unesco
Institute for Statistics

Laboratorio Latinoamericano Evaluación Calidad Educación

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

Rosetta Stone Analysis Report:
Establishing a Concordance between ERCE and TIMSS/PIRLS

4 QUALITY EDUCATION

IEA ROSETTA STONE
BOSTON COLLEGE

UNESCO

The constitution of the United Nations Educational, Scientific and Cultural Organization (UNESCO) was adopted by 20 countries at the London Conference in November 1945 and entered into effect on 4 November 1946. The 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/

ROSETTA STONE ANALYSIS REPORT:

Establishing a Concordance between ERCE 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 ERCE and the international TIMSS and PIRLS achievement scales in Latin American and Caribbean countries.

The Rosetta Stone study consists of two assessment parts. The first part is the ERCE assessment including the ERCE context questionnaire. The second part is the Rosetta Stone assessment comprising test booklets with item blocks and passages from TIMSS and PIRLS. Both assessment parts were administered in two ERCE 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 that account for the uncertainty of the measurement (measurement error).

The key findings can be summarized as follows:

- The overall difficulty of the selected TIMSS and PIRLS item blocks and passages (developed for 4th-grade students) are appropriate for the Rosetta Stone analysis and the goals of the study.
- Comparable ERCE 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 ERCE mathematics and TIMSS (r = .89-.90) and ERCE reading and PIRLS (r = .82-.86) suggest that constructs (while developed for different target grades and based on different assessment frameworks) are similar enough to enable a concordance.
- Population models were able to be estimated providing proficiency distributions for ERCE and Rosetta Stone scales.
- Plausible values (PVs) for ERCE 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 ERCE team (r = .94-.97 across countries for both ERCE mathematics and reading) indicating very good agreement of analytic processes.

- Population models were applied to the Rosetta Stone data to obtain posterior means and PVs for TIMSS mathematics/numeracy and PIRLS reading/literacy.
- Estimates from both assessments, ERCE and Rosetta Stone, were used to establish concordance tables that provide a conditional distribution on the TIMSS and PIRLS scales for a range of ERCE 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 ERCE 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 UN Sustainable Development Goal 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[1]. 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 UNESCO's Regional Comparative and Explanatory Study (ERCE) in Latin American and Caribbean countries and TIMSS and PIRLS. ERCE assesses student achievement in reading, writing, and mathematics at grade 3 as well as reading, writing, mathematics, and sciences at grade 6. It is conducted by the UNESCO's Latin American Laboratory for the Evaluation of the Quality of Education (LLECE), which is managed by the Regional Bureau for Education for Latin America and the Caribbean (OREALC/ UNESCO Santiago).

1 Please refer to the following document: http://unesdoc.unesco.org/images/0026/002606/260607E.pdf

In contrast to the ERCE assessment which is conducted at grades 3 and 6, PIRLS is an assessment conducted at grade 4 and TIMSS is an assessment conducted at grades 4 and 8. This is, the assessments are based on different frameworks targeting different populations as defined by their intended focal grade levels.

To construct the concordance, the 2019 ERCE assessment was administered to students in the sixth grade together with the Rosetta Stone linking booklets that contained items from the fourth grade TIMSS and PIRLS assessments. The content of ERCE's mathematics assessment was expected to align well with the TIMSS fourth grade assessments in numeracy and mathematics. Similarly, the content of ERCE'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 ERCE 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 ERCE is a collaborative project between the UNESCO Institute for Statistics (UIS), the ERCE study center (LLECE), IEA, and the TIMSS & PIRLS International Study Center at Boston College, as well as the national teams of the participating countries Colombia and Guatemala. Questions about linking design, the data analyses, and the report for the Rosetta Stone study for ERCE 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 anyone 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. The Rosetta Stone study comprises two assessment parts. The first part is the ERCE assessment including the ERCE achievement items and ERCE context questionnaires. The second part is the centerpiece of the study, the Rosetta Stone assessment part consisting of test booklets with TIMSS mathematics item blocks and PIRLS

reading passages. More precisely, items come from TIMSS 2015, TIMSS Numeracy 2015, PIRLS 2016, and PIRLS Literacy 2016. In total, eight mathematics/numeracy item blocks and four reading/literacy passages were selected (note that TIMSS Numeracy and PIRLS Literacy items are relatively easier than regular TIMSS and PIRLS items). Exhibit 3.1 provides the number of items and source for each item block. Both assessment parts were administered as paper-based assessments to the same students during the ERCE 2019 administration. Each student was administered (one or) two ERCE booklets on the first day and one Rosetta Stone booklet on the second day. The ERCE reading and mathematics domains were administered in separate booklets with most students taking both (and only a few students taking one). A description of the ERCE 2019 assessment can be found in the ERCE 2019 executive report (ERCE, 2021).

Exhibit 3.1: Rosetta Stone Linking Item Blocks and Passages

	Source	Number of Items
TIMSS Blocks		
N01	TIMSS Numeracy 2015 – N01	13
N02	TIMSS 2015 – M01	14
N03	TIMSS Numeracy 2015 – N07	13
N04	TIMSS 2015 – M02	11
N05	TIMSS Numeracy 2015 – N09	13
N06	TIMSS 2015 – M05	14
N07	TIMSS Numeracy 2015 – N10	13
N08	TIMSS 2015 – M08	11
	Total Mathematics Items	**102**
PIRLS Passages		
L01	PIRLS 2016 – Flowers on the Roof (Literary)	13
L02	PIRLS Literacy 2016 – African Rhinos (Informational)	17
L03	PIRLS Literacy 2016 – The Pearl (Literary)	15
L04	PIRLS 2016 – Green Sea Turtle (Informational)	16
	Total Reading Items	**61**

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 mathematics/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 ERCE and Rosetta Stone scales across student populations (sections 6 and 7). Third, population models were used to impute plausible values (PVs) separately for ERCE 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 two ERCE countries using sample weights provided to the TIMSS & PIRLS International Study Center. During data processing, questions arose concerning cases that had no responses to ERCE reading or math items but had received sample weights. These issues could be clarified in close collaboration with the ERCE team and updated data files were provided to the TIMSS & PIRLS International Study Center.

