populations based on broad coverage of the achievement domain. This usually translates into a large number of achievement items, only a fraction of which can be administered to any one student given the available testing time. Therefore, Rosetta Stone is based on a matrix-sampling booklet design where each student was administered only a subset of the selected item pools. Moreover, a subset of less difficult TIMSS and PIRLS item blocks and passages was used to best target the difficulty of the assessment for participating countries. The Rosetta Stone study comprises two assessment parts. The first part is the PASEC assessment including the PASEC achievement items and PASEC context questionnaire. The second part is the centerpiece of the study, the Rosetta Stone assessment part consisting of test booklets with easier TIMSS item blocks and passages and easier PIRLS passages. More precisely, items come from TIMSS Numeracy 2015, TIMSS 2019 less difficult (LD) and PIRLS Literacy 2016. In total, eight less difficult mathematics item blocks and four literacy passages were selected. Exhibit 3.1 provides the number of items and source for each item block and passage. Both assessment parts were administered as paper-based assessments to the same students. Each student was administered one PASEC booklet on the first day and one Rosetta Stone booklet on the second day. A description of the PASEC 2019 booklet design can be found in the related PASEC 2019 report (PASEC, 2020).

Exhibit 3.1: Rosetta Stone Linking Item Blocks and Passages

	Source	Number of Items
TIMSS Blocks		
N01	TIMSS Numeracy 2015 – N01	13
N02	TIMSS 2019 LD – MN04	14
N03	TIMSS 2019 LD – MN07	13
N04	TIMSS 2019 LD – MN05	13
N05	TIMSS 2019 LD – MN01	13
N06	TIMSS 2019 LD – MN14	14
N07	TIMSS 2019 LD – MN03	13
N08	TIMSS 2019 LD – MN09	12
	Total TIMSS Items	**105**
PIRLS Passages		
L01	PIRLS Literacy 2016 – *Baghita's Perfect Orange* (Literary)	16
L02	PIRLS Literacy 2016 – *African Rhinos* (Informational)	17
L03	PIRLS Literacy 2016 – *The Pearl* (Literary)	15
L04	PIRLS Literacy 2016 – *Ants* (Informational)	14
	Total PIRLS Items	**62**

Exhibit 3.2 illustrates the design for the Rosetta Stone assessment part, which was arranged into eight linking booklets. Each block or passage appeared twice in a balanced incomplete block design. The numeracy blocks appeared in different positions (at the beginning or the end of a booklet) to counterbalance possible position effects. Students had 40 minutes to complete each part of the linking booklet, with a short break in between.

Exhibit 3.2: Rosetta Stone Linking Booklet Design

Booklet	Part 1		Part 2	
1	N01	N02	L01	
2	L02		N02	N03
3	N03	N04	L03	
4	L04		N04	N05
5	N05	N06	L02	
6	L03		N06	N07
7	N07	N08	L04	
8	L01		N08	N01

4. Analysis Overview and Sample

To establish concordance tables, the analysis of the data proceeded in four steps. These steps are briefly described here and then in more detail in sections 5 to 11. First, data quality was evaluated based on classical item statistics and an analysis of nonresponse (section 5). Second, IRT models were used to further examine the psychometric quality of the assessment booklets and for constructing comparable PASEC and Rosetta Stone scales across student populations (sections 6 and 7). Third, population models were used to impute plausible values (PVs) separately for PASEC and Rosetta Stone (sections 8 and 9). Fourth, concordance tables were established based on posterior means and PVs from the population models (sections 10 and 11). The analysis was performed on data from three PASEC countries using sample weights provided to the TIMSS & PIRLS International Study Center.

Exhibit 4.1 provides the sample sizes for each country available for the scaling and population modeling. Cases with sample weights and responses to achievement items (PASEC items, Rosetta Stone items, or both) were included in the analysis while cases with responses only to the PASEC context questionnaire items were excluded. The sample size and number of schools per country in the Rosetta Stone study are smaller in comparison to the full TIMSS and PIRLS assessments where the approximate minimum sample includes 150 schools and 4,500 students for most countries.

Exhibit 4.1: Rosetta Stone Sample Sizes per Country

Country	Number of Students	Number of Schools	Number of Classes
Burundi	2,304	100	100
Guinea	2,252	100	100
Senegal	2,072	99	99
Total	**6,628**	**299**	**299**

The main goal of the IRT scaling was to establish comparable scales across countries and across the Rosetta Stone and the TIMSS/PIRLS assessments as the basis for a concordance. While PASEC items were already calibrated by the PASEC team, which also provided the PVs for PASEC, the TIMSS & PIRLS International Study Center performed IRT scaling and population modeling for the Rosetta Stone linking items. For validation and replication purposes, the PASEC items were re-calibrated as well. The following IRT models were estimated:

1. *Comparability of PASEC items across countries:* For evaluating the psychometric properties and cross-country comparability of the PASEC items, common item parameters were estimated across countries and item fit statistics were examined for all item-by-country combinations. Resulting item parameters were used to replicate and validate the PASEC PVs that were received from the PASEC team.

2. *Comparability of linking items across countries and assessments*: To achieve comparable scales across Rosetta Stone and TIMSS/PIRLS, item parameters for linking items were borrowed from TIMSS and PIRLS and fixed in the analysis for all countries. Item fit was examined for all item-by-country combinations.

3. *Comparability of PASEC and Rosetta Stone constructs*: Through multidimensional IRT models, latent correlations between PASEC and Rosetta Stone scales were estimated to evaluate whether the PASEC mathematics and reading scales are sufficiently similar to the TIMSS and PIRLS scales for establishing a meaningful concordance between them.

The estimated item parameters from the IRT scaling were used in the population models together with context variables from the PASEC background questionnaire for imputing PVs. The population modeling was performed at the country-level and separately for PASEC and Rosetta Stone linking data. After the comparability and accuracy of the population modeling approaches used in PASEC and in the Rosetta Stone study was confirmed (by re-estimating the PASEC PVs), the posterior means and PVs from the population models were used for constructing concordance tables, one for reading and one for mathematics. Sections 6 to 9 provide a more detailed description of all IRT and population models, and their application to the Rosetta Stone and PASEC data.

5. Data Quality Evaluation

Data quality was evaluated using classical item statistics (percent correct and item-total correlations) and examining item-level nonresponse variability. Exhibits 5.1 and 5.2 provide the average percent of correct responses and the average item-total correlation for each Rosetta Stone and PASEC item block and passage by country. The percent of correct responses show that the TIMSS and PIRLS item blocks and passages are more difficult for the PASEC population than the PASEC mathematics and reading item blocks. The item-total correlations indicate that TIMSS and PIRLS item blocks and passages exhibit similar medium discriminations as PASEC item blocks.

Exhibit 5.1: Average Item Difficulty (percent correct) and Discrimination (point-biserial correlation) by Item Block/Passage and Country for Reading/Literacy

	Burundi		Guinea		Senegal	
Item Block/Passage	Average Percent Correct	Average Point-Biserial	Average Percent Correct	Average Point-Biserial	Average Percent Correct	Average Point-Biserial
Rosetta Stone PIRLS Literacy						
L01	33.4	0.29	44.9	0.40	56.0	0.38
L02	27.3	0.33	34.5	0.40	53.7	0.45
L03	26.9	0.30	34.0	0.41	48.3	0.41
L04	29.3	0.34	40.2	0.44	54.9	0.48
Average	**29.2**	**0.32**	**38.4**	**0.41**	**53.2**	**0.43**
PASEC Reading						
RA	49.7	0.29	51.1	0.44	69.6	0.41
RB	46.9	0.32	55.7	0.46	73.0	0.43
RC	54.6	0.34	55.0	0.49	73.0	0.44
RD	54.0	0.36	52.1	0.45	70.5	0.41
Average	**51.3**	**0.33**	**53.5**	**0.46**	**71.5**	**0.42**

Exhibit 5.2: **Average Item Difficulty (percent correct) and Discrimination (point-biserial correlation) by Item Block and Country for Mathematics/Numeracy**

Item Block	Burundi		Guinea		Senegal	
	Average Percent Correct	Average Point-Biserial	Average Percent Correct	Average Point-Biserial	Average Percent Correct	Average Point-Biserial
Rosetta Stone TIMSS Numeracy and LD						
N01	35.7	0.33	32.9	0.34	54.5	0.39
N02	22.2	0.31	21.1	0.27	34.5	0.40
N03	38.3	0.35	34.8	0.38	49.9	0.45
N04	26.4	0.33	33.2	0.41	49.7	0.41
N05	38.5	0.28	35.3	0.33	51.9	0.35
N06	35.7	0.24	31.9	0.26	48.5	0.35
N07	31.4	0.28	32.8	0.29	44.8	0.38
N08	40.2	0.27	32.7	0.31	47.7	0.35
Average	**33.6**	**0.30**	**31.8**	**0.32**	**47.7**	**0.39**
PASEC Mathematics						
MA	54.4	0.32	40.3	0.30	55.6	0.37
MB	58.6	0.34	40.6	0.33	64.9	0.40
MC	57.0	0.31	39.7	0.28	58.2	0.34
MD	55.8	0.33	38.5	0.33	54.8	0.37
Average	**56.5**	**0.33**	**39.8**	**0.31**	**58.4**	**0.37**

Exhibits 5.3 and 5.4 illustrate the average item difficulty (P+) by item block and passage averaged across countries for PIRLS literacy and PASEC reading and for TIMSS numeracy and PASEC mathematics, respectively. In both figures the blue dots indicate the average P+ for the specific item blocks and passages while the red line marks the 50% level as means of comparison. Both figures, as well as the table in Exhibit 5.1, show that TIMSS and PIRLS item blocks and passages tend to be somewhat more difficult than PASEC item blocks and passages within and across countries, but that the difficulty is at an appropriate level for the Rosetta Stone analyses.

Exhibit 5.3: Average Item Difficulty (percent correct) by Item Block/Passage for PIRLS Literacy and PASEC Reading

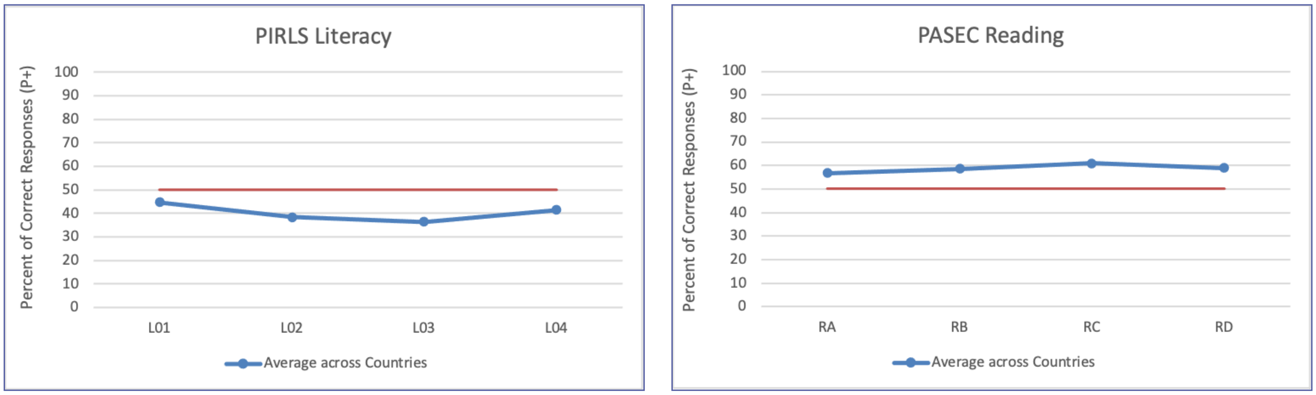
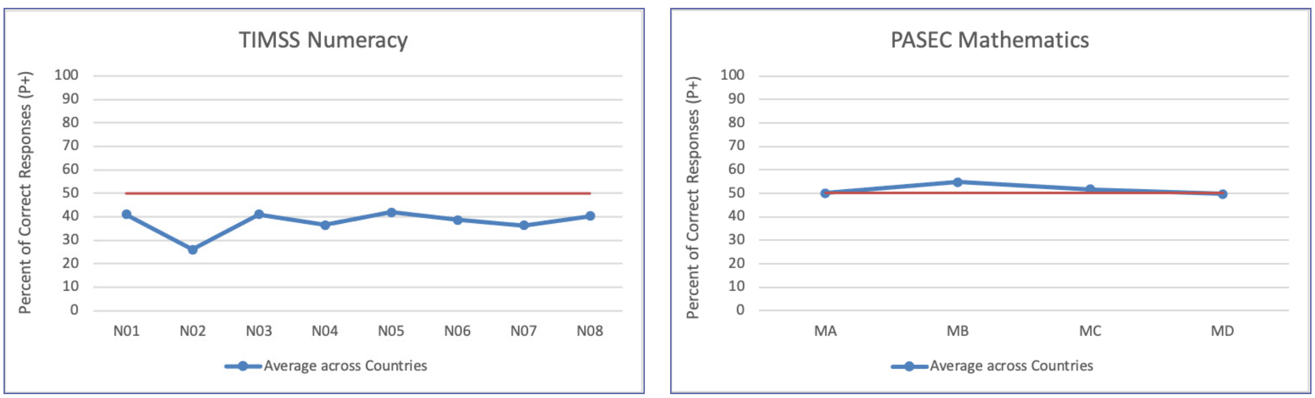


Exhibit 5.4: Average Item difficulty (percent correct) by Item Block for TIMSS Numeracy and PASEC Mathematics



Exhibits 5.5 and 5.6 illustrate the average percent of omitted (OM) and not reached (NR) items for each PASEC and Rosetta Stone item block and passage. The NR rates are small enough and consistent enough across countries and item blocks/passages to not be of any concern. OM rates are higher for Rosetta Stone item blocks across all countries, with the exception of literacy passages in Senegal. Senegal has the lowest OM rates across Rosetta Stone item blocks/passages, while Burundi has the lowest OM rates across PASEC item blocks. Guinea has the highest OM rates across all item blocks/passages.

Exhibit 5.5: Average Percentage of Omitted and Not Reached Items by Item Block/Passage and Country for Reading/Literacy

	Burundi		Guinea		Senegal	
Item Block/Passage	**Average Percent Omitted**	**Average Percent Not Reached**	**Average Percent Omitted**	**Average Percent Not Reached**	**Average Percent Omitted**	**Average Percent Not Reached**
Rosetta Stone (PIRLS) Literacy						
L01	9.1	0.3	10.9	1.5	3.7	0.2
L02	15.2	2.1	14.8	1.4	4.0	0.3
L03	12.2	1.0	15.4	1.0	4.2	0.6
L04	20.4	1.8	17.7	2.1	4.9	0.3
Average	**14.2**	**1.3**	**14.7**	**1.5**	**4.2**	**0.4**
PASEC Reading						
RA	1.6	0.7	8.8	1.2	3.8	0.3
RB	1.7	0.6	7.8	1.2	3.8	0.1
RC	2.2	0.4	7.2	0.6	2.6	0.2
RD	2.4	1.0	8.3	1.8	2.8	0.3
Average	**2.0**	**0.7**	**8.0**	**1.2**	**3.3**	**0.2**

Exhibit 5.6: Average Percentage of Omitted and Not Reached Items by Item Block and Country for Mathematics/Numeracy

	Burundi		Guinea		Senegal	
Item Block	Average Percent Omitted	Average Percent Not Reached	Average Percent Omitted	Average Percent Not Reached	Average Percent Omitted	Average Percent Not Reached
Rosetta Stone (TIMSS) Numeracy						
N01	12.5	0.1	17.9	3.0	5.6	0.2
N02	12.0	2.2	16.0	1.9	6.4	0.3
N03	12.3	0.4	13.6	3.2	5.3	1.2
N04	8.7	0.3	9.8	1.9	3.3	0.1
N05	4.5	0.2	7.4	1.1	2.3	0.1
N06	17.2	1.4	23.3	0.6	8.6	0.1
N07	7.3	0.6	8.9	1.3	3.2	0.1
N08	4.9	0.2	9.2	0.4	2.9	1.0
Average	**9.9**	**0.7**	**13.3**	**1.7**	**4.7**	**0.4**
PASEC Mathematics						
MA	1.2	0.1	6.2	1.5	2.5	0.1
MB	0.7	0.1	4.4	0.9	1.5	0.0
MC	0.5	0.0	4.5	0.8	1.6	0.1
MD	1.1	0.0	5.2	0.9	1.8	0.3
Average	**0.9**	**0.1**	**5.1**	**1.0**	**1.9**	**0.1**

6. IRT Models

Section 6 describes item response theory (IRT) models and the estimation of item parameters and student proficiencies, in general. This is followed by section 7 which describes the application of IRT scaling in Rosetta Stone specifically and the PASEC item re-calibration.

6.1 IRT Scaling in Large-Scale Assessments

Given the complexities of the data collection and the need to describe student achievement on a scale that represents the entirety of the assessment frameworks, large-scale assessments such as TIMSS, PIRLS, or Rosetta Stone rely on IRT scaling to provide accurate measures of student proficiency distributions. Item Response Theory (IRT; Lord & Novick, 1968) has become one of the most important tools of educational measurement as it provides a flexible framework for estimating proficiency scores from students' responses to test items. IRT is particularly well suited to handle data collection designs in which not all students are tested with all items. The assumptions made for enabling IRT methods to handle these types of designs, commonly known as balanced incomplete block designs (e.g., von Davier, Sinharay, Oranje & Beaton, 2006; von Davier & Sinharay, 2013), can be described and tested formally (e.g., Fischer, 1981; Zermelo, 1929).

In terms of the mathematical notation used in this report, the item response variables on an assessment are denoted by for items $i = 1, \ldots, I$. The set of responses to these items is $\boldsymbol{x_v} = (x_{v1}, \ldots, x_{vi})$ for student v. For simplicity, we assume $x_{vi} = 1$ denotes a correct response and $x_{vi} = 0$ denotes an incorrect response. The expected achievement is assumed to be a function of an underlying latent proficiency variable, often in IRT denoted by $\boldsymbol{\theta_v}$, a real-valued variable. Then, we can write:

$$P\left(\boldsymbol{x_v} \middle| \boldsymbol{\theta_v}\right) = \prod_{i=1}^{I} P\left(x_{vi} \middle| \theta_v ; \boldsymbol{\zeta_i}\right) \tag{6.1}$$

where $P(x_{vi} | \theta_v ; \boldsymbol{\zeta_i})$ represents the probability of an either correct or incorrect response of a respondent with ability $\boldsymbol{\theta_v}$ and an item with a certain characteristic $\boldsymbol{\zeta_i}$. In IRT, these item-specific effects are referred to as item parameters. Equation (6.1) is a statistical model describing the probability of a set of the observed response given ability $\boldsymbol{\theta_v}$. This collective probability is the product of the individual item probabilities.

Many IRT models used in educational measurement can be understood as relatively straightforward generalizations of the approach shown in equation (6.1). While PASEC uses the Rasch model, TIMSS and PIRLS use the 3PL model for multiple-choice items, the 2PL model for constructed-response items worth 1 score point, and the GPCM for constructed-response items worth more than 1 score point. The following section describes these models in more detail.

6.2 IRT Models for Dichotomous Items: Rasch Model, 2PL Model and 3PL Model

The Rasch model and the two- and three-parameter logistic (2PL and 3PL) models are suitable for items with only two response categories (i.e., dichotomously scored items). The 2PL model (Birnbaum, 1968, in Lord & Novick, 1968) is a generalization of the Rasch model (Rasch, 1960), which assumes that the probability of a correct response to item i depends only on the difference between the ability level $\boldsymbol{\theta_v}$ of respondent v and the difficulty of the item b_i. But in addition, the 2PL allows that for every item, the association between this difference and the response probability can depend on an additional item discrimination (or slope) parameter a_i, characterizing its sensitivity to proficiency. The 3PL model (Birnbaum, 1968, in Lord & Novick, 1968) generalizes the 2PL model by additionally assuming a pseudo guessing parameter c_i. Under the 3PL model the response probability to an item is given as a function of the person parameter and the three item parameters; and it can be written as follows:

$$P\left(x = 1 \middle| \boldsymbol{\theta_v} ; \boldsymbol{\zeta_i}\right) = c_i + (1 - c_i) \; \frac{\exp\left(a_i\left(\theta_v - b_i\right)\right)}{1 + \exp\left(a_i\left(\theta_v - b_i\right)\right)} \tag{6.2}$$

The 3PL is a popular choice for binary scored multiple-choice items. If c_i is set to 0.0, equation (6.2) yields the 2PL model for 1-point constructed response items.

6.3 IRT Model for Polytomous Items: GPCM

A model frequently used for binary and polytomous ordinal items (items worth up to 2 points in TIMSS and items worth up to 3 points in PIRLS) is the generalized partial credit model (GPCM; Muraki, 1992), given by:

$$P_i(x|\boldsymbol{\theta_v}) = \frac{exp(a_i(x\theta_v - b_{ix}))}{1 + \Sigma_{z=1}^{m_i} exp(a_i(z\theta_v - b_{iz}))} \tag{6.3}$$

assuming a response variable with $m_i + 1$ ordered categories. Very often, the threshold parameters are split into a location and normalized step parameters, $b_{ix} = \delta_i - \tau_{ix}$, with $\Sigma_x \tau_{ix} = 0$.

The proficiency variable $\boldsymbol{\theta_v}$ is sometimes assumed to be normally distributed, that is, $\boldsymbol{\theta_v} \sim N(\mu,\sigma)$. In TIMSS, a normal distribution is used to obtain initial proficiency estimates, as the 3PL model requires constraints of this and other types for identification (Haberman, 2005; San Martín, González, & Tuerlinckx, 2015; von Davier, 2009). Subsequently, this normality constraint can be relaxed and other types of distributions utilized (Haberman, von Davier & Lee, 2008; von Davier & Sinharay, 2013; von Davier et al. 2006; von Davier & Yamamoto, 2004; Xu & von Davier, 2008).

The following sections address the central assumptions of IRT models such as unidimensionality, conditional independence and monotonicity of item-proficiency regressions.

6.4 Unidimensionality

Large-scale assessments measure students' achievement on several items they receive. Let I denote the number of items and let the response variables be denoted by $x = (x_1, \ldots, x_I)$. Unidimensionality means that a single quantity is sufficient to describe the probabilities of these responses to each of the items and that this quantity is the same regardless of the selection of items a student received from within an assessment domain. Denote P_{iv} and P_{jv} as the probability of person v scoring 1 on items i and j.

$$P_{iv} = P_i(X = 1|\boldsymbol{\theta_v}) \tag{6.4}$$

and

$$P_{jv} = P_j(X = 1|\boldsymbol{\theta_v}) \tag{6.5}$$

with the same real-valued $\boldsymbol{\theta_v}$ in each expression. Unidimensionality ensures that the same underlying proficiency is measured by all the test items in the domain. This of course holds only if the assessment development aims at producing a set of items that are indeed designed to assess the same assessment domain and that test developers diligently refer to the content specifications outlined in the assessment framework.

6.5 Conditional Independence

The assumption of population *independence* states that the probabilities of producing a correct response for a given level of proficiency are not dependent on the group to which a test taker belongs. In international large-scale assessments, this independence is important for inferences across countries, but also within countries for inferences across different student groups. Formally population independence holds if

$$P(X_i = x_i \mid \theta, g) = P(X_i = x_i \mid \theta) \tag{6.6}$$

for any contextual variable g. This also holds for groups defined by performance on x_j on items $j < i$ that precede the current item response x_i. The response to a preceding item can be considered a grouping variable as well, as it splits the sample into those that produced a correct response and those who did not, in the simplest case. Applying the assumption of population independence, this yields

$$P(x_i, x_j \mid \theta) = P(x_i \mid x_j, \theta)\, P(x_j \mid \theta) = P(x_i \mid \theta)\, P(x_j \mid \theta) \tag{6.7}$$

The assumption of local independence directly follows. It states that the joint probability of observing a series of responses, given a student's proficiency level θ, can be written as the product of the item level probabilities. For a set of responses, local independence takes the form

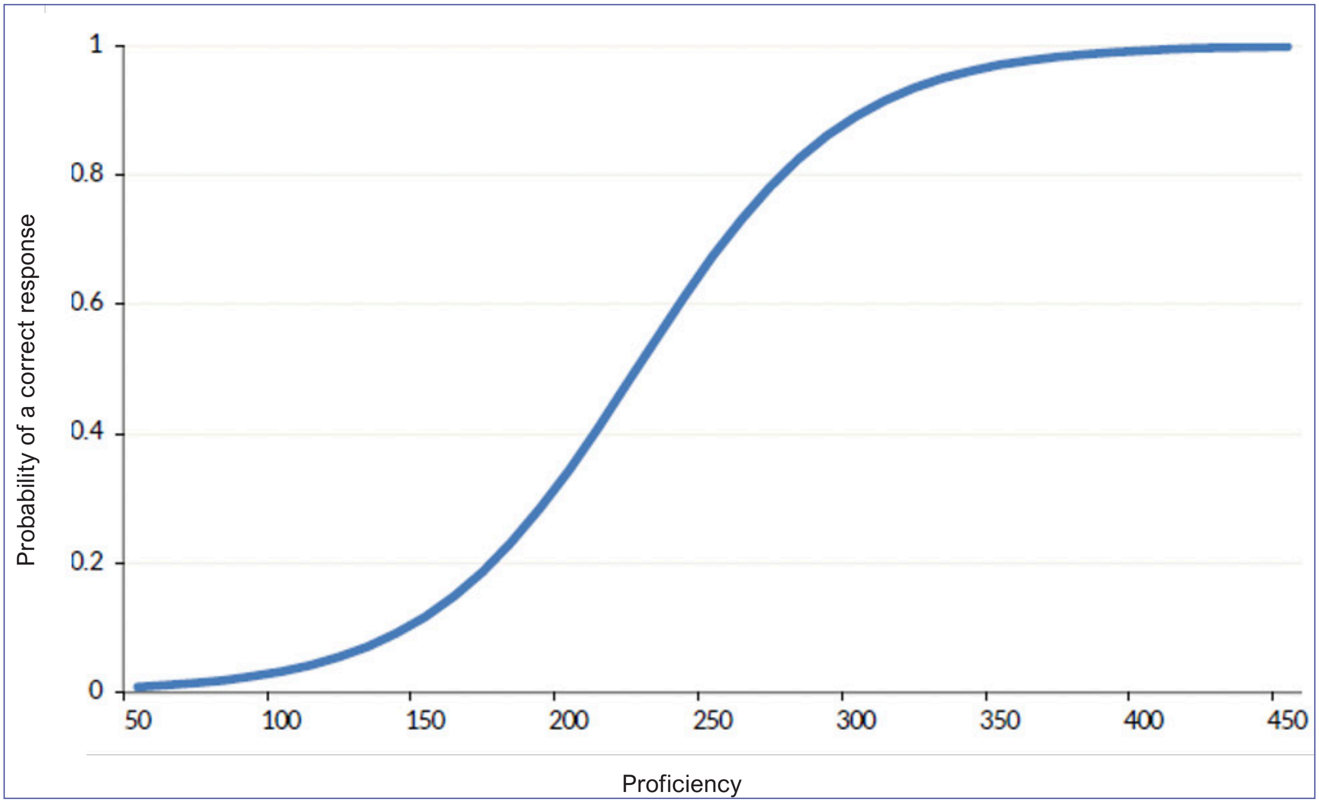
$$P(\mathrm{X} = x_1, \ldots, x_I \mid \theta) = \prod_{i=1}^{I} P_i\,(X = 1 \mid \theta)^{x_i}\,[1 - P_i\,(X = 1 \mid \theta)^{1-x_i} \tag{6.8}$$

According to the assumption of population invariance and local independence, if the model fits the data (and, for example, no learning occurs) and only one single proficiency is 'responsible' for the probability of giving correct responses, then no other variables (including language of the assessment, citizenship, gender, and other contextual variables) are helpful in predicting a respondent's answer to the next item. In this sense, the assumption of local independence and population invariance encapsulate the goal that there is only one variable that needs to be considered and that estimates of this variable will fully represent the available information about proficiency.

6.6 Monotonicity of Item-Proficiency Regressions

One important assumption of IRT models used for achievement data is the (strict) monotonicity of item functions. As seen in Exhibit 11.1, the Rasch model (but also the 2PL and 3PL IRT models) assumes that the probability of a correct response increases with increasing proficiency.

Exhibit 6.1: Example Item Characteristic Curve



This is represented in the following inequality

$$P(X_i = 1 \mid \boldsymbol{\theta_v}) < P(X_i = 1 \mid \boldsymbol{\theta_w}) \longleftrightarrow \boldsymbol{\theta_v} < \boldsymbol{\theta_w} \tag{6.9}$$

for all items i. This assumption ensures that the proficiency 'orders' the success on the items the students receive and implies that students with a higher level of proficiency will also have a higher probability of success on each of the items in the achievement domain. By implication, there is also a strict monotonic relationship between the expected achievement scores and proficiency θ:

$$E(S \mid \boldsymbol{\theta_v}) = \sum_{i=1}^{I} P(X_i = 1 \mid \boldsymbol{\theta_v}) < E(S \mid \boldsymbol{\theta_w}) = \sum_{i=1}^{I} P(X_i = 1 \mid \boldsymbol{\theta_w}) \longleftrightarrow \boldsymbol{\theta_v} < \boldsymbol{\theta_w} \tag{6.10}$$

The equation above shows that a person with a greater skill level $\boldsymbol{\theta_w}$ compared to a lesser skill level $\boldsymbol{\theta_v}$ will in terms of expected score $E(S|\boldsymbol{\theta_w})$ obtain a larger number of correct responses. This monotonicity ensures that the items and test-takers are ordered as one would expect, namely that higher levels of proficiency are associated with higher expected achievement — a larger expected number of observed correct responses — for any given item or item block measuring the same domain in an assessment booklet.

6.7 Multidimensional IRT Models

In multidimensional IRT (MIRT) models, the model can be specified for multiple scales. It is assumed that the IRT holds, with the qualifying condition that it holds with one or more ability parameters for each of a set of distinguishable subsets (scales) of items (Reckase, 2009; von Davier, Rost, and Carstensen 2007). For the case of a multidimensional 2PL, for example, with between-item multidimensionality (each item loads on only one scale), the probability of response (X_{iv}=1) to item i in scale k by respondent v can be defined as:

$$P(x_{iv} = 1 \mid \boldsymbol{\theta_v}, \beta_i, \alpha_i) = \frac{\exp\left[\sum_{k=1}^{K} \alpha_{ik}(x_{iv}\theta_{vk} - \beta_i)\right]}{1 + \exp\left[\sum_{k=1}^{K} \alpha_{ik}(x_{iv}\theta_{vk} - \beta_i)\right]}, \tag{6.11}$$

where $\boldsymbol{\theta_v}$ is a vector of latent variables and α_i is a vector of the item loadings for item i on scale k with the restriction that each item loads on only one scale. Unidimensional IRT models used in our analysis may be treated as special case of MIRT where $\boldsymbol{\theta_v} = \theta_v$ that is one latent dimension is assumed (K=1).

The following section will describe how the IRT models illustrated above were applied to the Rosetta Stone study data to estimate item parameters and to examine their cross-country and cross-assessment invariance.

7. IRT Model Application to PASEC and Rosetta Stone Data

This section describes the application of IRT scaling to Rosetta Stone linking items in particular as well as the PASEC item re-calibration performed by the TIMSS & PIRLS International Study Center. An overview of the specific model applications, and the examination of item-by-country interactions are followed by the results for Rosetta Stone linking and PASEC items.

7.1 Establishing Comparability through IRT Scaling

The comparability across assessments and countries for the Rosetta Stone linking items was evaluated by fixing the parameters to the published TIMSS and PIRLS item parameters for all three countries. More precisely, the item parameters used came from the TIMSS 2019 less difficult IRT calibration and the PIRLS Literacy 2016 IRT calibration (both assessments were linked to TIMSS and PIRLS) and were estimated based on the 2PL, 3PL and GPCM (Martin, von Davier & Mullis, 2020). The comparability of the PASEC items across countries was evaluated by estimating common item parameters across countries based on the Rasch model, in accordance with PASEC analysis procedures (PASEC, 2020). All IRT models were applied as multiple group models with countries as groups and estimated using the open-source package *mirt* (Chalmers, 2012) available in the R statistical programming language (R Core Team, 2013).

Separate *unidimensional multiple group IRT models* (with countries as groups) were estimated for each assessment domain resulting in four models:

- Model 1 (M1) was estimated for the 105 TIMSS numeracy items.
- Model 2 (M2) was estimated for the 62 PIRLS literacy items.
- Model 3 (M3) was estimated for the 84 PASEC mathematics items.
- Model 4 (M4) was estimated for the 98 PASEC reading items.

While M1 and M2 use the published TIMSS and PIRLS item parameters as fixed values, item parameters for M3 and M4 were estimated. In a first step, common item parameters were assumed across countries in each model. The fit of these common parameters was examined for all item-by-country combinations. That is, item-by-country interactions were examined as a possible result of differential item functioning (DIF). To set the scale, a reference group constraint was used when all item parameters were estimated in the model (M3 and M4) while no reference group constraint was used if item parameters were fixed in the model (M1 and M2).

Item-level model-fit analyses are a critical part of the scaling analyses described above. Different types of DIF statistics can be used to evaluate the extent to which the IRT model applied to a group fits the response data collected from that group. In the context of the IRT models used in the Rosetta Stone study, item-level model fit was examined using a robust approach to identifying misfit (von Davier & Bezirhan, 2021) based on the root mean squared deviation (RMSD).

The *RMSD* quantifies the extent to which the model-based item characteristic curve (ICC; computed using equations 6.2 or 6.3) and the empirical ICC can differ with regard to both the item difficulty parameters and item slope parameters. The ICC characterizes the relationship between a person and item parameters. The RMSD is defined as:

$$RMSD = \sqrt{\int (P_o(\theta) - P_e(\theta))^2 f(\theta)\, d\theta} \tag{7.1}$$

where $P_o(\theta)$ and $P_e(\theta)$ are the observed and expected probability of a correct response given proficiency θ; and $f(\theta)$ is the country-specific density (Khorramdel, Shin, & von Davier, 2019; von Davier, 2005). The observed probability correct is based on the pseudo counts from the EM algorithm that is used to estimate the model (Bock & Aitkin, 1981), while the expected probability correct is based on the estimated item function.

The *median absolute deviation (MAD)* is a robust measure of dispersion which can be used as a flagging rule to detect misfitting items. MAD classifies an observation as an outlier if the difference to the median of the absolute distances of all other observations exceeds a certain boundary. MAD is calculated as:

$$MAD = b\, M_i(|x_i - M_j(x_j)|) \tag{7.2}$$

where, x_j is the n original observations and M_i is the median of the series (Leys et al., 2013). b is the reciprocal of 0.75 quantiles of the underlying distribution. Under the assumption of normality of the data $b = 1/\,Q(0.75) = 1.4826$. A threshold (k) should be defined to identify the misfitting observations. Then we can write the decision criterion as:

$$\frac{x_i - M}{MAD} > |\pm k|. \tag{7.3}$$

In the Rosetta Stone scaling, the MAD outlier detection approach was applied to the RMSD values for all country-by-item combinations to identify misfitting items. Any value obtained in (7.3) exceeding a threshold of 1.96 was flagged as an outlier of the RMSD distribution (i.e., as misfitting item).

Item misfit relative to the TIMSS and PIRLS item parameters in M1 and M2 indicates that item characteristics (such as item difficulty and discrimination) differ across the data collections. In such cases, new common item parameters were estimated across countries and the item fit was evaluated again. Item misfit to new common item parameters in M1, M2, M3, and M4 indicates that item characteristics differ across PASEC countries. In such cases, items were excluded from the scaling.

After PASEC and Rosetta Stone items were scaled with separate unidimensional IRT models, *multidimensional IRT models* were used to examine how similar or different the measured constructs of the different assessments are. More precisely, the latent correlations from the multidimensional models were used to investigate the relationship between the PASEC mathematics and TIMSS numeracy scales and between the PASEC reading and PIRLS literacy scales. Hence, the following 2-dimensional IRT models were estimated:

- Model 5 (M5) was estimated with the PASEC mathematics items assigned to one factor/scale and TIMSS items assigned to a second factor/scale.
- Model 6 (M6) was estimated with the PASEC reading items assigned to one factor/scale and PIRLS items assigned to a second factor/scale.

The item parameters in M5 were fixed to the item parameter values obtained from M1 and M3, while the item parameters in M6 were fixed to the item parameter values in M2 and M4.

To establish a meaningful concordance between the PASEC scales and the TIMSS or PIRLS scales, these need to measure highly similar constructs, which is evaluated by means of the magnitude of the latent correlations estimated in models M5 and M6.