Exhibits 4.1a and 4.1b provide the sample sizes for each country and assessment available for the different analysis steps. Cases with sample weights and responses to achievement items (ERCE items, Rosetta Stone items, or both) were included in the analysis while cases with responses only to the ERCE context questionnaire items were excluded. The originally targeted samples sizes for the Rosetta Stone study were reduced as 1,357 students in Colombia and 176 students in Guatemala did not receive Rosetta Stone booklets; these cases could only be included in the IRT re-scaling of the ERCE items. For the population models, only students who participated in both the Rosetta Stone linking assessment and the ERCE assessment were included in the analysis (3,108 students in Colombia and 4,716 students in Guatemala); four more cases were available for the IRT scaling of the Rosetta Stone linking items. No impact of the reduced sample size on results from the population model for Colombia was found (see detailed results in section 9). For the main analysis of the report aimed at constructing the

concordance tables (i.e., the concordance donor selection; see more details in section 11), only students who participated in all four tests, ERCE math, ERCE reading, TIMSS linking, and PIRLS linking were included, illustrated in the last column of Exhibit 4.1b (2,619 students in Colombia and 3,902 students in Guatemala).

Exhibit 4.1a: Sample Sizes per Country and Assessment

Country	Number of Students in			Total Number of Students	Number of Schools/ Classes*
	ERCE only	Rosetta Stone only	ERCE & Rosetta Stone		
Colombia	1,357	2	3,108	4,467	145
Guatemala	176	2	4,716	4,894	234
Total	**1,533**	**4**	**7,824**	**9,361**	**379**

* Note: One class per school was tested

Exhibit 4.1b: Rosetta Stone Sample Sizes for each Analysis Step

Country	Number of Students in		
	IRT Scaling of Rosetta Stone Linking Items	Population Modeling	Concordance Donor Selection
Colombia	3,110	3,108	2,619
Guatemala	4,718	4,716	3,902
Total	**7,828**	**7,824**	**6,521**

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 ERCE items were already calibrated by the ERCE team, which also provided the PVs for ERCE, the TIMSS & PIRLS International Study Center performed IRT scaling and population modeling for the Rosetta Stone linking items. For validation and replication purposes, the ERCE items were re-calibrated as well. The following IRT models were estimated:

1. *Comparability of ERCE items across countries:* For evaluating the psychometric properties and cross-country comparability of the ERCE 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 ERCE PVs that were received from the ERCE 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 ERCE and Rosetta Stone constructs:* Through multidimensional IRT models, latent correlations between ERCE and Rosetta Stone scales were estimated to evaluate whether the ERCE 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 ERCE background questionnaire for imputing PVs. The population modeling was performed at the country-level and separately for ERCE and Rosetta Stone linking data. After the comparability and accuracy of the population modeling approaches used in ERCE and in the Rosetta Stone study was confirmed (by re-estimating the ERCE PVs), the posterior means and PVs from the population models were utilized 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 ERCE data.

5. Data Quality Evaluation

Data quality was evaluated using classical item statistics (percent correct and item-total or point-biserial 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 ERCE item block by country. The percent of correct responses show that on average the difficulty of the TIMSS item blocks and PIRLS passages are comparable to the ERCE mathematics and reading item blocks for the ERCE population, with the tendency of the TIMSS and PIRLS items to be slightly easier. This may be due to the fact that TIMSS and PIRLS target fourth grade students while the ERCE items are targeting sixth graders. The point-biserial correlations indicate that item blocks and passages exhibit medium discriminations across the different assessments, with the tendency of TIMSS and PIRLS item blocks and passages showing slightly higher discrimination than ERCE item blocks.

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

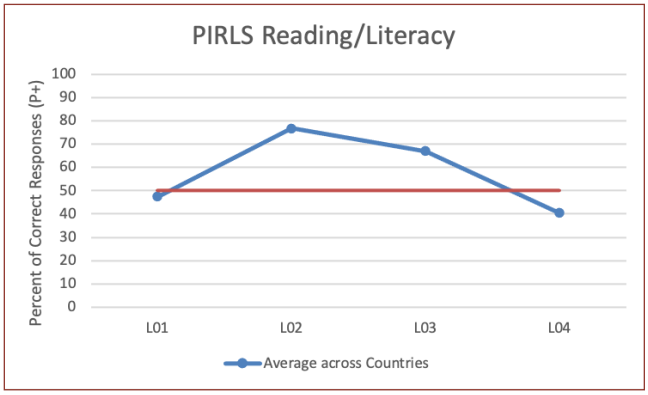
	Colombia		Guatemala	
Item Block/Passage	**Average Percent Correct**	**Average Point-Biserial**	**Average Percent Correct**	**Average Point-Biserial**
Rosetta Stone PIRLS Reading/Literacy				
L01	57.9	0.46	36.8	0.49
L02	82.7	0.44	70.8	0.50
L03	75.4	0.46	58.6	0.50
L04	48.7	0.50	32.0	0.46
Average	**66.2**	**0.47**	**49.6**	**0.49**
ERCE Reading				
L1	56.2	0.36	46.8	0.34
L2	62.3	0.44	49.1	0.41
L3	61.0	0.43	49.6	0.40
L4	61.7	0.41	50.3	0.39
L5	61.6	0.37	50.7	0.35
L6	63.6	0.39	49.6	0.36
Average	**61.1**	**0.40**	**49.4**	**0.38**

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

	Colombia		Guatemala	
Item Block	Average Percent Correct	Average Point-Biserial	Average Percent Correct	Average Point-Biserial
Rosetta Stone TIMSS Mathematics/ Numeracy				
N01	59.1	0.43	54.7	0.44
N02	50.4	0.41	44.8	0.39
N03	57.7	0.40	54.1	0.40
N04	39.5	0.42	28.9	0.41
N05	56.3	0.39	49.7	0.38
N06	42.6	0.35	35.5	0.35
N07	64.3	0.41	54.5	0.44
N08	35.5	0.42	31.8	0.37
Average	**50.7**	**0.40**	**44.2**	**0.40**
ERCE Mathematics				
B1	42.0	0.32	35.1	0.27
B2	43.7	0.32	34.9	0.29
B3	42.9	0.33	37.0	0.28
B4	44.9	0.29	39.0	0.25
B5	46.8	0.36	40.0	0.37
B6	45.6	0.36	38.0	0.35
B7	47.0	0.33	38.8	0.28
Average	**44.7**	**0.33**	**37.5**	**0.30**

Exhibits 5.3 and 5.4 illustrate the average item difficulty (P+) by item block averaged across countries for PIRLS reading/literacy and ERCE reading and for TIMSS mathematics/numeracy and ERCE 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 Exhibit 5.1 show that TIMSS and PIRLS item blocks and passages tend to be either more or less difficult than ERCE item blocks within and across countries. Overall, the difficulty of the TIMSS linking item blocks and the difficulty of two of the PIRLS passages is at an appropriate level for the Rosetta Stone analyses. Two of the PIRLS linking passages (L02 and L03) appear relatively easy for the assessed ERCE student population. This is likely due to the fact that TIMSS and PIRLS linking items were developed for students at grade 4 but administered to ERCE students at grade 6.