7.2 Results for Unidimensional IRT Models

The unidimensional IRT models showed high levels of comparability across countries and across assessments for the Rosetta Stone scales (M1, M2) and across countries for PASEC scales (M3, M4) providing a solid basis for establishing a concordance. The tables in Exhibits 7.1 and 7.2 show the percentages of common (fixed and new) and excluded item parameters for all item-by-country combinations in each of the unidimensional IRT models.

Results for M1 and M2 showed high levels of agreement of item functioning across countries and assessments. In M1 and M2, the TIMSS numeracy and PIRLS literacy item parameters showed a good fit for the majority of item-by-country pairs (86.3% and 81.2% respectively). For a very small subset of items, new common item parameters needed to be estimated (12.4% and 8.1% for numeracy and literacy respectively) which, therefore, do not serve as link items to the TIMSS and PIRLS scales but are still comparable across Rosetta Stone countries. In some cases of item-by-country pairs, items needed to be excluded from the analysis (1.3% and 10.7% for numeracy and literacy, respectively); items were either excluded for all or for single countries.

Results for M3 and M4 showed high levels of agreement of item functioning across countries as well. In the vast majority of item-by-country pairs for the PASEC mathematics and PASEC reading items, a good fit to the common item parameter estimates was achieved (92.5% and 91.5% respectively). In a very small number of cases of item-by-country pairs, items needed to be excluded from the analysis (7.5% and 8.5% for mathematics and reading respectively); again, items were either excluded for all or single countries.

Exhibit 7.1: Percentages of Item Parameter Estimates for Item-by-Country Combinations (Pairs) in Model 1 and Model 2

Item Parameters	TIMSS-Numeracy (Model 1)	PIRLS-Literacy (Model 2)
Fixed	86.3%	81.2%
New Common	12.4%	8.1%
Excluded	1.3%	10.7%

Exhibit 7.2: Percentages of Item Parameter Estimates for Item-by-Country Combinations (Pairs) in Model 3 and Model 4

Item Parameters	PASEC Math (Model 3)	PASEC Reading (Model 4)
Common	92.5%	91.5%
Excluded	7.5%	8.5%

A graphical overview of the proportions of fixed and common (invariant) item parameters and excluded items in each domain is given in the figures in Exhibits 7.3 to 7.6. In Exhibits 7.3. and 7.4, dark green indicates the fixed TIMSS and PIRLS item parameters (common item parameters across assessments), light green indicates new common item parameters (common across PASEC countries), and orange indicates excluded items for specific item-by-country pairs. In Exhibits 7.5 and 7.6, dark green indicates common item parameter estimates (common across PASEC countries) and orange indicates excluded items for specific item-by-country pairs. Note that item parameters were ordered for visualization purposes and that the grouping of colors in the figures does not indicate any specific pattern. No particular pattern could be observed for item-by-country interactions with regard to item type or content.

Exhibit 7.3: Distribution of Model 1 Items with Common Item Parameters versus Excluded Items

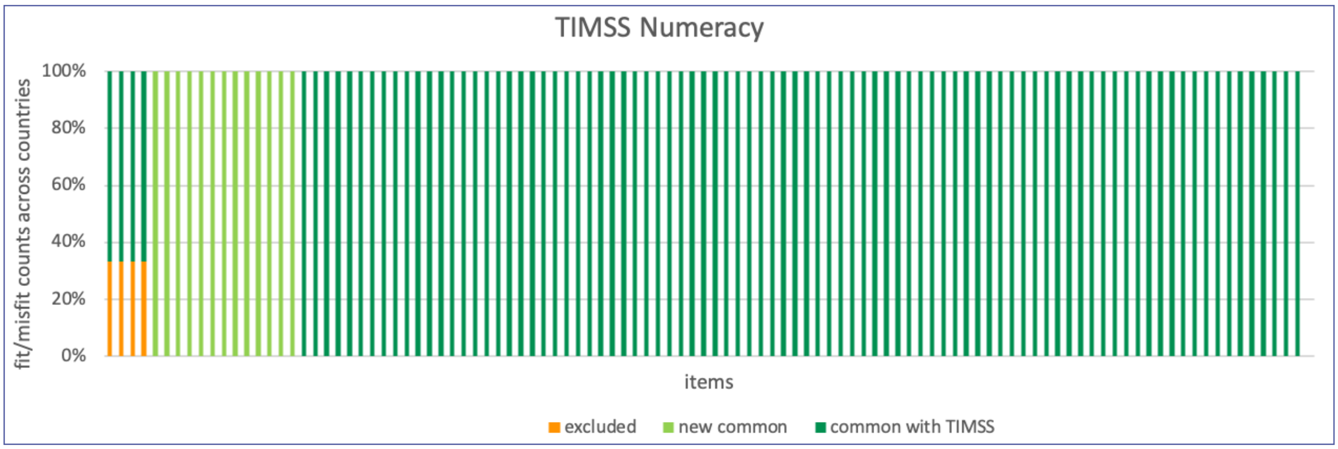


Exhibit 7.4: Distribution of Model 2 Items with Common Item Parameters versus Excluded Items

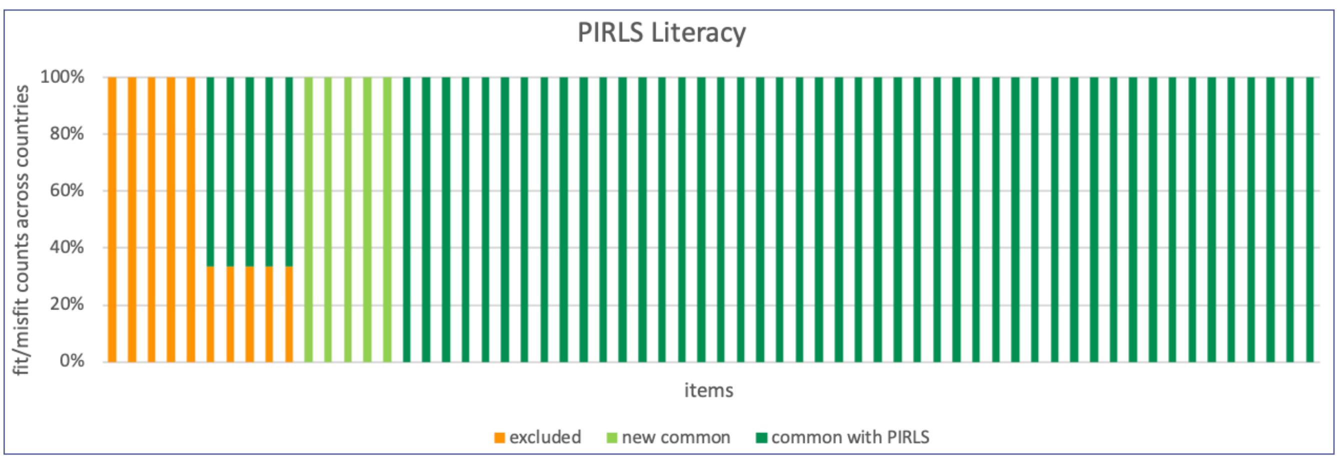


Exhibit 7.5: Distribution of Model 3 Items with Common Item Parameters versus Excluded Items

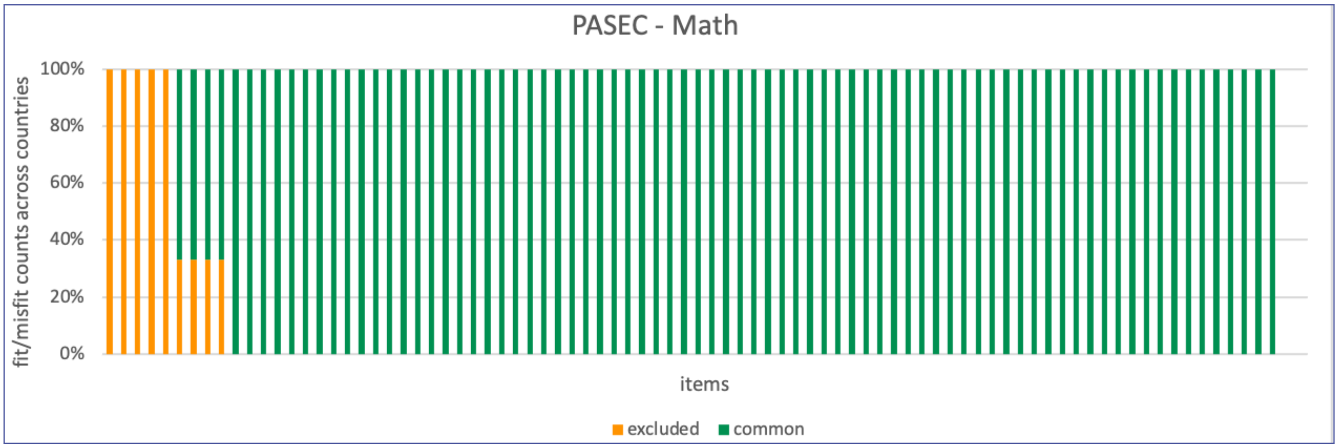
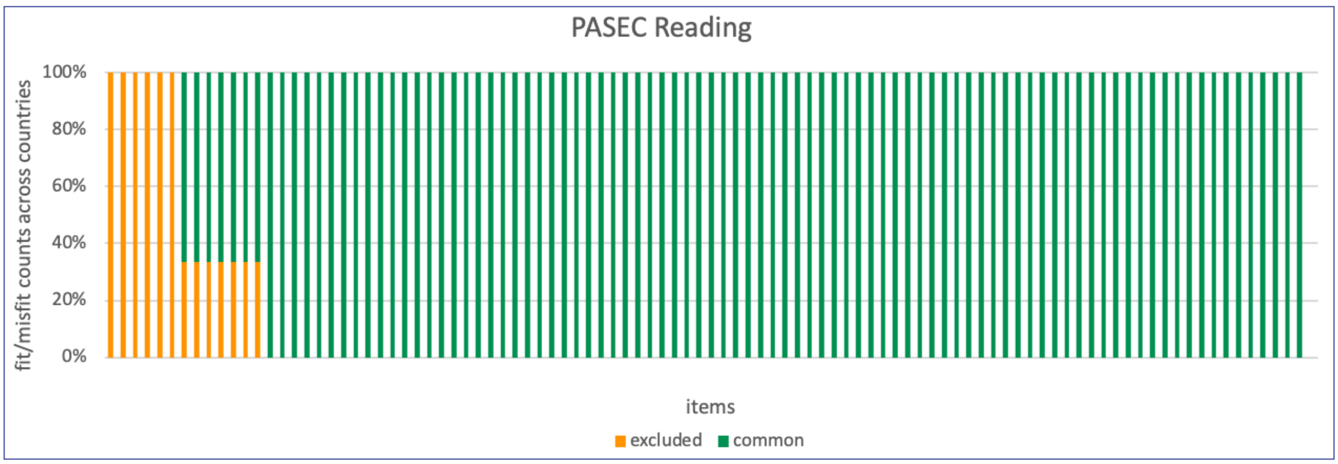


Exhibit 7.6: Distribution of Model 4 Items with Common Item Parameters versus Excluded Items



Given the small number of PASEC countries participating in Rosetta Stone and the smaller sample sizes in each country compared to customary TIMSS and PIRLS samples, the *uncertainty in the estimation* of common item parameters in M3 and M4 needed to be examined. That is, the effects of single countries on the item parameter estimation. This was done by conducting and comparing different rounds of item parameter estimation, separately for M3 and M4, using the *leave one "country" out (LOO) method*. More precisely, M3 and M4 were estimated by leaving one country out at a time of the estimation in each iteration. To obtain the final estimates from the calibrations, item parameters were pooled, and variability was estimated. Exhibits 7.7 and 7.8 illustrate the Rasch model-based item difficulties for M3 and M4 for the different estimation rounds: the colored lines indicate the estimates for each round with one country left out at a time while the black dots indicate the mean difficulties across all estimation rounds and the related mean standard errors as indicated by the intervals. Note that items in both figures were ordered by difficulty for visualization purposes.

Overall, mathematics items (M3) estimates show larger variability compared to the reading items (M4). For LOO with M3, it was also observed that when Burundi was left out of the estimation, difficulty estimates were consistently larger for mathematics items compared to when either Senegal or Guinea was left out. For LOO with M4, no clear pattern was observed for the item parameter estimation. Overall, the effect of a single country on item parameter estimates was minimal, especially for PASEC reading items.

Exhibit 7.7: Rasch Model-Based Item Difficulties for LOO Estimation Rounds – Model 3 (PASEC Mathematics)

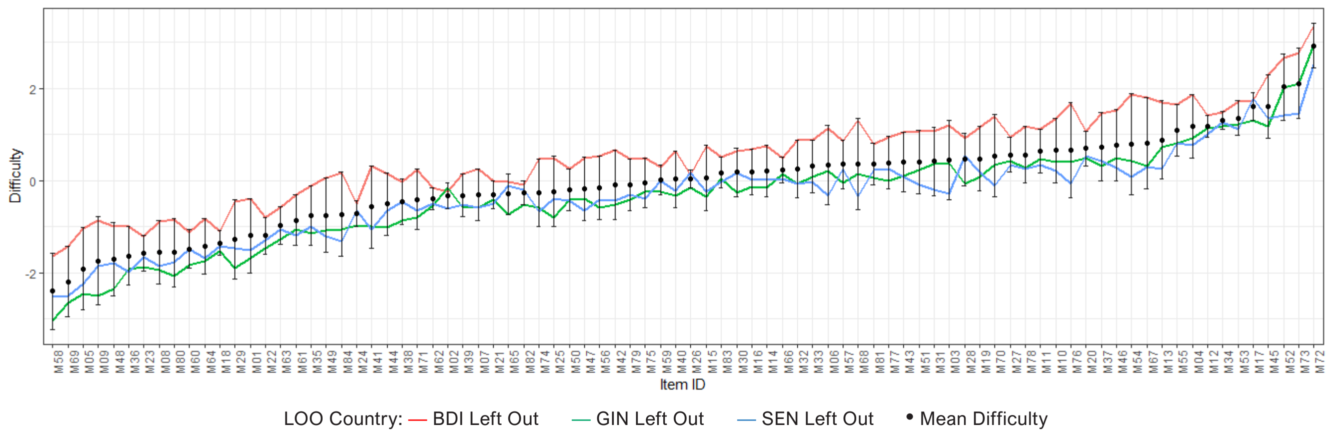
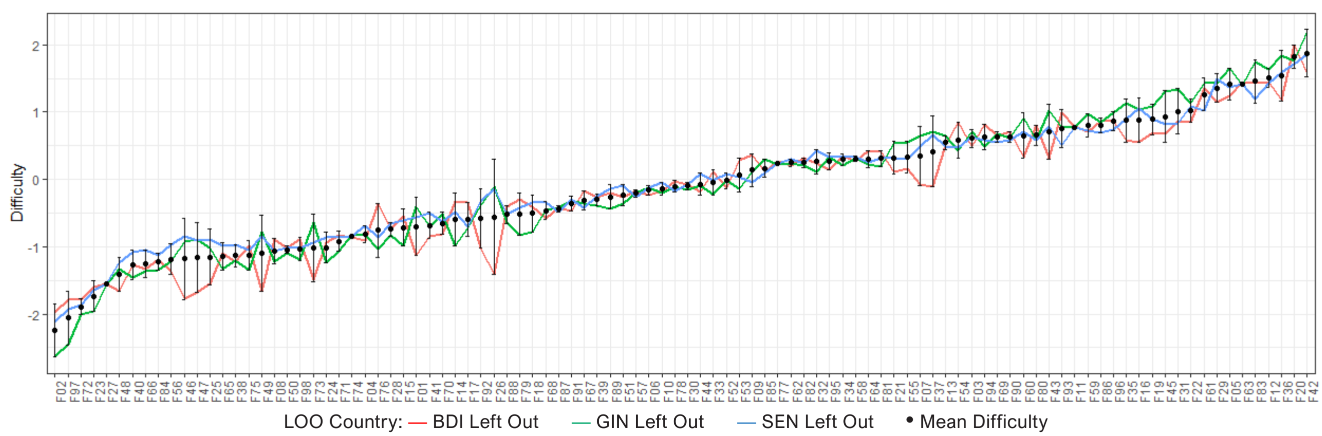


Exhibit 7.8: Rasch Model-Based Item Difficulties for LOO Estimation Rounds – Model 4 (PASEC Reading)



7.3 Results for Multidimensional IRT Models

The 2-dimensional IRT models (M5 and M6) provided information about the relation and similarity of the different constructs. The latent correlations between dimensions in both 2-dimensional IRT models showed to be substantial but not perfect ranging from .81 to .86 across countries in M5, and from .78 to .83 across countries in M6, see exhibit 7.9. This indicates that the corresponding Rosetta Stone and PASEC scales measure constructs that are not the same but similar enough to enable a meaningful concordance for the projection of score distributions.

Exhibit 7.9: Latent Correlations between PASEC and Rosetta Stone Scales

Country	PASEC Mathematics with TIMSS (M5)	PASEC Reading with PIRLS (M6)
Burundi	.81	.78
Guinea	.81	.79
Senegal	.86	.83

8. Population Models

Section 8 describes the general principles followed for the population modeling and the imputation of plausible values (PVs).

8.1 Integrating Achievement Data and Context Information

Rosetta Stone uses a latent regression or population model to estimate distributions of proficiencies. The population model is based on the likelihood function of an IRT model, as introduced in section 6 of this report, and a linear, latent regression of the proficiency on contextual data collected in background or context questionnaires (von Davier et al., 2006; von Davier et al., 2009). This approach can be viewed as an imputation model for the unobserved proficiency distribution that aims at obtaining unbiased group-level proficiency distributions by utilizing information about the extent to which background or context variables are related to the proficiency variable. Population models use a large number of context variables in the latent regression to avoid the omission of any useful information (von Davier et al., 2006; von Davier et al., 2009; von Davier & Sinharay, 2013).

To reduce the number of context variables and avoid overparameterization, a principal component analysis (PCA) is used to eliminate collinearity by identifying a smaller number of orthogonal predictors that account for most of the variation in the background variables.

To facilitate the estimation procedure, the data from the context questionnaires are combined with the responses obtained from the achievement items. The complete observed data for a person n can be expressed as $d_n = (x_{n1}, ..., x_{nI}, g_n, z_{n1}, ..., z_{nB})$, where $z_{n1}, ..., z_{nB}$ represent the context information, $x_{n1}, ...,$

x_{nI} represent the answers to the achievement items, and g_n represents the country or population the respondent was sampled from.

The estimation of student-level posterior proficiency distributions with IRT models utilizes an estimate of the proficiency distributions in the population of interest. A population model that incorporates contextual data utilizes this information by specifying a second-level model that predicts the distribution of proficiency as a function of contextual variables. The conditional expectation in this model is given by

$$\mu_n = \sum_{b=1}^{B} \beta_{g(n)b} z_{nb} + \beta_{g(n)0} \tag{8.1}$$

This expectation uses the available information on how context variables relate to the proficiency. The distribution of proficiency is assumed to be normally distributed around this conditional expectation, namely $\boldsymbol{\theta_n} \sim N(\mu_n, \sigma)$.

Together with the likelihood of the responses expressed by the IRT model, this provides a model for the posterior distribution of proficiency given the context data $z_{n1}, \ldots, z_{nB}$ and the responses to the items. In other words, the model implements the assumption that the posterior distribution of proficiency depends on the context data as well as on the observed item responses. Therefore, if background variables are selected so that correlations with proficiency are likely, one obtains a distribution around the expected value given the conditional expectation in (8.1) that is noticeably more accurate than a country-level distribution of proficiency.

8.2 Group-Level Proficiency Distributions and Plausible Values

The goal of population modeling is to produce posterior distributions of proficiencies from which plausible values (PVs) can be drawn. Integrating the IRT models described in section 7 of this report with the regression model introduced at the beginning of this section, we can estimate the probability of the responses, conditional on context information, as

$$P_g(x_n \mid z_n) = \int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni} \mid \theta)\, \Phi\left(\theta ; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right) d\theta \tag{8.2}$$

This equation provides the basis for the imputation of proficiency estimates that are commonly known as PVs (Mislevy, 1991). To allow a more compact notation, we use

$$P_{ig}(x_{ni} \mid \theta) = P_{ig}(X = 1 \mid \theta)^{x_{ni}} \left[1 - P_{ig}(X = 1 \mid \theta)\right]^{1-x_{ni}} \tag{8.3}$$

The model given in 8.2 enables inferences about the posterior distribution of the proficiency θ, given both the TIMSS assessment items $x_1, \ldots, x_I$ and the context information $z_1, \ldots, z_B$. The posterior distribution of the proficiency given the observed data can be written as

$$P_g(\theta | x_v, z_n) = \frac{\prod_{i=1}^{I} P_{ig}(x_{ni} | \theta)\, \Phi(\theta; \Sigma_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma)}{\int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni} | \theta)\, \Phi(\theta; \Sigma_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma)\, d\theta} \quad (8.4)$$

An estimate of where a respondent n is most likely located on the proficiency dimension can be obtained by

$$E_g(\theta | x_n, z_n) = \int_\theta \theta P_g(\theta | x_n, z_n) = d\theta \quad (8.5)$$

The posterior variance, which provides a measure of uncertainty around this expectation, is calculated as follows:

$$V_g(\theta | x_n, z_n) = E_g(\theta^2 | x_n, z_n) - [E_g(\theta | x_n, z_n)]^2 \quad (8.6)$$

Using these two estimates (the posterior mean and variance) to define the posterior proficiency distribution, it is possible to draw a set of PVs from this distribution for each student. PVs are the basis for all reporting of proficiency data in large-scale assessments such as TIMSS, PIRLS or PASEC, allowing reliable group-level comparisons.

Note that the correlations between context variables and proficiency are estimated separately in each country so that there is no bias or inaccurate attribution that could affect the results. Although the expected value of the country-level proficiency is unchanged whether context information is used or not, the advantage of including context information plays out when making group-level comparisons. It can be shown analytically and by simulation (von Davier et al., 2009) that including context information in a population model greatly reduces bias in group-level comparisons using this information, and using country-specific population models with context variables ensures there is no bias in country-level average proficiency data.

In summary, the PVs used in TIMSS, PIRLS, PASEC, and other large-scale assessments are random draws from a conditional normal distribution

$$\tilde{\theta}_{ng} \sim N\left(E_g(\theta | x_n, z_n \sqrt{V_g(\theta | x_n, z_n)}\right) \quad (8.7)$$

that depends on response data x_n as well as context information z_n estimated using a group-specific model for each country g. That means two respondents with the same item responses, but different context information will receive a different predicted distribution of their corresponding latent trait. Although this may seem potentially unfair to individual test takers – and would not be adequate to assign test scores to individual students – it is important to remember that large-scale assessments are population surveys, not individual assessments, and that it is necessary to include context information in order to achieve unbiased comparisons of population distributions (e.g., Little & Rubin, 1987; Mislevy, 1991; Mislevy et al., 1992; Mislevy & Sheehan, 1987; von Davier et al., 2009). Consequently, PVs are not and should never be used or treated as individual test scores.

9. Population Model Application to PASEC and Rosetta Stone Data

Section 9 describes the application of population models in Rosetta Stone specifically as well as the replication of PASEC PVs for validation purposes.

9.1 Applied Population Models

The population model, as described above, is a multivariate model that incorporates the available student context variables from the PASEC student questionnaire, as well as the Rosetta Stone linking item parameters and the PASEC item parameters from the IRT scaling, respectively.

For Rosetta Stone, two 2-dimensional models were used:

- Population Model 1: was estimated for TIMSS numeracy and PIRLS literacy
- Population Model 2: was estimated for PASEC math and PASEC reading

Population Model 1 follows the practice established by TIMSS and PIRLS of using principal components analysis for reducing collinearity and dimensionality of predictors while retaining 90% of their common variance. It was calculated separately for each of the three countries that participated in the Rosetta Stone study. Latent regression parameters were estimated while the item parameters obtained from the IRT scaling (described in section 7) were assumed to be fixed and known.

In addition to the principal components, students' gender (dummy coded) and an indicator of the classroom in the school to which a student belongs (criterion scaled) were included as primary conditioning variables. Exhibits 9.1 provide details on the counts of variables used in the latent regression used for proficiency estimation of the Rosetta Stone linking data.

Exhibit 9.1: Counts of Conditioning Variables used for the Rosetta Stone Linking Data

Country	Number of Primary Conditioning Variables	Number of Principal Components Available	Number of Principal Components Retained	Percentage of Variance Explained
Burundi	2	233	90	90
Guinea	2	234	89	90
Senegal	2	234	103	90

The same analysis steps and conditioning variables were used for Population Model 2. Note that this model was only estimated for evaluation purposes and is, therefore, not described in detail here. The PASEC PVs that were provided by the PASEC team were used for constructing the concordance tables after the validity could be confirmed based on the results of Population Model 2.

9.2 Generating Plausible Values and PASEC Score Validation

Educational Testing Service's DGROUP program (Rogers et al., 2006) was used to estimate the latent regression models and generate PVs. A useful feature of DGROUP is its ability to estimate multi-dimensional latent regression models using the responses to all items across the proficiency scales and the correlations among the scales to improve the reliability of estimates (e.g., von Davier, Sinharay, Oranje & Beaton, 2006).

Following the procedures in TIMSS and PIRLS (Foy, Fishbein, von Davier, & Yin, 2020; Foy & Yin, 2017), five PVs were drawn from the conditional distribution for each domain and each student. A predictive distribution of PVs was produced for the TIMSS numeracy and the PIRLS literacy domains (Population Model 1) as well as for the PASEC mathematics and reading domains (Population Model 2).

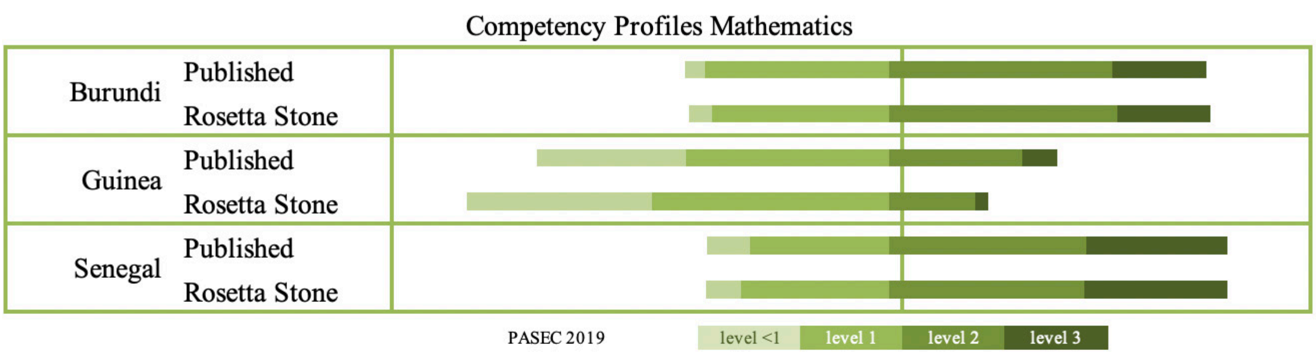
The PASEC PVs received from the PASEC team were evaluated in two steps. First, the distributions for PASEC scales based on these PVs were compared to the PASEC 2019 published results. Exhibits 9.2 and 9.3 show that both sets of results are very similar indicating that the Rosetta Stone student sample is comparable to the PASEC 2019 student sample. The standard errors (SEs) for the 2019 published results generally are smaller than SEs for the Rosetta Stone sample because they were estimated based on larger national samples.

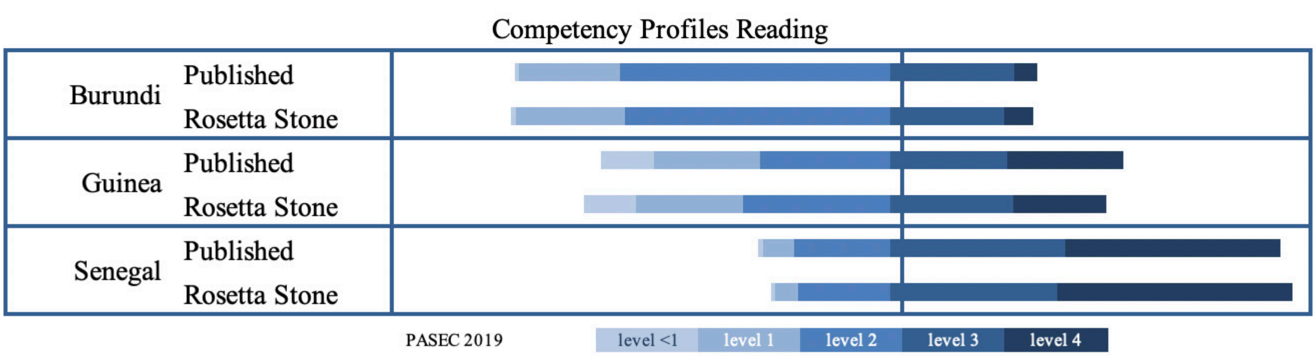
Second, they were compared to the re-estimated PVs from Population Model 2. Very high correlations between both sets of PVs could be observed (ranging from .95 to .97, and from .96 to .98 for mathematics and reading respectively) indicating very good agreement of analytic processes.

Exhibit 9.2: Comparison of Published (2019) and Rosetta Stone Study Results for PASEC Scales

Statistics		Burundi		Guinea		Senegal	
		Published	Rosetta Stone	Published	Rosetta Stone	Published	Rosetta Stone
Mathematics	Mean	546 (3.2)	546 (4.3)	482 (4.7)	459 (6.4)	558 (4.7)	557 (7.7)
	Std. Dev.	71 (2.2)	73 (2.3)	85 (3.4)	73 (3.9)	91 (2.8)	85 (4.4)
	Level 0	3.8 (0.6)	4.3 (0.8)	28.7 (2.1)	35.5 (3.4)	8.2 (1.2)	6.6 (1.2)
	Level 1	35.3 (1.6)	34.0 (2.1)	38.9 (1.7)	45.5 (2.1)	26.7 (1.4)	28.4 (2.9)
	Level 2	42.9 (1.4)	43.8 (1.8)	25.6 (1.7)	16.6 (2.8)	37.8 (1.5)	37.6 (2.7)
	Level 3	18.0 (1.5)	17.9 (1.9)	6.8 (1.2)	2.4 (0.9)	27.2 (2.0)	27.4 (3.8)
Reading	Mean	490 (2.7)	488 (3.3)	503 (6.0)	495 (7.3)	576 (4.9)	586 (9.5)
	Std. Dev.	58 (2.4)	63 (2.5)	115 (4.5)	105 (4.1)	90 (3.4)	91 (4.1)
	Level 0	0.6 (0.2)	0.9 (0.3)	10.0 (1.4)	9.9 (1.6)	1.0 (0.4)	0.9 (0.4)
	Level 1	19.4 (1.1)	20.9 (2.0)	20.4 (1.5)	20.4 (1.9)	6.0 (0.9)	4.3 (0.8)
	Level 2	51.8 (1.3)	50.8 (1.9)	24.9 (1.4)	28.2 (2.2)	18.3 (1.3)	17.6 (2.4)
	Level 3	23.7 (1.2)	21.8 (2.1)	22.5 (1.5)	23.6 (1.8)	33.6 (1.7)	32.0 (3.1)
	Level 4	4.5 (1.0)	5.5 (0.8)	22.2 (1.7)	17.9 (2.4)	41.1 (2.2)	45.2 (4.9)

Exhibit 9.3: Graphical Comparison of Published (2019) and Rosetta Stone Study Results for PASEC Scales





9.3 Transforming the Plausible Values to TIMSS and PIRLS Scales

The numeric scales of the PVs that were drawn using the model parameters of each population model were set by means of the IRT scaling and had to be transformed to the TIMSS and PIRLS reporting metric. This was accomplished through a set of linear transformations given by:

$$\boldsymbol{PV}^{*} = \boldsymbol{A_{ik}} + \boldsymbol{B_{ik}} \times \boldsymbol{PV_{ik}} \tag{9.1}$$

Where $\boldsymbol{PV_{ik}}$ is the plausible value i of scale k (mathematics or reading) prior to transformation; $\boldsymbol{PV^{*}_{ik}}$ is the plausible value i of scale k after transformation; and $\boldsymbol{A_{ik}}$ and $\boldsymbol{B_{ik}}$ are the linear transformation constants.

For the Rosetta Stone linking data, the linear transformation constants for numeracy and literacy were obtained from TIMSS 2019 less difficult (Foy et al., 2020) and PIRLS Literacy 2016 (Foy and Yin, 2016). There are five sets of transformation constants for each scale or subject, one for each plausible value (Exhibit 9.4).

Exhibit 9.4: Transformation Constants for Rosetta Stone (TIMSS and PIRLS) Linking Data

Plausible Value (PV)	TIMSS		PIRLS	
	A	B	A	B
PV1	404.448	113.863	516.968	96.598
PV2	404.156	113.749	516.163	97.544
PV3	405.574	112.539	515.765	97.534
PV4	404.177	114.003	515.905	97.571
PV5	403.994	114.170	516.014	97.267

The following two sections describe how posterior means and PVs produced for Rosetta Stone data were used to establish concordance tables for PASEC mathematics and TIMSS numeracy as well as for PASEC reading and PIRLS literacy.

10. Establishing an Enhanced Concordance between Scales

Scale concordance refers to establishing a relationship between scores on different assessments or tests that measure similar (but not identical) constructs. It aims to provide a projection onto a target scale score from a source scale score. In Rosetta Stone, a range of TIMSS and PIRLS scores is predicted or projected from PASEC mathematics and reading scores respectively. That is, PASEC mathematics and PASEC reading represent the source test θ and TIMSS and PIRLS represent the target test ϑ. This prediction can be displayed as a concordance table and provide useful information to stakeholders, researchers, or institutions who need to compare test scores.

A technically sound concordance allows students and professionals to compare scores from similar assessments to inform decisions. However, concorded scores are not true predictions of how students would perform on the other test as they do not provide a direct link between tests. While predictions or equating of scores includes uncertainty due to measurement error, concordance-based projections include an additional source of uncertainty, the error due to projecting from one construct to another. In addition, concordance tables are dependent on the characteristics of the sample and include uncertainty due to sampling. Hence, the uncertainty of the prediction has to be taken into consideration when using and interpreting concordance tables.

The method used for establishing scale concordance in the Rosetta Stone study directly takes the uncertainty of the proficiency estimates on source and target test forms into account and thus appropriately controls for potential construct differences between the tests. More specifically, the proposed method is based on predictive mean matching (PMM; Little, 1988; Rubin, 1986) as well as imputation methodology (PVs). It provides a method for score projections where equating methods are not defensible as they would make unrealistic assumptions such as equivalency of constructs and reliability levels.

10.1 Predictive Mean Matching (PMM)

Predictive mean matching (*PMM*) (Little, 1988; Rubin, 1986) finds a small number of 'donor' observations based on a predicted value generated by an imputation model. Assume that a number of observed variables is available as a predictor set $Z_1,...,Z_K$ and that an imputation model was specified to predict the conditional distribution of a variable θ so that we can write the predictive distribution as

$$\Phi_z(\theta) = P(\theta \,|\, Z_{1v}, \ldots, Z_K)\,. \tag{10.1}$$

PMM replaces a missing observation $\boldsymbol{\theta_v}$ of a respondent v by defining the predictive mean of this respondent as

$$\boldsymbol{\hat{\theta}_v} = E\,(\theta \,|\, Z_{1v}, \ldots, Z_{Kv}) \tag{10.2}$$

finding a small number of 'donor' observations by selecting these. $m_{1v} = m_1(\boldsymbol{\hat{\theta}_v}),...,\mathrm{m}_{Lv} = \mathrm{m}_L(\boldsymbol{\hat{\theta}_v})$ based on their distance to $\hat{\theta}_v$. That is, the goal is to find the set of L donors with the smallest distances to the predicted mean so that

$$|\boldsymbol{\theta_m} - \boldsymbol{\hat{\theta}_v}|\; for\; m \in \{m_{1v}, \ldots, m_{Lv}\} < |\boldsymbol{\theta_m} - \boldsymbol{\hat{\theta}_v}|\; for\; m \in \{1, \ldots, N\} \setminus \{m_{1v}, \ldots, m_{Lv}\} \tag{10.3}$$

This can simply be achieved by sorting all observations according to this distance and choosing the L observations with the smallest differences. Finally, these closest observations

$$\{\theta_{m_{1v}}, \ldots, \theta_{m_{Lv}}\} \tag{10.4}$$

are taken as the imputed values for the missing observation $\boldsymbol{\theta_v}$.