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



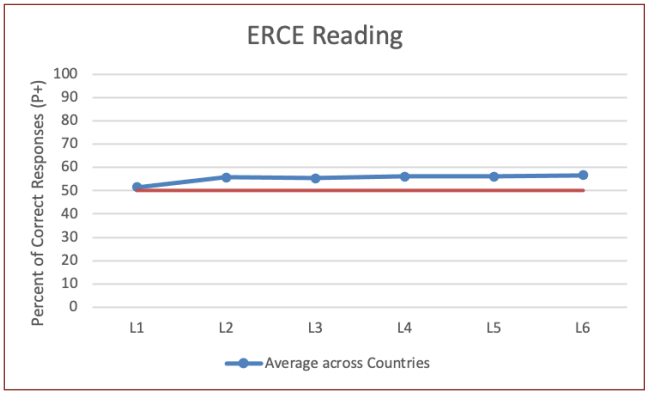
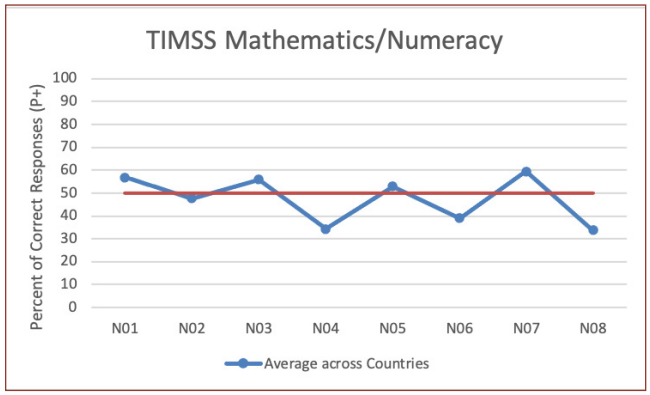
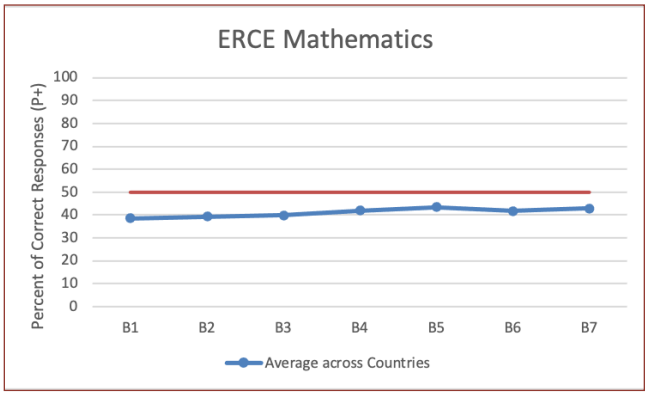


Exhibit 5.4: Average Item difficulty (percent correct) by Item Block for TIMSS Mathematics/ Numeracy and ERCE Mathematics





Exhibits 5.5 and 5.6 illustrate the average percent of omitted (OM) and not reached (NR) items for each ERCE and Rosetta Stone item block. The NR and OM rates are small enough and consistent enough across countries and item blocks/passages to not be of any concern.

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

	Colombia		Guatemala	
Item Block/Passage	**Average Percent Omitted**	**Average Percent Not Reached**	**Average Percent Omitted**	**Average Percent Not Reached**
Rosetta Stone (PIRLS) Reading/Literacy				
L01	3.2	0.3	4.0	0.8
L02	1.7	0.2	2.6	0.2
L03	2.2	0.7	2.4	0.4
L04	5.7	1.2	6.8	1.1
Average	**3.2**	**0.6**	**4.0**	**0.6**
ERCE Reading				
L1	0.6	0.7	2.0	4.6
L2	0.6	0.5	2.1	3.9
L3	0.6	0.8	2.2	4.8
L4	0.5	0.6	1.6	4.2
L5	0.6	0.4	2.2	4.0
L6	0.8	1.1	2.2	5.6
Average	**0.6**	**0.7**	**2.1**	**4.5**

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

	Colombia		Guatemala	
Item Block	Average Percent Omitted	Average Percent Not Reached	Average Percent Omitted	Average Percent Not Reached
Rosetta Stone (TIMSS) Mathematics/ Numeracy				
N01	3.5	0.6	6.0	1.0
N02	6.6	0.4	7.7	0.8
N03	3.6	0.6	4.2	0.6
N04	3.4	0.3	2.9	0.5
N05	3.3	0.5	3.5	1.2
N06	2.9	0.4	6.1	0.6
N07	1.9	0.4	2.0	0.6
N08	7.3	0.3	8.5	1.0
Average	**4.1**	**0.4**	**5.1**	**0.8**
ERCE Mathematics				
B1	1.8	0.2	5.1	1.2
B2	2.0	0.2	6.4	1.1
B3	1.4	0.3	4.0	0.9
B4	1.1	0.3	3.5	0.8
B5	3.4	0.2	5.9	1.1
B6	3.0	0.4	5.3	0.9
B7	2.5	0.3	4.9	1.0
Average	**2.2**	**0.3**	**5.0**	**1.0**

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 ERCE 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 $i = 1, ..., I$. The set of responses to these items is $\boldsymbol{x_v} = (x_{v1}, ..., x_{vi})$ 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:

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

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 ERCE 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:

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

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

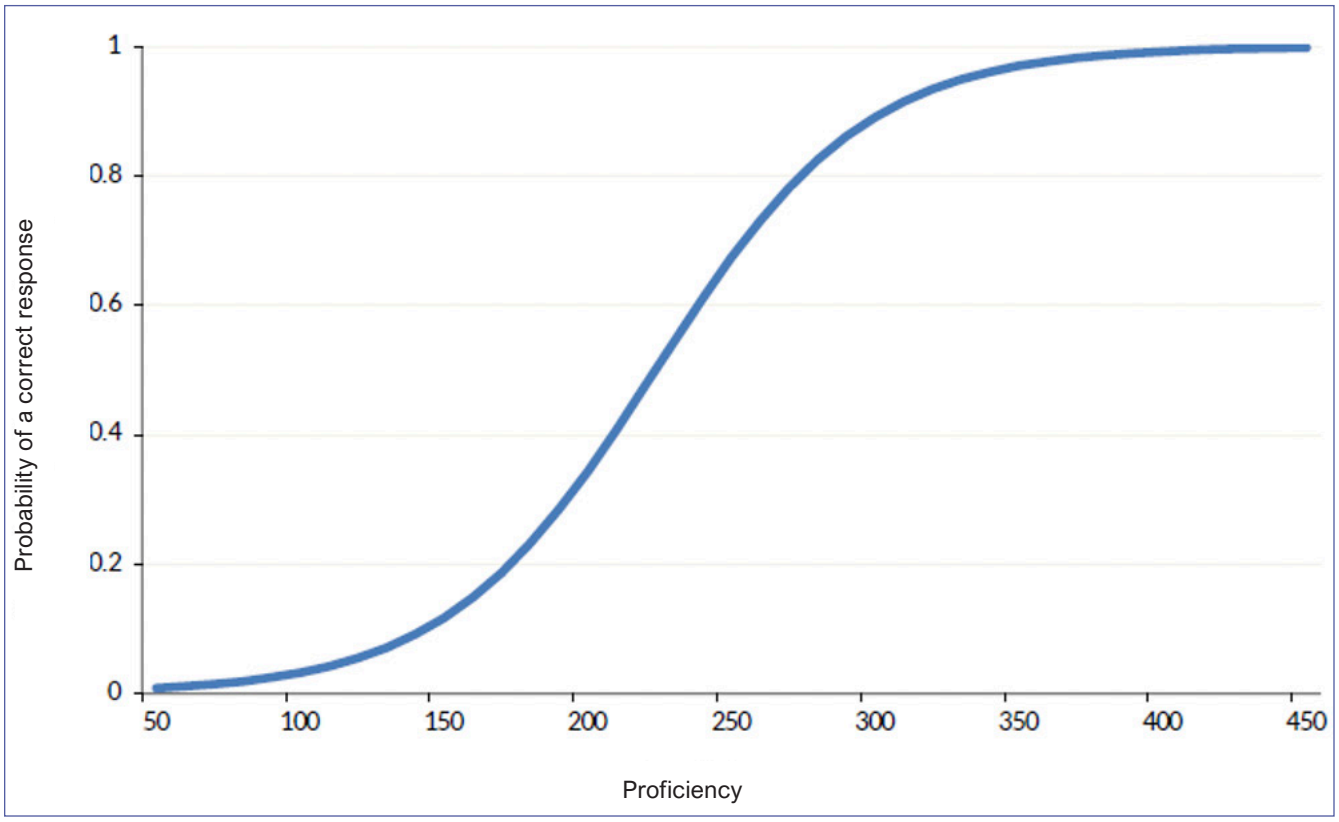
$$P(\mathrm{X} = x_1, \ldots, x_I \mid \theta) = \prod_{i=1}^{I} P_i\,(X = 1 \mid \theta)^{x_i}\,[1 - P_i\,(X = 1 \mid \theta)^{1-x_i} \tag{6.8}$$

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:

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

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 ERCE and Rosetta Stone Data

This section describes the application of IRT scaling to Rosetta Stone linking items in particular as well as the ERCE 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 ERCE 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 two countries. More precisely, the item parameters used came from the TIMSS 2015 and TIMSS Numeracy 2015 IRT calibration and the PIRLS 2016 and PIRLS Literacy 2016 IRT calibration (TIMSS Numeracy 2015 and PIRLS Literacy 2016 were linked to TIMSS 2015 and PIRLS 2016 respectively) and were estimated based on the 2PL, 3PL and GPCM (Martin, von Davier & Mullis, 2020). The comparability of the ERCE items across countries was evaluated by estimating common item parameters across countries based on the

Rasch model, in accordance with ERCE analysis procedures. 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 102 TIMSS mathematics/numeracy items.
- Model 2 (M2) was estimated for the 61 PIRLS reading/literacy items.
- Model 3 (M3) was estimated for the 112 ERCE mathematics items.
- Model 4 (M4) was estimated for the 95 ERCE reading items.

While M1 and M2 utilize 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 ERCE countries. In such cases, items were excluded from the scaling.

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

- Model 5 (M5) was estimated with the ERCE mathematics items assigned to one factor/scale and TIMSS items assigned to a second factor/scale.
- Model 6 (M6) was estimated with the ERCE 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 ERCE 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 ERCE 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 mathematics/numeracy and PIRLS reading/literacy item parameters showed a good fit for the majority of item-by-country pairs (91.2% and 95.1% respectively). For a small subset of items, new common item parameters needed to be estimated (7.8% and 1.6% 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 a very small number of cases of item-by-country pairs, items needed to be excluded from the analysis (1.0% and 3.3% for numeracy and literacy, respectively); items were either excluded for all or 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 ERCE mathematics and ERCE reading items, a good fit to the common item parameter estimates was achieved (98.7% and 95.8% respectively). In a very small number of cases of item-by-country pairs, items needed to be excluded from the analysis (1.3% and 4.2% 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	91.2%	95.1%
New Common	7.8%	1.6%
Excluded	1.0%	3.3%

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

Item Parameters	ERCE Math (Model 3)	ERCE Reading (Model 4)
Common	98.7%	95.8%
Excluded	1.3%	4.2%

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 ERCE 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 ERCE 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

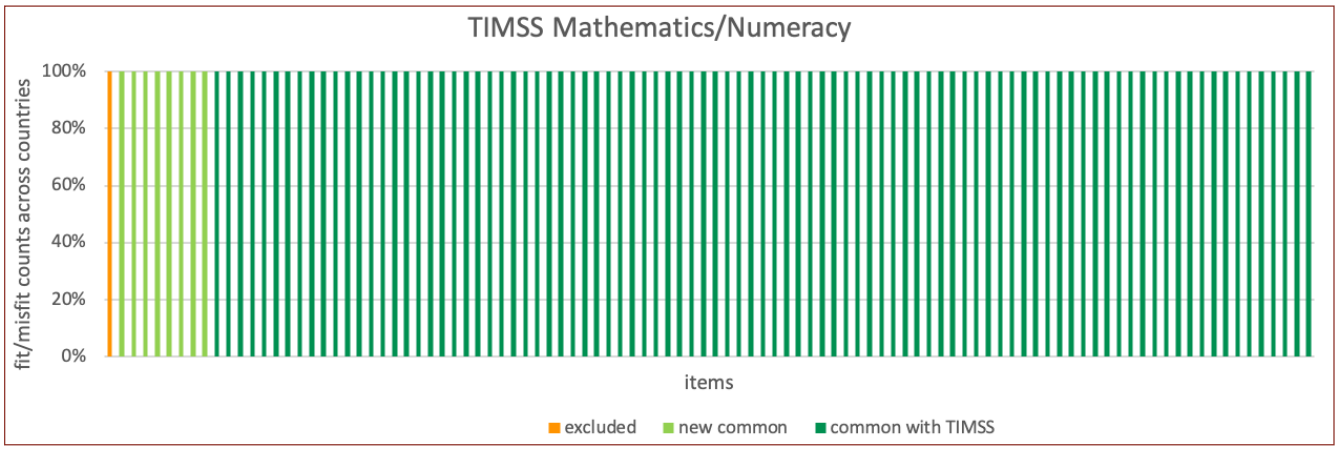


Exhibit 7.4: Distribution of Model 2 Items with Common Item Parameters versus Excluded Items