The advantage of PMM over other methods of imputation can be described as the 'realism' in the imputed values. The predicted mean given in equation (10.2) can be out of range, say if a constrained range sum score on a test is imputed, while the imputed (donated) set of values given in (10.4) is not only guaranteed to be within range, but also to follow other features of the observed distribution. For example, if the sum score is discrete, either if classical test theory (CTT) or a Rasch model was used, the donated values will be discrete scores as well, while the predicted conditional means, and the draws from the posterior used for imputation will, in general, not be discrete. Or, if the 'true' observed distribution is censored, skewed, or bimodal, the donated values will mimic these features, while this is typically not the case when using a parametric form for the posterior distribution selected for generating imputations.

10.2 Technical Procedure for Establishing Concordance Tables

The technical procedures described in this section draw on the statistical principles of conditioning models used in PASEC, TIMSS and PIRLS, (e.g., von Davier, Gonzalez & Mislevy, 2009; von Davier & Sinharay, 2013). This allows constructing a *concordance enhanced by conditional variance estimates* to properly account for uncertainty and can be described as follows:

1. The predictive means of source test score θ and target test score ϑ are derived utilizing population models as described in the previous sections 8 and 9. The expected value given responses and context is given by

$$\hat{\vartheta} = E(\vartheta | Y_1, \ldots, Y_J, Z_1, \ldots, Z_K) \text{ and } \hat{\theta} = E(\theta | X_1, \ldots, X_I, Z_1, \ldots, Z_K) \qquad (10.5)$$

2. The conditional distribution is available for generating imputations for ϑ for those cases where only test $X_1,\ldots, X_I$ is given together with the context variables can be constructed if ϑ is known for a sample, so that the conditional distribution

$$P(\vartheta | X_1, \ldots, X_I, Z_1, \ldots, Z_K) \qquad (10.6)$$

   becomes available for generating imputations.

3. For a concordance, the full population model using individual responses and context variables is often impractical. Practitioners want to use a score on one test to make inferences about the likely score range on another test. Note this is always projection-based using joint or conditional distributions, and the use of just a point estimate on the target test form given the source test score would be ignoring the uncertainty around this projected score. Therefore, the approach used here utilizes PVs (obtained from population models) to account for the uncertainty of the score projection.

2. The observed joint distribution of source and target test latent variable estimates can be used to create a conditional (predictive) distribution of the target test's latent variable given the source test's variable, $P(\vartheta | \theta)$. Based on a sample of respondents $v = 1, ..., N$, plugging in the posterior means and PVs allows us to approximate this conditional distribution. Instead of the full population model

$$\hat{\vartheta} \sim P(\vartheta | X_1, \ldots, X_I, Z_1, \ldots, Z_K) \tag{10.7}$$

an approximate imputation model $P(\vartheta|\theta)$ based on the source and target latent variables only is used and estimated using the two full population models

$$\hat{\vartheta} \sim E(\vartheta | Y_1, \ldots, Y_J, Z_1, \ldots, Z_K) \tag{10.8}$$

and

$$\bar{\theta} \sim E(\theta | X_1, \ldots, X_I, Z_1, \ldots, Z_K) \tag{10.9}$$

to generate an estimate of the conditional distribution

$$P(\hat{\vartheta} | \hat{\theta}) \approx P(\vartheta | \theta) \tag{10.10}$$

5. Then, the concordance is essentially given by

$$P(\hat{\vartheta} | E(\theta | X_1, \ldots, X_I, Z_1, \ldots, Z_K)) \tag{10.11}$$

and provides a projected distribution on the target test form given a function of the context variables and observed responses on the source test form.

6. A practical implementation of estimating this concordance can be implemented as:
   a. Draw $m = 1, \ldots, M$ PVs $\hat{\vartheta}_{mn}$ on the target test form for all respondents $n = 1, \ldots, N$.
   b. Estimate the posterior means

$$\bar{\theta}_n = E(\theta | X_{1n}, \ldots, X_{In}, Z_{1n}, \ldots, Z_{Kn}) \tag{10.12}$$

   for all respondents n = 1,..., N.

   c. Select a concordance range of source scores $\Omega = \{\theta_0 < \theta_1 < \cdots < \theta_{R-1} < \theta_R\}$ that covers 99% or more of the $\bar{\theta}$, i.e., so that $P(\theta_0 < \bar{\theta} < \theta_R) > 0.99$.

   d. For each of these concordance table scores $\theta_r \in \Omega$, select a set of L donors $d_{1r}, \ldots, d_{Lr}$ that have the smallest distances to the concordance table score θ_r. That is $|\theta_m - \theta_r| \; for \; m \in \{d_{1r}, \ldots, d_{Lr}\} < |\theta_k - \theta_r| \; for \; k \in \{1, \ldots, N\} \setminus \{m_{1r}, \ldots, m_{Lr}\}$.

e. Use the PVs $\hat{\vartheta}_{1d_{1r}},\ldots,\hat{\vartheta}_{md_{1r}},\hat{\vartheta}_{1d_{2r}},\ldots,\hat{\vartheta}_{md_{2r}},\hat{\vartheta}_{1d_{Lr}},\ldots,\hat{\vartheta}_{md_{Lr}}$ as the predictive distribution of scores on the target test ϑ given concordance score θ_r.

10.3 Advantages of the Enhanced Concordance Method

The *enhanced concordance method* described above provides an estimate of the conditional distribution $P(\vartheta|\theta)$, using imputed scores (PVs) on the target test ϑ, given a model-based point estimate on the source test θ. This model-based point estimate is a posterior mean given the available information of the source test and condenses a complex imputation model for the target test score into a single value that can be used in a concordance table.

The use of PMM finds donors in each sample that are nearest neighbors to the concordance table scores and assigns their target test PVs as the projection of scores based on the closest estimates obtained when only taking the source and target test forms, respectively. This approach ensures that score uncertainty due to measurement error and due to the imperfect correlation between source and target test are appropriately taken into account. In addition, when aggregating multiple population-based concordances, the uncertainty due to variability among countries is appropriately incorporated.

An additional advantage of the approach is that no functional form is assumed for the concordance other than those used to estimate the imputation models for source and target test forms. Commonly used equating and linking methods assume that the construct being measured is the same in source and target test forms, and project a point estimate on the source form onto a point estimate of the target form. Even if a transformed standard error would be used in addition, this would still assume that the constructs are essentially the same. In the proposed method, however, the estimated conditional distribution based on within subject repeated measurement of different tests is used, so that the dependencies (or lack thereof) between source and target test forms are directly incorporated in the enhanced concordance.

Moreover, there is no linearity assumption, and no other functional relationships between source and target test scores assumed other than the one that comes 'naturally' by utilizing multiple donors that are closest neighbors to the concordance table scores. The number of donors and how they are weighted and selected can increase smoothing effects, and the approach followed here uses 5 nearest neighbors per country-specific sample.

11. Establishing an Enhanced Concordance between PASEC and TIMSS/PIRLS

This section describes the procedures used to construct the Rosetta Stone concordance tables which provide a projection of the scores on the PASEC source assessment on the scales of the TIMSS and PIRLS target assessments.

11.1 Relationship between PASEC data and Rosetta Stone Linking data

As the PASEC mathematics and reading PVs provided by the PASEC team were on the IRT logit metric, they first needed to be transformed to the PASEC reporting metric to make the concordance tables more interpretable and meaningful. This was done by applying the PASEC linear transformation constants. The means of the transformed PASEC PVs for all three countries were compared to the corresponding values in the PASEC 2019 technical report (PASEC, 2020) for quality assurance. Overall, the transformed PASEC PVs were highly consistent with the reported PASEC 2019 scores at the individual country level with mean differences being in the range between 0 to 10 points. Only one country (Guinea) had a slightly larger mean difference between the transformed PASEC mathematics mean and the reported mathematics score in the PASEC 2019 technical report, which is most likely due to sampling variability.

To check the relationship between the data from source and target assessments, the correlations between the posterior means of PASEC data and Rosetta Stone linking data for mathematics/numeracy and reading/literacy were examined. For the PASEC mathematics and reading tests, the posterior means were not available and needed to be approximated. This was done by averaging the five PVs from the PASEC mathematics scale and the five PVs from the PASEC reading scale, respectively. The correlations between the posterior means of PASEC data and Rosetta Stone linking data are presented in the table in Exhibit 11.1.

Exhibit 11.1: Correlations between PASEC Data and TIMSS and PIRLS Linking Data

Country	PASEC Mathematics with TIMSS	PASEC Reading with PIRLS
Burundi	0.73	0.73
Guinea	0.70	0.74
Senegal	0.80	0.81

Correlations in Exhibit 11.1 approach the latent correlations from the multidimensional IRT models illustrated in section 7.3 and indicate that PASEC and Rosetta Stone scales measure different but similar constructs; that is, correlations are reasonably high for constructing a concordance.

For quality control, the cumulative distributions of the PASEC and the Rosetta Stone posterior means were approximated by averaging the five PVs of the corresponding assessments for each country and are illustrated in Exhibit 11.2. and Exhibit 11.3. Dis-ordinal interactions are shown between the posterior means of PASEC data and Rosetta Stone linking data among the three countries. This finding is consistent with the finding in Exhibit 5.1 and, according to feedback from the participating countries, could potentially be due to curricula differences among countries, and differences between test language and language spoken at home or in the classroom. In addition, the score ranges are quite different across countries which may also contribute to variability of the results across countries.

Exhibit 11.2: Cumulative Distributions of PASEC Mathematics and TIMSS Linking Data

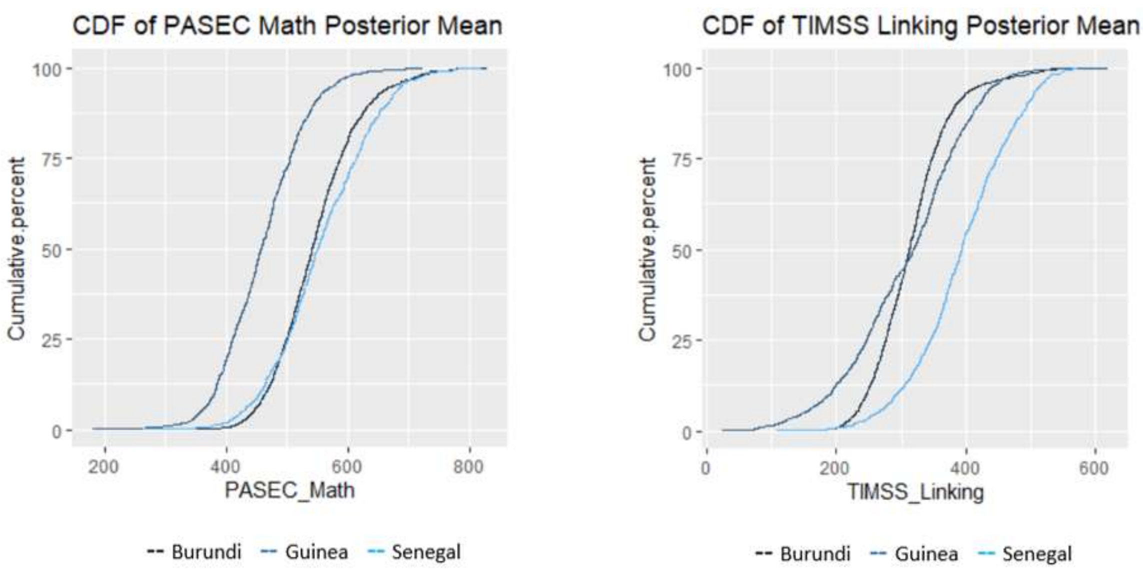
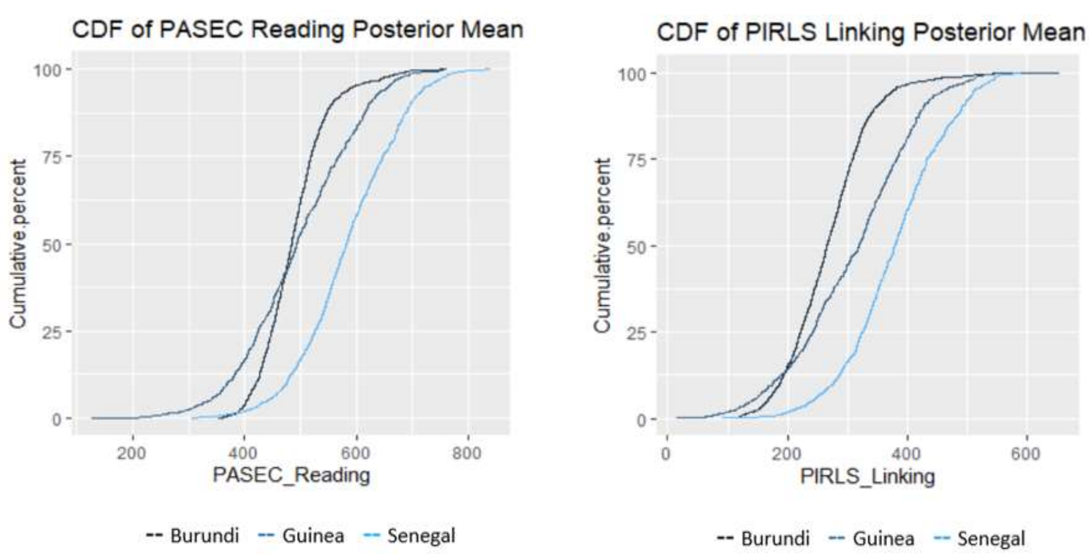


Exhibit 11.3: Cumulative Distributions of PASEC Reading and PIRLS Linking Data



A joint concordance table was constructed by aggregating the data across countries as country-level differences should not affect projected score averages but be reflected in the variability of projections. As a tool for international comparable assessments, the concordance should form the basis for comparisons regardless of the countries used to construct the projection table. This was done by using PVs for all participating countries in a combined table, one for mathematics and one for reading. Joint concordance tables account for the uncertainty in the measurement (i.e., the measurement error), country-specific effects due to sampling and other nuisance variables, and the imperfect correlation between PASEC data and Rosetta Stone linking data.

11.2 Creating Preliminary Concordance Tables

The concordance scores and levels were identified based on estimated PASEC posterior means using the combined data of the three countries. The score ranges of the posterior means of the PASEC mathematics and reading scales were either rounded up or down to cover almost all the data of the three countries and to be as symmetric as possible around the overall mean of the PASEC scale (which is 500). For both PASEC scales, mathematics and reading, scores range from about 200 to 800 (covering almost 100% of the data) with very few data points beyond the range of 260 to 760 (covering about 99.5% of the data). Therefore, the following description of creating the concordance tables primarily focusses on the scores within the range from 260 to 760.

For both PASEC scales, mathematics and reading, 20 points on the PASEC reporting metric 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 260 to 760.

For each identified concordance score level, PMM was used to select 5 donors from each of the three countries so that each country contributes equally to each of the concordance tables. Each of the donors donated 5 PVs on the target tests. This selection was achieved by selecting the 5 smallest absolute differences of students' posterior mean on the PASEC test to each specified concordance score for each country. The mean and standard deviation of the donors' PVs from the Rosetta Stone linking data were calculated based on the total 75 donated PVs (3 countries × 5 donors × 5 PVs) at each concordance level. Note that these steps were implemented separately for PASEC mathematics and reading.

Preliminary concordance tables for PASEC mathematics and PASEC reading were created by assigning the estimated mean and standard deviation of each set of 75 PVs based on the TIMSS and PIRLS linking data, respectively, to each concordance score level in the specified range of PASEC mathematics and PASEC reading.

11.3 Smoothing and Extrapolating the Concordance Tables

To examine the distribution of the donated PVs on the target tests, boxplots of each set of 75 donated PVs were produced for each concordance score level between the range of 260 and 760 on the PASEC source tests. They are presented in Exhibits 11.4 and 11.5 for mathematics and reading, respectively.

The conditional means of the donated PVs on the target scales show that generally higher means are related to higher concordance scores for both mathematics and reading. Because of the volatility due to the limited number of countries, the smaller sample sizes and the dis-ordinal interaction effects among countries, a smoothing procedure was used to better represent the underlying projected conditional means and standard deviations on the target scales.