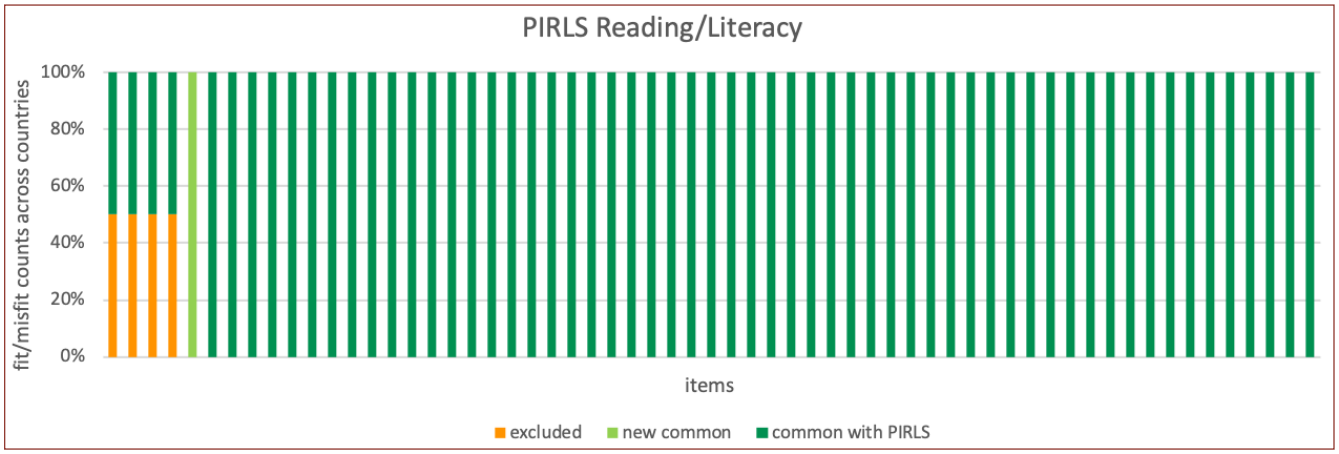


Exhibit 7.5: Distribution of Model 3 Items with Common Item Parameters versus Excluded Items

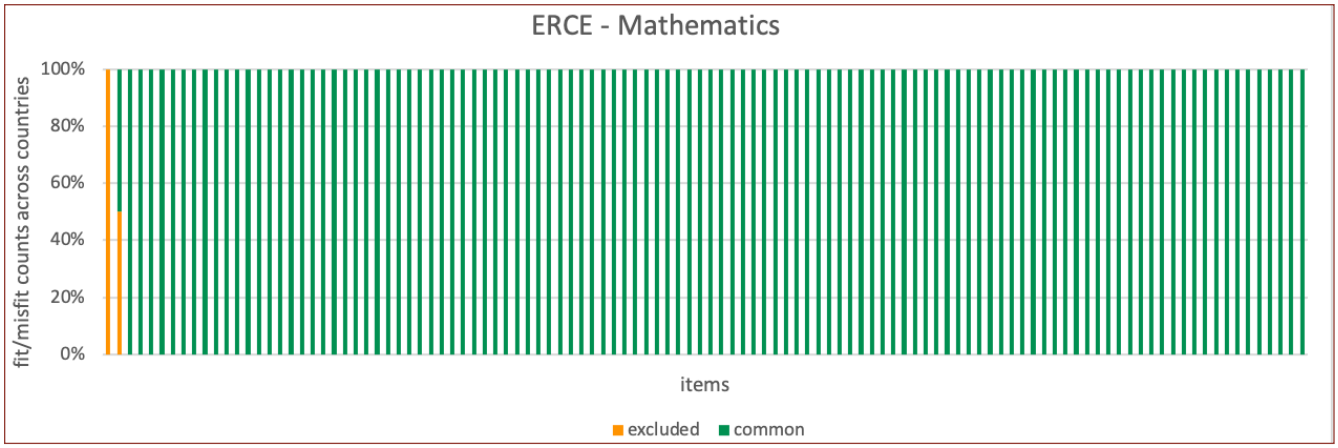
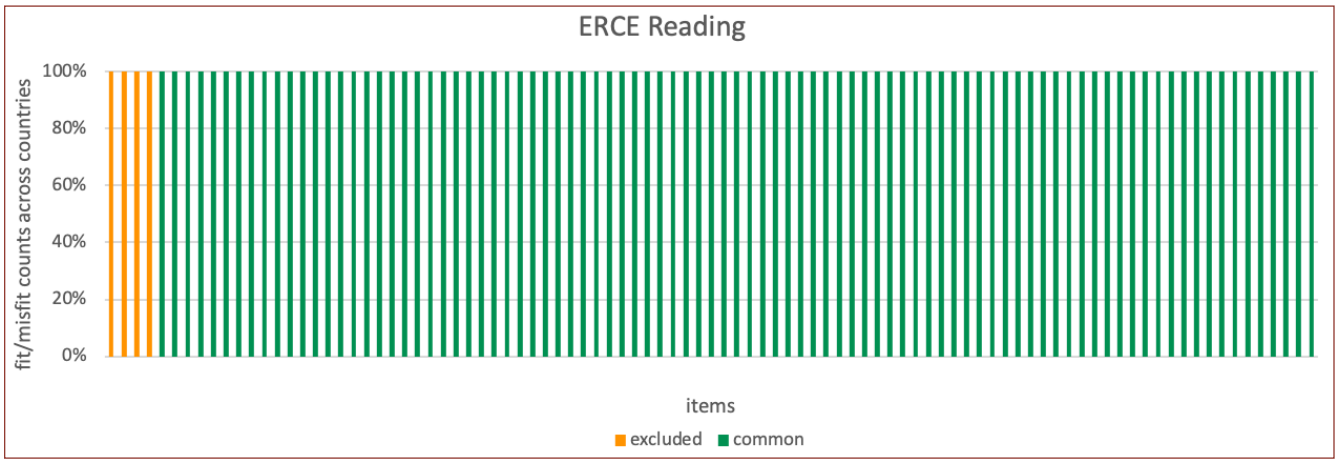


Exhibit 7.6: Distribution of Model 4 Items with Common Item Parameters versus Excluded Items



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 ranging from .82 to .90 across countries and models, see Exhibit 7.7. This indicates that the corresponding Rosetta Stone and ERCE scales measure constructs that are highly correlated and thus enable a meaningful concordance for the projection of score distributions. Nevertheless, constructs are not identical as they were developed based on different frameworks and by different assessment development teams, and for different target grades (grade 4 versus grade 6).

Exhibit 7.7: Latent Correlations between ERCE and Rosetta Stone Scales

Country	ERCE Mathematics with TIMSS (M5)	ERCE Reading with PIRLS (M6)
Colombia	0.89	0.82
Guatemala	0.90	0.86

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 utilize 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

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

This expectation utilizes 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

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

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

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

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

$$E_g(\theta \mid x_n, z_n) = \int_\theta \theta P_g(\theta \mid x_n, z_n) = d\theta \tag{8.5}$$

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

$$V_g(\theta \mid x_n, z_n) = E_g(\theta^2 \mid x_n, z_n) - [E_g(\theta \mid x_n, z_n)]^2 \tag{8.6}$$

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 ERCE, 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, ERCE, and other large-scale assessments are random draws from a conditional normal distribution

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

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 ERCE and Rosetta Stone Data

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

9.1 Applied Population Models

The population model, as described above, is a multivariate model that incorporates the available context variables from the ERCE student and home questionnaires, as well as the Rosetta Stone linking item parameters and the ERCE item parameters from the IRT scaling, respectively. Only data from students who responded to both the ERCE and the Rosetta Stone assessments were included in the population models.

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

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

Population Model 1 follows the practice established by TIMSS and PIRLS of using principal components analysis for reducing collinearity and dimensionality of predictors. In TIMSS and PIRLS, those principal components accounting for 90 percent of the variance of all context variables are retained for use as conditioning variables. In addition, the number of principal components retained is also limited to no more than 5% of a country's student sample size to avoid over-specification of the conditioning model. In the Rosetta Stone study for ERCE, both rules were used. That is, principal components were selected that either retained 90% of common variance or 5% of the sample size, whichever lead to fewer principal components. The population model was calculated separately for each of the two 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. As shown in the exhibit, the percentage of variance accounted for is much lower than 90%, especially for Colombia where only 3,108 students responded to both the ERCE and Rosetta Stone assessment and were included in the population model (compared to 4,716 students in Guatemala). The smaller percentage of explained variance in the population model implies that borrowing strength from the context variables for proficiency estimation is affected to a certain degree.

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
Colombia	2	798	155	58
Guatemala	2	769	235	72

The same analysis steps and conditioning variables were used for Population Model 2. Note that Model 2 was only estimated for evaluation purposes and is, therefore, not described in detail here. The ERCE PVs that were provided by the ERCE 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 ERCE 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, 2016, 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 mathematics/numeracy and the PIRLS reading/literacy domains (Population Model 1) as well as for the ERCE mathematics and reading domains (Population Model 2).

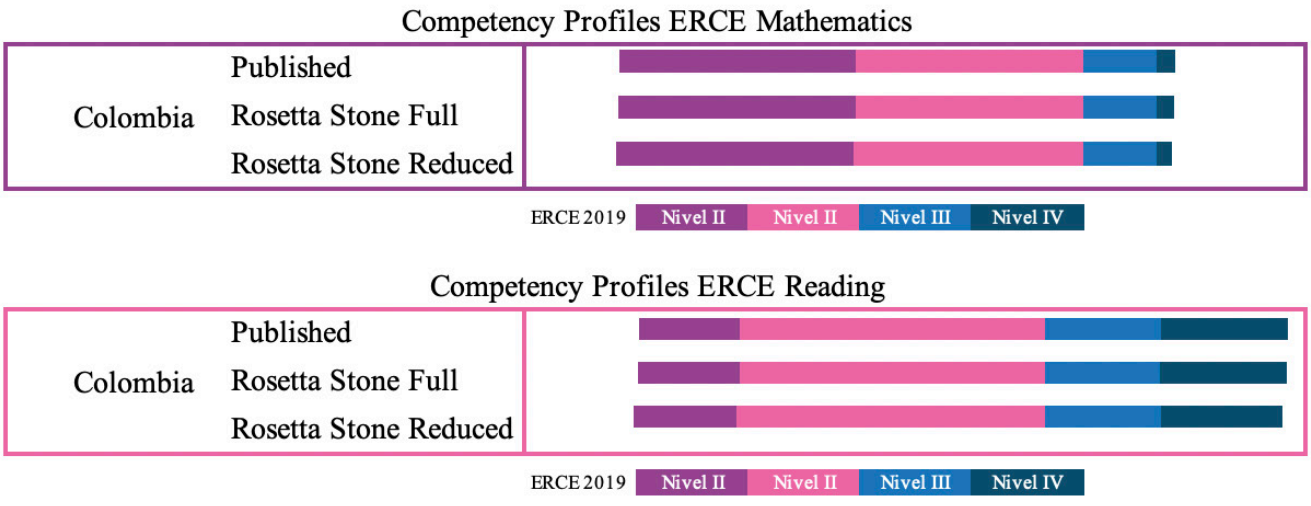
The ERCE PVs received from the ERCE team were evaluated by comparing them to the re-estimated PVs from Population Model 2. Very high correlations between both sets of PVs could be observed (ranging from .94 to .97 across countries for both mathematics and reading) indicating very good agreement of analytic processes.

To examine whether the reduced sample size in Colombia had any impact on the results of the population modeling, mean PVs based on the full sample size (N=4,467), including students with responses to ERCE items only, were compared to mean PVs based on the reduced sample size (N=3,108), including students who responded to both ERCE and Rosetta Stone items, for both the ERCE mathematics and reading scale. Findings are presented in Exhibits 9.2 and 9.3 and show very similar results based on both sample sizes.