Exhibit 11.4: Boxplots of Plausible Values (PVs) from Selected Donors for Mathematics

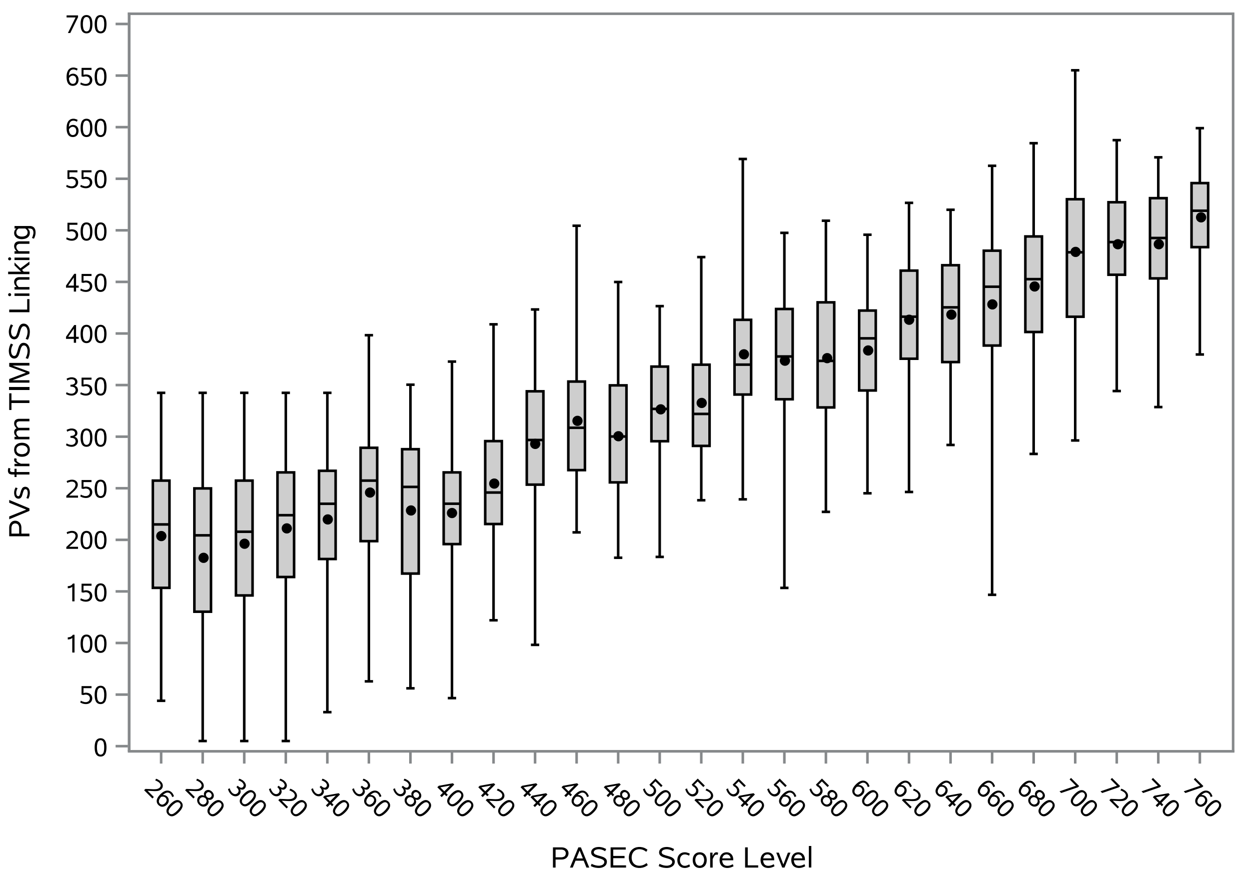
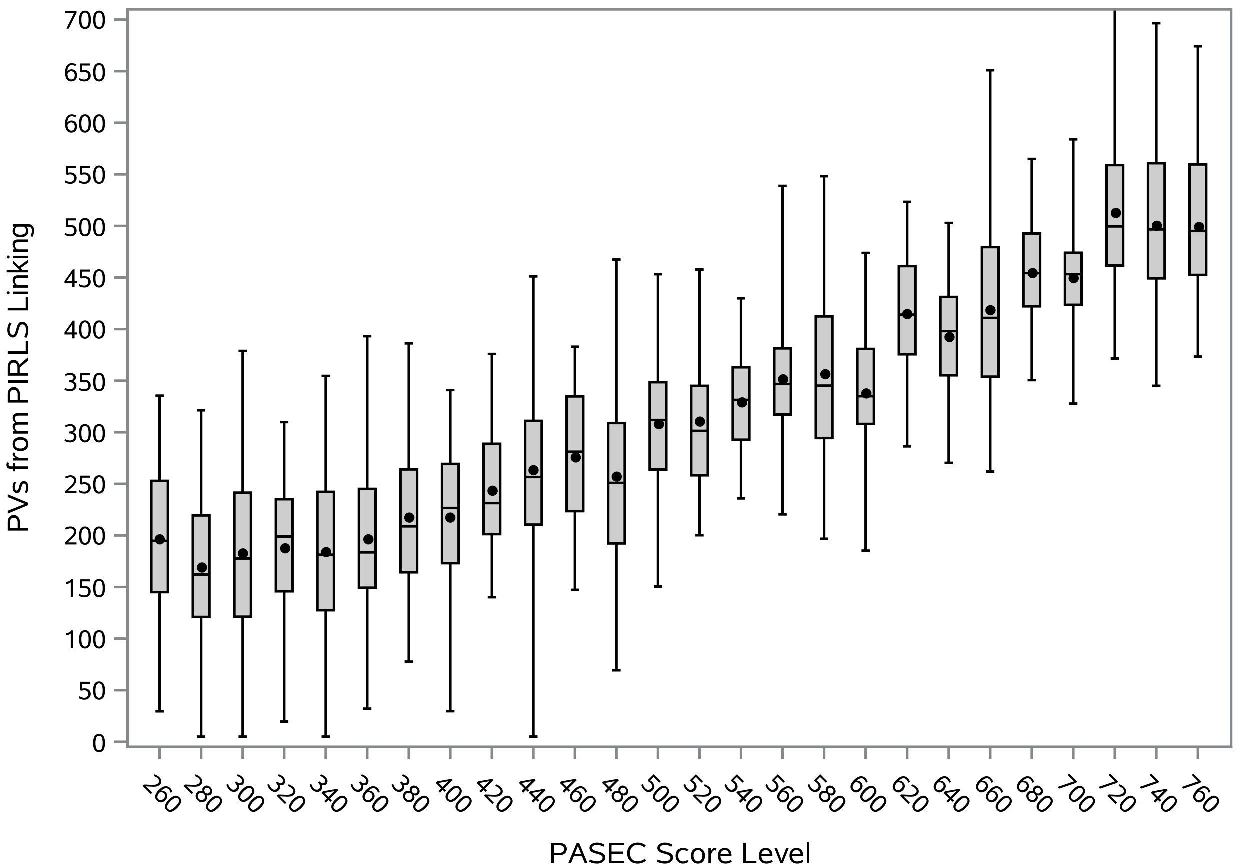


Exhibit 11.5: **Boxplots Plausible Values (PVs) from Selected Donors for Reading**



For each concordance score point, the mean of the donated PVs was smoothed by applying a simple moving average (e.g., Isnanto, 2011) using a window of 7 score points. The standard deviation of PVs of each score point was smoothed in a similar way as the means of PVs, using a moving geometric mean of variances of each set of the 7 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.

To obtain a robust prediction for PASEC concordance scores beyond the range of 260 to 760, where only a very small number (less than 0.5%) of students was observed, a non-parametric regression method called Sen's slope estimator (or the Thiel-Sen estimator; Sen, 1968) was used to extrapolate for two more concordance score levels at both extreme ends. To calculate the Sen's slope estimator for the predicted mean, the median of all slopes for all pairs of ordered (ordinal) PASEC score levels and the smoothed means were used to predict the conditional means of the likely posterior distributions at the concordance score levels 220, 240, 780, and 800. Similarly, the median of all slopes for all pairs of ordered score levels and the smoothed standard deviations were used to predict the conditional standard deviations of the likely posterior distributions at the two tails of the distribution.

Exhibits 11.6 and 11.7 show the final concordance tables for PASEC mathematics and PASEC reading, respectively. The first column of each table shows the PASEC concordance score levels, either PASEC mathematics or PASEC reading. The second and third columns show the projected means and standard deviations of the conditional distribution of the latent variable on the TIMSS or PIRLS scale given the PASEC score level. The fifth and sixth columns show the lower and upper bounds of the range in which 68% of the students should fall on the TIMSS and PIRLS scale for a given PASEC score level. The fourth and seventh columns show the lower and upper bounds of the range in which 95% of the students should fall on the TIMSS and PIRLS scale for a given PASEC score level.

Exhibit 11.6: **Concordance Table for PASEC Mathematics**

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

Exhibit 11.7: Concordance Table for PASEC Reading

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

As an example of the usefulness of the concordance tables, the percentages of students in each country reaching or exceeding the TIMSS[1] and PIRLS[2] lower benchmarks for mathematics and reading at grade 4 (indicated by a score of ≥ 400) were estimated and are illustrated in Exhibit 11.8. Percentages were estimated for two sets of PVs: the PVs generated based on the Rosetta Stone assessment part (TIMSS and PIRLS linking booklets) and the projected PVs based on the concordance tables.

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.

Overall, Exhibit 11.8 shows that while there is variability in countries' separate estimated percentages when comparing the concordance-based estimates with the Rosetta Stone part (TIMSS and PIRLS linking booklets) based estimates, the average percentages across countries provides highly comparable results. The different trend between the two parts of the table we see for Burundi (compared to the other countries) matches the patterns we see in Exhibits 5.1, 11.2 and 11.3. Without in-depth analyses of country experts, we cannot speculate regarding the source of these observed differences based on the available data. However, based on feedback from the participating countries, potential sources could be curricula differences among countries, and differences between test language and language spoken at home or in the classroom. It also should be noted that the distributions of PVs across countries are very different with Burundi showing a narrower range of imputed scores.

1 A description of the TIMSS 2019 grade 4 mathematics benchmarks can be found here: https://timss2019.org/reports/achievement/#math-4

2 A description of the PIRLS 2016 grade 4 reading benchmarks can be found here: http://timssandpirls.bc.edu/pirls2016/international-results/pirls/performance-at-international-benchmarks/pirls-2016-international-benchmarks/

The variations across countries seen in Exhibit 11.8 could be related to the following limitations. First, the constructs that are measured with the Rosetta Stone assessment are not identical with the constructs measured by the PASEC assessment as indicated by the imperfect correlations between the scales. Second, the Rosetta Stone linking booklets mainly covered lower difficulty levels to adjust the assessment to participating countries. Third, the estimates are based on three countries only and somewhat smaller sample sizes per country (approximately 2,000 students), which are commonly used in national samples, compared to the full TIMSS and PIRLS assessments (approximately 4,500 students). Fourth, the scaling approach does not account for potential linking error. Therefore, the concordance should be interpreted with caution. Larger national sample sizes and adding more countries in the Rosetta Stone study would likely stabilize this estimated concordance more.

12. How to Use and Interpret the Concordance Tables

Concordance tables 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 tests 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, assuming we have approximately normal conditional score distributions, 68% of the generated PVs on the TIMSS scale would likely fall in the score range of 273 and 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 Exhibit 11.6. Appendix A and Appendix B provide examples of 100 randomly generated PVs based on the projected means and standard deviations of the conditional distributions in the PASEC concordance table for mathematics and reading.

Besides making inferences about the likely score range on TIMSS or PIRLS scales given a PASEC score, practitioners could also generate the likely PVs for individual students on the TIMSS and PIRLS scales by using the projected means and standard deviations from the concordance tables. To generate random PVs for the students who participated in the PASEC assessments, first, the posterior mean of the conditional distribution for each student from the PASEC population model needs to be obtained and transformed onto the PASEC reporting metric. Next, the posterior means are rounded to the nearest PASEC score levels as shown in the first column of Exhibits 11.6 and 11.7, so that the projected means and standard deviations could be assigned to individual students according to the rounded PASEC score levels. Then, the PVs are imputed based on the assigned projected mean and standard deviation of the conditional distribution for each student. There are a few ways to impute PVs based on these projected conditional means and standard deviations. In the examples shown in Appendix A and Appendix B, PVs were imputed using the "inverse of normal cumulative distribution" function in Excel. PVs for individual students can also be imputed using a normal distribution with the corresponding conditional mean and standard deviation in SAS, R Packages, and other software tools.

Concordance tables can only provide likely projections of distributions of source test scores on a target test and, therefore, have to be understood and interpreted with caution. Differences in the measured constructs, differences in construct coverage, smaller sample sizes, linking error or curricular differences across countries result in larger conditional variance in the projections compared to equated scores on two essentially equivalent test forms that measure the same construct. Nevertheless, concordance tables provide useful and valuable information when used and interpreted correctly. Countries that participated in the Rosetta Stone linking study and administered the Rosetta Stone linking booklets can project their students' PASEC score distributions on the TIMSS and PIRLS scales. For countries which did not participate in this study and did not administer the linking booklets, the use of the concordance tables provided in this report will be an extrapolation and comes with some added uncertainty that cannot be accounted for without also conducting a Rosetta Stone data collection. Therefore, such countries are encouraged to contact IEA for possible participation in a Rosetta Stone study to obtain updated concordance tables that account for their student-specific variability in the measurement. Moreover, larger national sample sizes and adding more countries in the Rosetta Stone study will further improve the estimated concordance.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443-459. https://doi.org/10.1007/BF02293801

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the environment. *Journal of statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Fischer, G. H. (1981). On the existence and uniqueness of maximum-likelihood estimates in the Rasch model. *Psychometrika*, *46*(1), 59-77. https://doi.org/10.1007/BF02293919

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 12.1-12.146). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-12.html

Foy, P., & Yin, Y. (2017). Scaling the PIRLS 2016 Achievement Data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in PIRLS 2016* (pp. 12.1-12.38). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-12.html

Haberman, S. J. (2005). *Identifiability of parameters in item response models with unconstrained ability distribution* (ETS Research Report Series RR-05-24). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2005.tb02001.x

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions.* Educational Testing Service RR-08-45. Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02131.x

Isnanto, R.R. (2011). Comparation on Several Smoothing Methods in Nonparametric Regression. *Jurnal Sistem Komputer*, 1(1), 41-47.

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software mdltm including parallel EM algorithm. In M. von Davier & Y. S. Lee (Eds.), *Handbook of psychometric models for cognitive diagnosis* (pp. 603–628). Springer. https://doi.org/10.1007/978-3-030-05584-4_30

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, *49*(4), 764-766. doi: 10.1016/j.jesp.2013.03.013

Little, R. J. A. (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics, 6* (3), 287–296. doi:10.2307/1391878.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: J. Wiley & Sons. https://psycnet.apa.org/record/1968-35040-000

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley. https://psycnet.apa.org/record/1968-35040-000

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. https://timssandpirls.bc.edu/timss2019/methods

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196. https://doi.org/10.1007/bf02296272

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-162. https://doi.org/10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal Estimation Procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983–84 technical report* (No. 15-TR-20, pp. 293–360). Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress. https://files.eric.ed.gov/fulltext/ED288887.pdf

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

PASEC (2020). *PASEC 2019 Qualité des systèmes éducatifs en Afrique Subsaharienne Francophone*. CONFEMEN: Dakar (Sénégal). Retrieved from: https://www.confemen.org/wp-content/uploads/2020/12/RapportPasec2019_Web.pdf

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen & Lydiche (Expanded edition, Chicago, University of Chicago Press, 1980).

Reckase, M. D. (2009) Multidimensional Item Response Theory (Statistics for Social and Behavioral Sciences), New York, NY: Springer.

Rogers, A., Tang, C., Lin, M. J., & Kandathil, M. (2006). DGROUP [Computer software]. Princeton, NJ: Educational Testing Service.

Rubin, D. B. (1986). Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. *Journal of Business & Economic Statistics, 4* (1), 87–94. https://doi.org/10.2307/1391390

San Martín, E., González, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika, 80*(2), 450-467. https://doi.org/10.1007/s11336-014-9404-2

Sen, P. K. (1968), Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, Vol. 63, No. 324. 1379-1389. https://www.pacificclimate.org/~wernera/zyp/Sen%201968%20JASA.pdf

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*, 309–322. https://doi.org/10.2307/1390648

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05-16). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01993.x

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspectives*, 7(2), 110-114. https://doi.org/10.1080/15366360903117079

von Davier, M., & Bezirhan, U. (2021, December 23). A Robust Method for Detecting Item Misfit in Large Scale Assessments. https://doi.org/10.31234/osf.io/mnsdg

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments* (Vol. 2, pp. 9-36). Retrieved from https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf von Davier, M., Rost, R., & Carstensen, C. H. (2007). Introduction: Extending the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 1-12). New York, NY: Springer. https://doi.org/10.1007/978-0-387-49839-3

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item Response Theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155-174). Boca Raton, FL: CRC Press. https://doi.org/10.1201/b16061-12

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26: Psychometrics). Amsterdam, Netherlands: Elsevier. https://doi.org/10.1016/S0169-7161(06)26032-2

von Davier, M., & Yamamoto, K. (2004). *A class of models for cognitive diagnosis*. Paper presented at the Fourth Spearman Conference, Philadelphia, PA. https://www.researchgate.net/publication/257822207_A_class_of_models_for_cognitive_diagnosis

Xu, X., & von Davier, M. (2008). *Fitting the structured general diagnostic model to NAEP data* (ETS Research Report, RR-08-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02113.x

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1), 436-460. https://doi.org/10.1007/BF01180541

APPENDIX A

Example of Generated PVs based on the Concordance Table for PASEC Mathematics

PASEC Math	Projected Mean	Projected SD	PV 1	PV 2	PV 3	PV 4	PV 5	PV 6	PV 7	PV 8	PV 9	PV 10	PV 11	PV 12	PV 13	PV 14	PV 15	PV 16	PV 17	PV 18	PV 19	PV 20
220	197	76	73	233	170	108	80	182	199	218	298	192	228	133	246	137	237	144	181	142	212	248
240	210	76	384	345	246	193	235	203	136	170	231	189	226	162	96	210	310	185	250	89	203	188
260	223	75	143	243	269	64	235	217	314	292	311	167	282	222	283	138	308	146	155	26	112	252
280	224	75	183	300	89	254	223	307	262	156	148	194	248	216	255	130	332	201	262	239	194	364
300	229	77	217	313	99	315	297	184	254	189	141	232	164	151	285	229	248	180	144	310	145	177
320	231	76	203	140	236	312	192	149	345	261	253	341	272	181	233	107	220	264	100	211	180	221
340	235	75	169	252	298	264	306	394	267	203	138	287	133	429	302	174	108	306	266	167	210	288
360	241	71	213	203	223	235	227	215	205	208	186	367	349	272	215	57	308	198	158	313	245	278
380	254	69	187	171	225	205	333	250	294	180	280	183	232	324	218	195	362	228	321	179	288	256
400	265	68	143	370	270	264	287	274	265	293	189	180	190	340	277	364	261	290	295	288	151	369
420	273	68	253	284	272	306	145	268	228	362	168	288	308	265	374	284	268	401	234	194	330	203
440	284	65	225	342	245	172	345	214	226	255	327	319	400	241	252	144	291	214	274	338	154	231
460	297	62	395	280	444	313	259	253	364	319	268	288	358	348	289	355	286	242	331	255	294	312
480	315	61	256	293	231	343	254	265	282	188	223	323	325	373	409	229	240	279	251	273	386	242
500	336	63	342	369	233	204	322	437	414	476	409	429	366	338	300	358	288	376	304	414	336	294
520	344	65	362	343	291	268	328	379	368	466	400	415	378	347	408	291	331	279	331	437	371	418
540	355	64	231	263	274	398	300	289	477	388	334	392	232	432	342	427	373	403	326	408	393	349
560	371	63	348	399	303	396	374	433	452	363	276	381	402	413	419	408	337	339	327	332	331	320
580	382	66	378	393	404	380	446	427	260	450	349	468	354	390	427	259	391	299	285	289	397	450
600	395	70	473	361	388	290	352	411	424	445	389	376	273	402	339	347	396	377	413	272	456	392
620	403	71	293	274	433	368	462	325	339	428	310	617	558	360	425	468	416	368	351	340	344	357
640	417	71	367	527	437	439	404	416	380	348	417	408	409	437	258	338	354	387	318	503	542	275
660	437	69	509	360	357	376	515	458	466	351	452	450	423	461	395	466	418	505	345	384	440	415
680	453	69	383	508	399	496	298	348	466	482	516	423	492	458	512	377	522	465	534	511	401	473
700	469	67	510	446	506	443	329	502	548	513	483	610	522	598	525	398	427	514	468	567	449	439
720	484	64	481	470	613	445	554	544	560	503	471	488	382	492	490	529	520	494	457	400	555	433
740	500	57	529	547	415	503	525	554	534	501	512	523	462	484	550	479	516	477	410	431	446	438
760	513	52	473	589	475	512	358	393	536	394	488	470	553	481	504	429	503	625	514	453	428	448
780	526	52	464	495	502	540	434	528	561	453	591	565	477	553	588	508	481	609	517	443	644	567
800	539	51	561	391	491	517	545	619	461	512	554	539	507	551	553	527	519	509	617	486	556	494