Exhibit 9.2: Comparison of Published ERCE 2019 Results with Results Based on Full and Reduced Rosetta Stone Study Samples for Colombia

Statistics		Colombia		
		Published	Rosetta Stone Full (N=4467)	Rosetta Stone Reduced (N=3108)
Mathematics	Mean	707	707 (3.9)	706 (4.0)
	Standard Deviation		87 (2.7)	86 (2.6)
	Level 0	42.5	42.7 (2.0)	42.7 (2.1)
	Level 1	40.9	40.8 (1.4)	41.2 (1.5)
	Level 2	13.3	13.2 (1.1)	13.2 (1.3)
	Level 3	3.3	3.3 (0.8)	2.9 (0.7)
Reading	Mean	719	719 (4.6)	718 (4.9)
	Standard Deviation		103 (2.4)	102 (2.9)
	Level 0	15.5	15.6 (1.5)	15.7 (1.6)
	Level 1	47.0	47.0 (1.4)	47.6 (1.5)
	Level 2	17.8	17.7 (0.9)	17.9 (1.0)
	Level 3	19.7	19.6 (1.2)	18.8 (1.4)

Exhibit 9.3: Graphical Comparison of Published ERCE 2019 Results with Results Based on Full and Reduced Rosetta Stone Study Samples for Colombia



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 2015 (Foy & Yin, 2016) and PIRLS 2016 (Foy and Yin, 2017). 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	507.005	103.102	516.968	96.598
PV2	506.966	103.632	516.163	97.544
PV3	507.286	102.323	515.765	97.534
PV4	506.763	103.135	515.905	97.571
PV5	506.562	103.201	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 ERCE mathematics and TIMSS mathematics/numeracy as well as for ERCE reading and PIRLS reading/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 ERCE mathematics and reading scores respectively. That is, ERCE mathematics and ERCE 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

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

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

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

finding a small number of 'donor' observations by selecting these. $m_{1v} = m_1(\hat{\boldsymbol{\theta}}_{\boldsymbol{v}}),...,\mathrm{m}_{Lv} = \mathrm{m}_L(\hat{\boldsymbol{\theta}}_{\boldsymbol{v}})$ 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

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

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 ERCE, 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

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

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

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

   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.

4. 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 $m = 1,..., M$ 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 $|\theta_m - \theta_r|\ for\ m \in \{d_{1r},\ldots, d_{Lr}\} < |\theta_k - \theta_r|\ for\ k \in \{1,\ldots, N\}\backslash\{m_{1r},\ldots, m_{Lr}\}$.
   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 utilizes 5 nearest neighbors per country-specific sample.

11. Establishing an Enhanced Concordance between ERCE 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 ERCE source assessment on the scales of the TIMSS and PIRLS target assessments.

11.1 Relationship between ERCE data and Rosetta Stone Linking dat

To check the relationship between the data from source and target assessments, the correlations between the posterior means of ERCE data and Rosetta Stone linking data for mathematics/numeracy and reading/literacy were examined. For the ERCE 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 ERCE mathematics scale and the five PVs from the ERCE reading scale, respectively. The correlations between the posterior means of ERCE data and Rosetta Stone linking data are presented in the table in Exhibit 11.1.

Exhibit 11.1: Correlations between the Posterior Means of ERCE Data and Rosetta Stone (TIMSS and PIRLS) Linking Data

Country	ERCE Mathematics with TIMSS	ERCE Reading with PIRLS
Colombia	0.80	0.78
Guatemala	0.82	0.82

Correlations in Exhibit 11.1 approach the latent correlations from the multidimensional IRT models illustrated in section 7.3 and indicate that ERCE 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 ERCE 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.

Overall, Colombia shows a relatively higher performance than Guatemala on Rosetta Stone and ERCE assessments for both mathematics/numeracy and reading/literacy, which is consistent with the finding in Exhibit 5.1. The score ranges are similar across countries with Guatemala showing relatively lower scores at both ends on the ERCE mathematics, ERCE reading, and PIRLS Linking scales but not on the TIMSS linking scale.

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

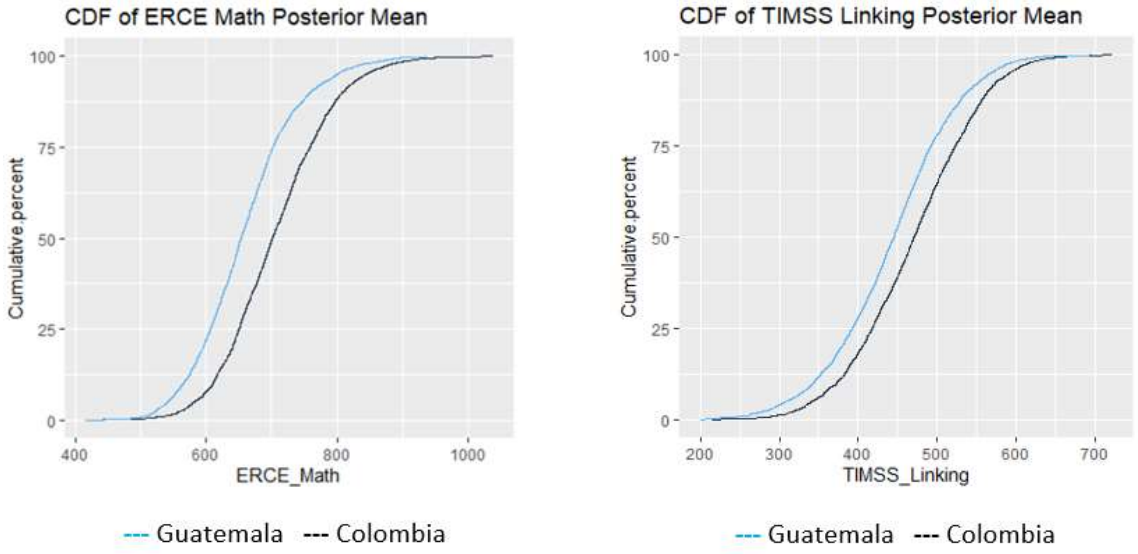
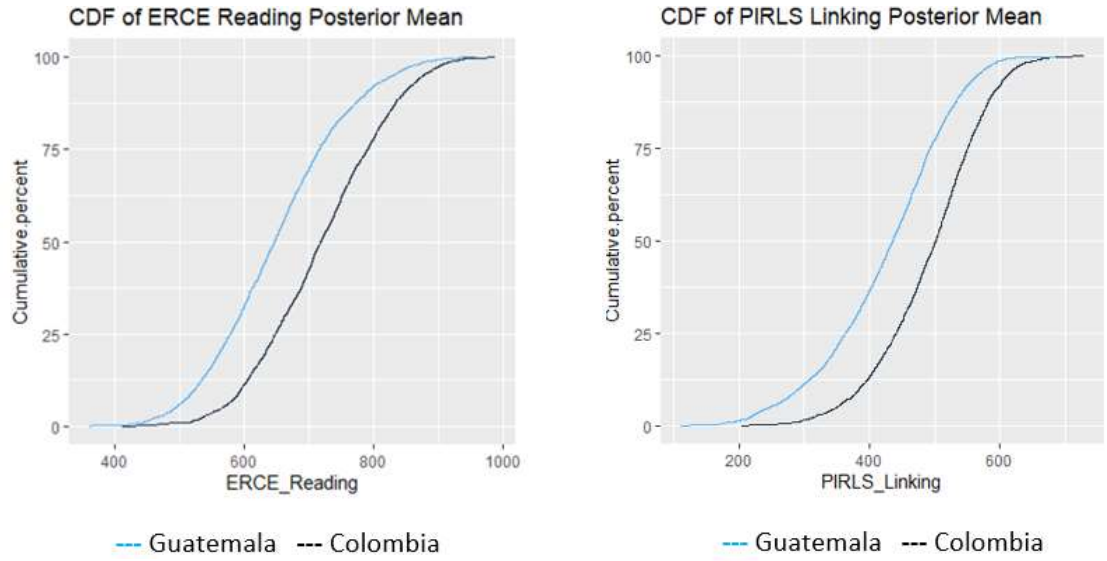