PASEC Math	Projected Mean	Projected SD	PV 21	PV 22	PV 23	PV 24	PV 25	PV 26	PV 27	PV 28	PV 29	PV 30	PV 31	PV 32	PV 33	PV 34	PV 35	PV 36	PV 37	PV 38	PV 39	PV 40
220	197	76	299	254	88	151	282	273	158	311	223	356	121	256	233	176	202	208	186	147	171	191
240	210	76	172	232	184	254	51	187	164	174	182	223	324	238	150	240	291	404	195	210	277	316
260	223	75	108	232	234	265	193	221	214	248	124	204	163	149	171	152	136	356	205	147	116	236
280	224	75	228	244	143	279	164	135	324	148	255	196	125	214	137	185	333	272	236	164	201	228
300	229	77	253	440	49	243	279	147	308	248	222	120	315	164	60	192	169	303	156	257	195	162
320	231	76	325	291	236	261	260	260	123	96	253	221	218	191	143	193	224	203	214	283	219	117
340	235	75	242	148	315	242	300	232	272	250	173	200	267	270	203	198	294	182	438	347	240	464
360	241	71	306	322	73	167	261	196	188	325	166	200	326	243	295	254	382	268	290	256	292	144
380	254	69	189	220	266	202	191	286	203	311	247	142	297	248	189	338	126	216	365	165	160	354
400	265	68	197	270	312	172	124	349	176	154	231	264	372	344	253	178	315	210	353	327	266	205
420	273	68	267	239	254	250	330	344	182	164	316	311	371	314	296	284	246	295	357	364	340	303
440	284	65	171	191	283	250	313	244	214	262	327	210	257	241	247	298	233	282	328	344	231	340
460	297	62	277	389	328	289	237	398	258	179	257	247	277	236	250	366	415	313	457	240	212	329
480	315	61	386	260	253	259	257	335	354	268	318	235	376	337	437	425	262	347	385	478	260	267
500	336	63	332	267	331	348	277	347	230	288	397	349	316	472	334	352	316	267	358	253	277	309
520	344	65	309	291	442	278	376	463	323	406	265	287	244	377	246	364	223	323	367	442	338	373
540	355	64	357	351	348	294	372	346	305	319	415	307	244	219	432	378	343	320	217	286	396	375
560	371	63	441	386	437	412	250	385	465	388	338	340	389	261	340	348	264	366	347	381	349	412
580	382	66	353	442	367	299	426	452	351	318	380	560	416	398	303	310	354	412	424	322	378	269
600	395	70	427	534	525	389	396	401	327	451	381	551	515	440	417	502	457	389	406	378	351	370
620	403	71	470	408	280	376	340	420	373	305	479	389	358	475	337	413	454	430	391	422	360	355
640	417	71	296	419	568	422	334	586	364	430	269	389	310	312	439	527	270	416	465	427	309	464
660	437	69	418	375	475	404	379	545	432	536	461	413	537	501	313	485	384	561	468	347	477	404
680	453	69	342	415	543	413	450	423	509	529	364	483	389	519	472	505	498	437	378	467	434	518
700	469	67	478	359	416	482	600	506	424	419	417	439	426	384	465	475	408	522	460	538	547	606
720	484	64	503	522	491	399	587	454	449	436	389	434	443	496	407	484	477	432	388	481	503	469
740	500	57	422	455	563	525	557	495	458	513	439	476	560	571	513	545	601	544	521	449	563	541
760	513	52	543	615	472	518	464	589	467	576	553	493	504	460	462	539	553	552	555	543	591	542
780	526	52	470	563	425	493	549	516	480	549	529	516	493	547	534	653	552	538	564	538	493	554
800	539	51	496	573	526	514	436	524	487	551	534	506	512	553	543	460	470	515	531	571	581	518

PASEC Math	Projected Mean	Projected SD	PV 41	PV 42	PV 43	PV 44	PV 45	PV 46	PV 47	PV 48	PV 49	PV 50	PV 51	PV 52	PV 53	PV 54	PV 55	PV 56	PV 57	PV 58	PV 59	PV 60
220	197	76	196	199	187	161	189	54	103	202	263	194	124	118	193	300	163	126	131	222	206	332
240	210	76	205	154	267	216	237	273	309	230	252	52	152	305	116	207	219	357	151	145	234	201
260	223	75	301	329	334	100	164	218	302	272	238	294	138	215	148	322	196	173	66	158	349	199
280	224	75	141	308	247	308	189	309	231	302	327	164	199	304	194	350	193	77	111	260	182	269
300	229	77	339	275	146	350	162	423	348	142	249	271	360	342	183	240	219	255	222	315	386	210
320	231	76	262	133	233	129	262	189	306	241	350	261	225	156	127	219	62	121	303	129	264	259
340	235	75	215	306	181	159	243	151	341	235	316	238	184	295	168	256	261	283	199	225	271	314
360	241	71	440	215	136	210	282	344	241	120	244	233	67	192	273	217	334	183	245	300	181	220
380	254	69	299	202	266	243	308	282	287	252	331	270	367	334	297	193	129	398	238	400	330	283
400	265	68	165	148	222	143	176	238	275	163	418	350	260	240	246	275	328	329	361	320	217	214
420	273	68	146	278	346	217	345	330	335	124	325	314	258	301	435	310	337	123	292	221	163	286
440	284	65	260	292	407	297	281	194	318	283	189	240	309	223	397	199	180	310	312	370	331	232
460	297	62	197	302	254	313	291	259	328	313	230	282	294	210	338	335	352	245	315	304	237	355
480	315	61	336	384	285	268	313	318	395	320	385	356	471	414	173	270	262	190	269	288	341	326
500	336	63	297	449	385	340	284	396	415	312	243	300	325	308	343	311	330	387	262	257	410	331
520	344	65	369	313	364	187	161	243	354	253	438	329	262	356	483	307	309	371	352	330	252	390
540	355	64	325	395	396	407	262	301	346	380	355	270	377	405	404	334	285	293	328	425	318	377
560	371	63	428	229	491	340	337	428	262	458	346	404	394	368	355	415	281	497	344	413	376	372
580	382	66	366	395	325	332	467	452	377	407	286	446	355	389	442	428	431	401	296	450	333	408
600	395	70	370	456	268	329	464	327	406	410	606	427	327	305	375	392	414	313	418	360	330	429
620	403	71	329	335	399	538	449	414	391	377	322	393	410	229	412	454	431	405	418	447	472	285
640	417	71	322	334	452	354	474	546	496	400	414	350	539	380	429	530	463	437	307	428	355	353
660	437	69	519	376	522	404	409	358	462	523	571	441	549	404	439	464	366	357	446	461	374	310
680	453	69	387	369	413	301	586	433	531	422	475	412	520	405	396	382	427	459	434	401	488	334
700	469	67	393	406	518	505	418	306	405	482	433	445	412	462	463	343	487	466	426	557	555	516
720	484	64	482	548	507	480	486	485	700	503	413	497	449	421	477	445	346	594	489	441	488	395
740	500	57	451	490	567	514	509	467	449	478	584	461	436	552	538	506	560	415	519	488	563	542
760	513	52	558	480	460	538	440	496	583	567	475	463	482	519	460	552	518	443	490	485	655	476
780	526	52	536	472	468	520	622	552	562	465	523	586	499	479	496	596	575	493	551	443	572	452
800	539	51	439	508	578	631	561	462	571	530	627	544	419	444	605	552	497	513	468	540	586	568

PASEC Math	Projected Mean	Projected SD	PV 61	PV 62	PV 63	PV 64	PV 65	PV 66	PV 67	PV 68	PV 69	PV 70	PV 71	PV 72	PV 73	PV 74	PV 75	PV 76	PV 77	PV 78	PV 79	PV 80
220	197	76	16	90	238	150	122	310	136	114	158	333	194	143	265	207	268	75	109	290	135	152
240	210	76	90	211	142	244	234	140	150	90	271	162	334	201	133	192	118	213	236	275	228	68
260	223	75	214	130	236	264	225	219	164	156	221	233	207	123	231	167	163	136	265	50	111	239
280	224	75	338	376	70	276	304	150	335	365	276	267	319	244	104	249	297	221	217	218	238	160
300	229	77	365	249	233	269	258	233	197	251	206	265	313	97	136	266	139	236	125	282	200	111
320	231	76	208	216	226	369	264	287	167	164	42	339	176	264	169	183	204	296	335	178	342	144
340	235	75	415	156	331	270	328	281	150	258	132	174	172	230	185	339	273	241	249	266	300	121
360	241	71	227	238	224	282	235	192	131	146	205	210	151	168	472	273	301	213	237	186	179	170
380	254	69	220	190	317	119	327	311	184	304	336	202	325	338	212	91	244	155	223	325	315	278
400	265	68	226	257	238	361	294	122	222	291	219	328	370	204	205	214	188	312	172	255	82	281
420	273	68	365	351	180	251	383	267	322	307	302	349	379	142	333	165	206	259	210	319	196	324
440	284	65	297	351	289	275	378	227	279	351	304	259	220	343	246	260	262	359	384	341	151	297
460	297	62	304	216	329	296	263	276	275	290	285	329	276	287	316	231	287	295	330	323	224	230
480	315	61	301	265	190	291	421	353	383	343	349	284	256	456	367	250	246	264	394	299	306	374
500	336	63	522	401	343	396	128	405	298	275	405	264	232	313	360	186	306	319	337	328	207	294
520	344	65	291	295	373	358	379	162	433	320	337	296	387	362	298	170	368	435	470	289	384	438
540	355	64	237	416	344	237	399	266	308	237	290	295	264	410	357	353	384	391	295	443	448	328
560	371	63	341	332	300	393	415	500	414	322	397	247	489	308	385	225	431	390	357	319	325	353
580	382	66	435	485	356	339	436	339	451	319	339	375	417	377	340	462	304	406	370	463	376	334
600	395	70	300	409	448	322	465	373	486	315	267	328	421	356	345	468	263	422	327	340	446	386
620	403	71	393	417	456	334	404	388	434	402	486	569	442	465	488	475	400	533	417	445	422	429
640	417	71	572	340	366	319	254	396	296	397	475	494	284	396	480	474	441	352	349	554	462	344
660	437	69	345	444	295	442	513	505	422	527	425	354	538	503	519	439	460	494	493	333	435	359
680	453	69	446	431	463	484	357	475	359	415	377	471	392	354	530	553	443	357	296	467	515	448
700	469	67	439	509	619	481	563	447	433	502	394	328	333	401	445	528	557	502	342	319	334	406
720	484	64	416	537	410	356	435	589	525	529	431	472	495	462	486	512	404	478	434	482	455	471
740	500	57	545	481	445	480	387	439	536	510	574	481	584	638	586	403	556	541	534	406	415	459
760	513	52	569	502	521	380	489	515	470	521	594	511	513	509	496	458	481	504	483	446	525	537
780	526	52	416	509	496	420	492	539	536	516	576	518	534	554	624	577	545	485	494	593	591	586
800	539	51	557	516	460	457	626	478	549	558	560	498	552	582	524	570	653	684	463	527	528	537

PASEC Math	Projected Mean	Projected SD	PV 81	PV 82	PV 83	PV 84	PV 85	PV 86	PV 87	PV 88	PV 89	PV 90	PV 91	PV 92	PV 93	PV 94	PV 95	PV 96	PV 97	PV 98	PV 99	PV 100
220	197	76	348	167	198	17	181	236	182	121	185	185	54	237	247	395	176	208	223	111	189	299
240	210	76	130	110	145	214	216	168	52	187	160	151	121	299	191	144	119	282	213	196	201	159
260	223	75	197	165	248	204	203	65	316	227	305	270	223	138	227	35	146	225	310	300	204	198
280	224	75	188	151	188	292	191	277	326	220	181	104	147	176	211	249	103	204	356	335	227	172
300	229	77	226	178	289	3	363	288	226	127	233	248	170	167	191	77	223	136	256	142	223	214
320	231	76	254	275	342	169	77	277	169	294	242	211	221	161	307	317	343	160	212	202	221	104
340	235	75	157	240	227	178	297	328	181	220	255	171	224	294	201	250	283	365	289	284	247	205
360	241	71	368	165	289	214	159	233	194	129	258	270	129	249	267	314	99	336	280	236	232	215
380	254	69	224	180	257	294	146	171	185	234	287	379	266	265	264	276	266	310	205	192	251	283
400	265	68	254	217	356	294	273	181	175	296	229	199	278	218	361	286	248	144	297	267	253	272
420	273	68	249	212	221	331	317	236	357	205	238	281	326	231	244	344	261	292	227	205	275	254
440	284	65	259	234	359	376	222	362	88	301	395	322	299	280	297	343	317	317	328	304	277	265
460	297	62	216	395	382	376	293	280	282	330	375	236	285	294	240	402	235	318	210	279	294	266
480	315	61	216	286	378	286	284	289	329	224	269	302	202	348	247	322	249	194	269	308	303	368
500	336	63	302	311	267	354	300	393	393	361	230	324	341	351	399	377	329	435	384	396	331	349
520	344	65	290	451	366	362	341	233	314	274	298	373	384	372	376	391	414	325	352	338	339	391
540	355	64	373	380	440	296	334	336	380	335	522	333	272	284	354	399	364	234	311	240	339	456
560	371	63	366	417	329	544	420	527	436	214	425	441	231	344	385	360	449	470	291	462	370	264
580	382	66	385	352	506	383	424	362	330	461	423	356	412	316	247	514	385	355	404	404	380	446
600	395	70	458	383	339	450	327	452	260	613	411	289	378	396	518	376	399	510	461	394	393	420
620	403	71	395	557	264	292	451	414	387	452	378	340	407	266	454	428	344	647	372	320	400	400
640	417	71	430	403	507	344	406	513	380	539	400	379	540	463	378	391	446	456	386	393	406	457
660	437	69	411	512	426	348	362	415	493	560	450	459	369	554	355	433	433	423	430	475	435	465
680	453	69	456	452	441	533	374	458	551	493	484	467	426	447	483	455	495	380	438	476	442	382
700	469	67	511	350	456	486	558	396	613	456	562	432	408	367	538	482	540	445	500	337	460	528
720	484	64	545	542	552	525	547	536	562	463	401	499	391	340	472	510	420	422	530	462	474	519
740	500	57	573	480	454	484	487	511	535	461	423	445	567	594	461	493	486	480	487	355	496	392
760	513	52	570	499	531	552	510	598	453	494	634	562	601	535	485	536	427	506	536	561	505	538
780	526	52	455	470	564	454	568	465	573	497	438	561	580	537	568	456	520	437	460	502	518	534
800	539	51	482	444	511	415	589	622	648	573	525	556	473	486	527	502	465	496	533	449	523	492

APPENDIX B

Example of Generated PVs based on the Concordance Table for PASEC Reading

PASEC Reading	Projected Mean	Projected SD	PV 1	PV 2	PV 3	PV 4	PV 5	PV 6	PV 7	PV 8	PV 9	PV 10	PV 11	PV 12	PV 13	PV 14	PV 15	PV 16	PV 17	PV 18	PV 19	PV 20
220	146	72	98	211	139	186	144	121	2	149	177	219	229	219	150	136	79	81	185	135	289	192
240	161	72	114	195	200	213	162	127	154	242	228	289	143	205	149	57	121	142	72	71	268	173
260	175	72	98	175	165	195	-39	97	247	198	42	221	150	263	205	3	210	166	359	187	88	146
280	178	72	148	144	162	178	166	120	151	121	246	95	166	174	313	164	149	157	276	156	126	259
300	181	72	125	192	176	136	181	127	311	184	149	178	179	251	261	234	54	36	124	107	168	77
320	190	71	98	167	180	311	145	204	202	242	146	172	211	194	175	208	209	174	191	79	147	69
340	196	72	179	173	247	100	213	210	357	150	213	140	133	263	133	90	194	273	120	184	307	234
360	205	71	239	113	136	142	210	200	162	253	194	312	223	170	222	181	247	182	158	152	209	208
380	216	72	181	202	253	363	163	170	104	241	221	187	153	322	181	320	138	109	233	177	220	281
400	228	72	123	299	170	280	196	130	266	201	267	159	301	177	129	312	287	88	90	311	147	135
420	238	76	216	265	218	262	336	168	207	308	139	202	241	297	140	135	196	223	198	142	319	136
440	253	74	257	224	236	250	289	258	127	170	267	141	326	262	187	341	329	196	280	306	398	369
460	265	73	216	144	348	230	305	79	223	285	220	353	314	226	314	446	170	165	207	261	232	201
480	280	71	189	336	125	261	287	203	307	212	277	257	252	294	385	364	201	328	334	307	283	239
500	297	71	213	192	273	317	343	277	380	222	114	300	233	280	283	226	285	403	341	383	338	343
520	317	73	353	380	246	212	249	364	315	367	322	342	344	304	443	295	311	401	276	296	287	317
540	330	72	414	340	368	397	409	336	251	286	225	202	174	298	429	269	315	387	327	347	323	405
560	351	66	392	377	389	341	405	347	347	304	223	384	204	321	322	451	257	407	410	235	301	303
580	364	66	399	303	371	447	354	419	523	441	248	399	316	343	254	254	534	391	286	341	453	323
600	377	68	352	442	463	514	321	399	294	346	344	360	329	367	323	242	303	420	278	320	411	335
620	392	69	318	398	360	411	375	432	502	477	428	301	257	496	439	445	431	415	387	508	411	397
640	405	67	401	368	471	357	351	417	403	428	425	488	464	406	374	572	333	485	389	254	316	442
660	420	63	399	362	477	426	466	395	468	395	381	455	368	368	393	454	380	418	408	361	351	419
680	444	66	544	432	499	383	399	402	504	436	613	412	531	409	449	560	316	541	404	396	494	382
700	456	69	528	497	333	453	478	451	400	450	422	493	442	495	446	464	523	381	407	401	361	499
720	473	71	421	455	533	558	438	451	504	448	513	650	478	340	431	481	500	599	425	549	391	590
740	486	70	393	502	439	402	461	493	356	423	506	528	476	440	574	584	419	417	479	504	507	470
760	492	72	510	526	588	532	464	465	513	518	388	419	435	462	522	444	500	463	692	467	548	579
780	507	72	520	495	545	537	516	433	538	555	361	432	562	532	527	474	521	563	486	456	593	619
800	521	72	559	515	423	514	544	562	518	337	535	413	585	531	446	666	431	408	604	557	554	508

PASEC Reading	Projected Mean	Projected SD	PV 21	PV 22	PV 23	PV 24	PV 25	PV 26	PV 27	PV 28	PV 29	PV 30	PV 31	PV 32	PV 33	PV 34	PV 35	PV 36	PV 37	PV 38	PV 39	PV 40
220	146	72	176	96	107	7	255	100	57	214	154	254	223	206	216	215	138	217	119	110	171	208
240	161	72	136	134	253	179	210	143	82	-25	239	203	270	162	160	173	237	175	232	233	138	231
260	175	72	159	236	209	246	234	227	265	170	237	254	140	254	181	55	227	207	216	187	149	85
280	178	72	150	257	340	151	108	282	139	242	290	176	253	176	81	88	247	307	391	216	222	132
300	181	72	253	134	216	177	331	310	51	148	129	215	151	310	2	154	256	243	252	201	101	90
320	190	71	182	240	219	152	170	274	260	218	196	208	136	389	189	108	297	178	167	336	86	269
340	196	72	246	181	230	240	267	286	242	121	84	314	139	250	154	218	169	298	57	164	278	208
360	205	71	246	135	255	193	335	184	309	277	251	195	220	166	264	145	174	139	285	83	101	194
380	216	72	207	273	53	174	65	111	252	324	264	443	216	179	222	170	296	331	300	238	92	322
400	228	72	156	235	251	205	208	182	215	273	224	268	185	282	278	182	138	343	236	166	274	262
420	238	76	207	145	91	77	237	185	120	268	200	310	363	292	269	257	212	312	275	76	167	229
440	253	74	251	199	132	150	259	246	339	305	188	236	233	209	134	181	322	203	195	356	265	311
460	265	73	316	240	306	253	239	165	305	311	343	233	395	346	293	214	309	236	375	357	309	259
480	280	71	185	256	339	384	285	135	320	353	357	291	290	169	247	187	265	290	342	190	256	279
500	297	71	213	258	239	187	277	232	314	370	327	287	395	415	269	303	249	393	221	291	422	385
520	317	73	265	284	328	301	347	306	361	357	347	299	280	271	298	373	393	201	313	343	289	297
540	330	72	257	289	292	335	341	353	197	346	411	470	279	294	151	237	264	303	326	267	437	262
560	351	66	346	312	315	281	271	341	384	283	372	318	435	326	287	283	311	326	328	300	397	335
580	364	66	438	378	491	406	378	426	353	315	406	365	217	391	306	447	276	485	406	351	371	414
600	377	68	553	419	411	376	371	369	273	357	338	258	422	287	262	486	303	467	344	267	467	445
620	392	69	380	394	381	464	467	413	453	353	318	333	512	573	391	402	423	317	436	441	326	532
640	405	67	369	437	281	421	436	427	366	464	373	459	330	380	387	397	428	345	303	412	362	342
660	420	63	448	408	389	404	397	482	417	630	410	373	528	434	345	418	457	466	333	443	362	350
680	444	66	517	339	482	533	414	469	552	461	385	346	504	317	361	478	376	439	432	414	500	531
700	456	69	496	543	353	495	409	512	371	412	373	448	506	516	462	430	476	472	372	324	500	530
720	473	71	428	454	380	481	571	453	576	421	587	469	601	437	543	567	372	507	489	493	465	600
740	486	70	413	464	408	450	494	345	352	528	325	568	481	454	442	539	655	362	506	402	378	445
760	492	72	573	424	407	505	441	608	447	473	507	408	570	461	539	463	563	364	439	490	601	593
780	507	72	459	514	505	587	340	633	479	600	433	525	479	574	565	437	562	509	456	578	513	584
800	521	72	479	471	434	518	462	544	624	603	505	577	465	439	552	580	604	489	547	446	389	553

PASEC Reading	Projected Mean	Projected SD	PV 41	PV 42	PV 43	PV 44	PV 45	PV 46	PV 47	PV 48	PV 49	PV 50	PV 51	PV 52	PV 53	PV 54	PV 55	PV 56	PV 57	PV 58	PV 59	PV 60
220	146	72	183	111	103	254	163	131	136	151	232	93	68	144	148	197	172	172	138	246	109	147
240	161	72	225	109	162	120	290	151	231	124	166	98	105	116	144	128	-14	191	137	150	195	164
260	175	72	108	272	77	200	220	168	158	72	248	307	154	325	125	311	121	84	197	226	244	133
280	178	72	271	364	174	71	286	225	185	184	241	50	237	91	112	144	151	213	117	169	93	206
300	181	72	332	190	154	117	120	266	253	91	227	227	245	184	125	49	141	225	142	149	157	178
320	190	71	232	274	209	173	206	202	290	141	217	81	229	86	335	239	287	210	202	135	108	312
340	196	72	250	304	289	177	258	257	200	91	48	101	182	62	299	167	259	144	181	251	178	141
360	205	71	214	250	197	138	123	135	184	187	138	173	139	293	114	210	201	137	243	161	206	231
380	216	72	205	173	183	216	215	182	144	311	193	323	198	266	207	224	229	190	195	349	262	170
400	228	72	240	297	152	192	191	240	213	294	164	205	271	223	138	153	221	172	332	317	281	286
420	238	76	173	246	252	304	285	243	123	260	277	229	250	291	272	184	285	126	192	251	248	181
440	253	74	316	213	212	238	373	259	119	256	275	229	127	418	216	284	125	194	376	113	266	223
460	265	73	323	449	359	315	138	278	227	212	240	263	358	175	204	250	273	226	228	265	357	263
480	280	71	250	190	267	219	268	182	206	347	321	246	180	153	232	286	227	314	336	315	310	493
500	297	71	244	315	171	209	363	307	218	348	325	281	238	244	369	251	341	301	297	328	368	316
520	317	73	318	236	205	329	493	388	331	251	362	392	352	296	138	354	332	299	207	401	197	304
540	330	72	292	429	313	267	397	258	383	376	468	332	321	451	234	243	392	322	281	313	336	363
560	351	66	302	264	356	312	214	241	472	290	367	299	268	237	320	481	427	374	327	419	405	417
580	364	66	322	394	359	420	312	385	234	426	329	389	331	309	403	358	418	363	278	330	351	307
600	377	68	312	272	301	470	224	363	450	457	161	353	375	451	290	324	470	381	374	200	343	298
620	392	69	304	342	449	296	442	324	373	431	473	451	510	398	537	398	345	462	429	273	302	489
640	405	67	403	393	381	432	391	424	466	264	339	331	409	442	586	508	440	473	431	548	341	462
660	420	63	449	512	437	454	455	349	414	407	347	388	341	377	372	453	482	383	412	356	413	348
680	444	66	488	498	499	309	419	519	431	418	492	422	411	480	624	418	588	543	403	384	294	548
700	456	69	459	542	375	540	344	407	572	445	537	418	556	517	569	375	397	480	472	602	482	497
720	473	71	460	482	467	640	482	406	585	488	347	445	298	487	585	439	473	538	495	437	539	481
740	486	70	619	494	502	437	488	484	516	697	500	578	506	416	437	617	407	490	487	461	469	332
760	492	72	514	427	520	554	381	444	502	409	467	396	441	480	481	475	548	471	526	456	707	457
780	507	72	513	595	342	523	564	361	610	470	537	560	452	357	624	450	439	498	454	490	389	437
800	521	72	591	538	507	372	478	519	531	414	427	352	546	618	422	642	630	415	488	544	460	439

PASEC Reading	Projected Mean	Projected SD	PV 61	PV 62	PV 63	PV 64	PV 65	PV 66	PV 67	PV 68	PV 69	PV 70	PV 71	PV 72	PV 73	PV 74	PV 75	PV 76	PV 77	PV 78	PV 79	PV 80
220	146	72	94	239	214	163	300	-10	61	161	120	112	21	359	221	126	66	167	11	276	217	101
240	161	72	177	142	264	141	7	224	90	212	256	228	121	184	152	50	136	103	253	182	201	216
260	175	72	265	187	66	167	169	105	126	255	89	249	192	267	57	254	263	127	182	31	155	289
280	178	72	135	111	163	186	195	146	15	131	118	170	52	189	199	182	226	229	112	193	239	125
300	181	72	141	125	135	234	187	158	205	175	253	158	169	331	169	93	134	116	203	204	167	239
320	190	71	117	166	115	132	230	-19	225	306	266	171	285	256	221	194	131	238	152	161	267	250
340	196	72	256	223	78	25	201	208	271	251	148	208	227	181	249	235	196	221	219	190	216	173
360	205	71	241	172	298	189	299	61	247	199	292	254	23	259	300	343	96	171	207	191	184	151
380	216	72	249	252	249	190	180	348	86	260	340	172	175	241	189	248	297	350	308	213	268	303
400	228	72	223	204	232	199	126	121	254	233	333	317	118	178	207	211	186	247	247	334	230	268
420	238	76	207	208	280	302	186	205	293	198	346	240	136	293	322	227	260	238	240	386	170	157
440	253	74	207	392	345	335	241	163	271	138	199	275	226	304	293	188	226	203	204	332	71	256
460	265	73	242	223	316	277	356	445	121	248	169	264	237	232	193	229	282	379	147	317	372	299
480	280	71	381	309	240	456	318	337	438	281	342	254	198	250	333	198	254	190	332	308	219	263
500	297	71	264	346	272	358	374	240	287	428	400	202	260	299	273	307	343	236	208	348	245	229
520	317	73	325	425	340	441	223	400	370	123	340	371	360	254	301	520	305	215	348	318	429	308
540	330	72	456	350	310	361	380	367	277	396	336	167	381	311	311	283	343	422	357	336	362	275
560	351	66	283	381	305	420	377	414	357	258	419	410	314	240	396	360	289	451	392	396	400	530
580	364	66	365	384	346	313	419	419	360	311	554	237	337	350	395	351	448	436	369	303	285	332
600	377	68	453	370	386	346	398	410	402	321	491	449	495	284	296	402	361	357	384	450	428	455
620	392	69	368	387	382	393	403	279	497	328	356	414	399	317	327	404	397	393	336	402	431	286
640	405	67	513	432	371	454	361	499	335	459	431	421	487	400	393	529	307	372	414	375	345	475
660	420	63	483	469	456	364	527	383	487	429	436	431	396	350	405	304	423	419	513	338	509	498
680	444	66	446	471	445	336	352	409	416	436	510	456	477	438	481	495	487	385	502	472	551	494
700	456	69	379	577	504	406	331	434	413	371	409	507	531	465	363	583	493	483	410	519	412	436
720	473	71	564	424	488	360	516	425	441	447	485	447	508	328	555	455	468	396	496	558	528	367
740	486	70	452	336	647	432	619	449	408	465	478	307	330	516	507	470	579	539	518	492	425	504
760	492	72	557	441	433	465	457	520	476	517	413	502	582	482	583	509	415	490	501	542	371	519
780	507	72	541	534	369	501	505	436	494	429	527	534	483	400	429	509	602	629	516	542	488	456
800	521	72	592	540	484	480	609	598	529	624	437	484	442	516	483	534	670	491	466	531	585	439

PASEC Reading	Projected Mean	Projected SD	PV 81	PV 82	PV 83	PV 84	PV 85	PV 86	PV 87	PV 88	PV 89	PV 90	PV 91	PV 92	PV 93	PV 94	PV 95	PV 96	PV 97	PV 98	PV 99	PV 100
220	146	72	224	106	145	112	137	180	136	231	223	97	383	169	138	209	310	150	132	38	158	216
240	161	72	246	168	152	130	31	200	169	162	224	54	152	268	147	167	288	262	242	116	166	209
260	175	72	213	206	132	266	202	224	232	116	176	271	171	219	213	166	101	210	239	136	181	184
280	178	72	190	270	178	213	192	187	208	61	128	150	211	211	217	277	44	109	143	300	180	263
300	181	72	257	225	208	187	148	206	305	86	232	208	338	125	127	190	203	105	73	275	179	184
320	190	71	154	179	134	99	318	135	222	309	250	177	148	134	220	117	164	71	182	190	193	311
340	196	72	203	144	222	248	211	182	111	92	229	286	271	172	247	142	136	165	137	111	194	224
360	205	71	163	206	310	165	235	223	124	319	101	181	166	160	253	138	257	165	185	175	197	304
380	216	72	162	91	142	114	166	284	159	97	157	293	219	171	249	166	257	279	229	229	219	164
400	228	72	373	234	292	233	137	208	194	276	304	259	178	262	148	257	85	191	259	179	220	184
420	238	76	291	245	143	301	339	301	73	303	250	250	62	194	321	295	388	185	212	195	229	158
440	253	74	275	266	251	415	236	265	350	264	146	303	134	373	418	98	391	197	158	269	248	373
460	265	73	261	342	380	266	238	273	286	245	258	317	129	396	302	247	167	172	294	286	267	265
480	280	71	218	356	226	302	295	387	352	327	264	180	386	249	290	295	187	251	286	306	276	305
500	297	71	342	236	410	141	253	319	246	303	319	207	230	391	488	285	365	295	255	270	293	234
520	317	73	246	350	194	300	245	315	370	151	428	319	439	455	360	398	293	469	335	236	319	291
540	330	72	334	350	338	508	298	375	389	279	344	251	376	387	338	327	382	288	432	363	329	373
560	351	66	304	332	447	276	157	452	498	156	303	350	329	224	332	406	323	402	333	433	339	264
580	364	66	410	409	333	290	344	188	361	408	288	387	448	534	379	414	404	474	417	495	368	382
600	377	68	378	438	427	445	357	438	402	391	308	283	319	397	387	367	417	330	239	336	366	435
620	392	69	380	443	501	489	363	322	519	429	504	352	305	326	485	383	387	256	434	410	398	432
640	405	67	468	535	315	425	325	338	495	383	425	410	296	315	425	397	376	483	500	476	406	391
660	420	63	357	518	369	437	537	350	451	394	360	347	466	346	537	443	459	455	512	451	417	498
680	444	66	481	409	444	508	395	581	478	365	511	444	389	384	355	463	296	525	398	295	444	422
700	456	69	352	495	529	451	498	468	464	491	394	367	479	344	524	454	400	522	482	461	452	452
720	473	71	455	430	381	486	451	484	437	530	446	547	385	374	595	369	482	587	372	301	471	467
740	486	70	548	412	464	447	465	571	479	493	413	567	423	280	503	417	466	488	439	527	468	420
760	492	72	447	330	540	533	432	485	479	493	498	593	462	500	433	470	455	551	601	441	486	557
780	507	72	523	542	535	373	456	575	381	514	496	460	616	515	571	567	383	546	475	412	497	564
800	521	72	482	434	556	388	539	520	495	492	565	512	342	501	594	500	466	491	628	432	505	508

APPENDIX C

Using the Rosetta Stone Concordance Tables – Analysis Steps

Using the Rosetta Stone concordance tables for projections of regional assessments is possible, but relies on a number of assumptions that cannot be tested unless a Rosetta Stone study is conducted for the country that utilizes the concordance. The estimation of percentages of students reaching TIMSS and PIRLS International Benchmarks described here must therefore be considered as extrapolation. The mechanics of generating such an extrapolation are:

Analysis Steps

1. Calculate the average PV based on the PASEC sample for each student in the domain of interest, either PASEC mathematics or PASEC reading, to obtain an approximate posterior mean on the PASEC scale for each student.
2. Find the closest PASEC level in the concordance table for each student (source test); the corresponding projected mean and standard deviation (SD) on the TIMSS or PIRLS scale for the closest PASEC level should be assigned to each student.

   Example: For a student with an average PV of 505 based on 5 PASEC mathematics PVs, the closest PASEC mathematics level is 500; the assigned projected mean and SD on the TIMSS scale are 336 and 63, respectively (see the concordance table for PASEC mathematics in Exhibit 11.6). For a student with an average PV of 505 based on 5 PASEC reading PVs, the closest PASEC reading level is 500; the assigned projected mean and SD on the PIRLS scale are 297 and 71, respectively (see the concordance table for PASEC reading in Exhibit 11.7).
3. Impute 5 new projected TIMSS mathematics or PIRLS reading PVs (target test) based on the assigned projected mean and SD (for mathematics or reading) of the conditional distribution for each student. PVs for individual students can be imputed using a normal distribution with the corresponding projected mean and SD in SAS, R Packages, EXCEL, and other software tools (this step is repeated five times to get 5 PVs).
4. These 5 sets of projected PVs can then be used to estimate the percentages of PASEC students reaching the TIMSS or PIRLS international benchmarks, Advanced (625), High (550), Intermediate (475), Low (400). The final percentage of reaching a benchmark, t_0, is the average of the 5 percentages, t_m, calculated based on 5 set of projected PVs:

$$t_0 = \frac{1}{5}\sum_{m=1}^{5} t_m \tag{C.1}$$

5. The standard error needs to be calculated using proper weights and formulas to include imputation variance and sampling variance. The imputation variance is simply the variance of the 5 percentages of reaching the benchmark (from step 4) then multiplied by a factor $\frac{6}{5}$:

$$Var_{imp}(t_0) = \frac{6}{5}\sum_{m=1}^{5}\frac{(t_m - t_0)^2}{4} \tag{C.2}$$

For each set of PVs, the sampling variance is the variance among the different percentages calculated by using each set of replicate sampling weights (which are usually included in the data file); n is the number of replicate weights:

$$Var_{smp}(t_m) = \sum_{h=1}^{n}(t_{mh} - t_m)^2 \tag{C.3}$$

The final sampling variance is the average of the sampling variance from the 5 set of projected PVs:

$$Var_{smp}(t_0) = \frac{1}{5}\sum_{m=1}^{5} Var_{smp}(t_m) \tag{C.4}$$

The square root of the sum of imputation variance and sampling variance is the standard error of the percentages of reaching international benchmarks:

$$SE = \sqrt{Var_{imp}(t_0) + Var_{smp}(t_0)} \tag{C.5}$$

6. Do all the steps for each domain of interest (mathematics or reading) separately using the (mathematics or reading) concordance table
7. The estimated percentages and standard errors can be reported noting that the projection for each new country relies on the concordance based on samples from only 3 other countries, not including the present country. Therefore, there are sources of error variance and bias that are not reflected in the projections.

UNESCO
unesco
Institute for Statistics

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

Analysis Report:
Establishing a Concordance between PASEC and TIMSS/PIRLS