Exhibit 11.3: Cumulative Distributions of ERCE Reading and PIRLS Linking Data



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 ERCE data and Rosetta Stone linking data.

11.2 Creating Preliminary Concordance Tables

The concordance scores and levels were identified based on estimated ERCE posterior means using the combined data of the two countries. The score ranges of the posterior means of the ERCE mathematics and reading scales were either rounded up or down to cover almost all the data of the two countries and to be as symmetric as possible around the overall mean of the ERCE scale (which is 700). For both ERCE scales, mathematics and reading, scores range from about 400 to 1000 (covering almost 100% of the data) with very few data points beyond the range of 440 to 940 (covering about 99.5% of the data). Therefore, the following description of creating the concordance tables primarily focuses on the scores within the range from 440 to 940.

For both ERCE scales, mathematics and reading, 20 points on the ERCE 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 440 to 940.

For each identified concordance score level, PMM was used to select 5 donors from each of the two 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 ERCE test to each specified concordance score for each country. Only students who participated in all four tests, ERCE math, ERCE reading, TIMSS linking, and PIRLS linking were included for the donor selection: 2,619 students in Colombia and 3,902 students in Guatemala. Because the ERCE PVs have only one decimal, in some cases more than one student had a fifth smallest absolute difference value based on the approximated posterior means. In such cases, a random number was assigned to the students so that one of the students was randomly selected as the fifth donor. The mean and standard deviation of the donors' PVs from the Rosetta Stone linking data were calculated based on the total 50 donated PVs (2 countries * 5 donors * 5 PVs) at each concordance level. Note that these steps were implemented separately for ERCE mathematics and reading.

Preliminary concordance tables for ERCE mathematics and ERCE reading were created by assigning the estimated mean and standard deviation of each set of 50 PVs based on the Rosetta Stone (TIMSS

and PIRLS) linking data, respectively, to each concordance score level in the specified range of ERCE mathematics and ERCE 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 50 donated PVs were produced for each concordance score level between the range of 440 to 940 on the ERCE 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 and the smaller sample sizes (not all students could be included in the population modeling and donor selection), a smoothing procedure was utilized 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

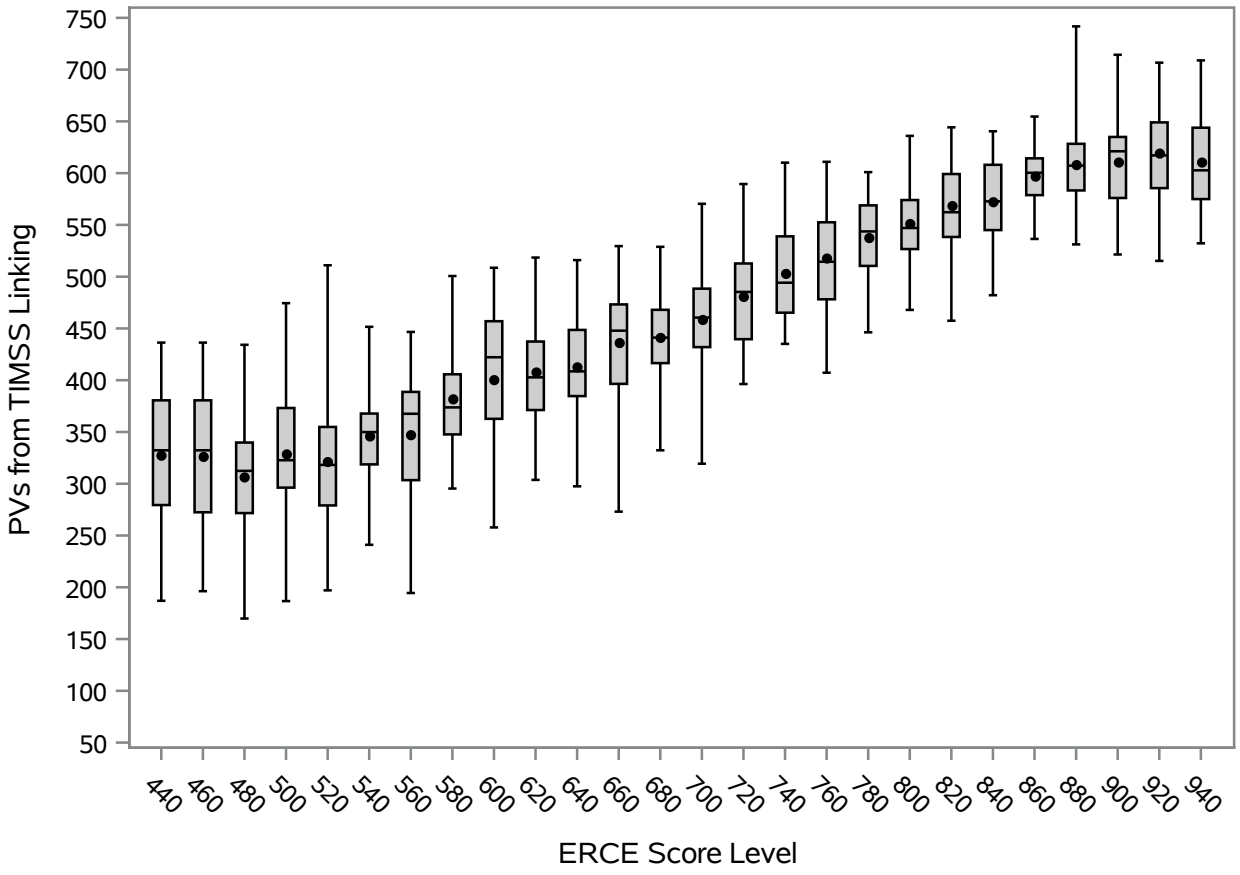
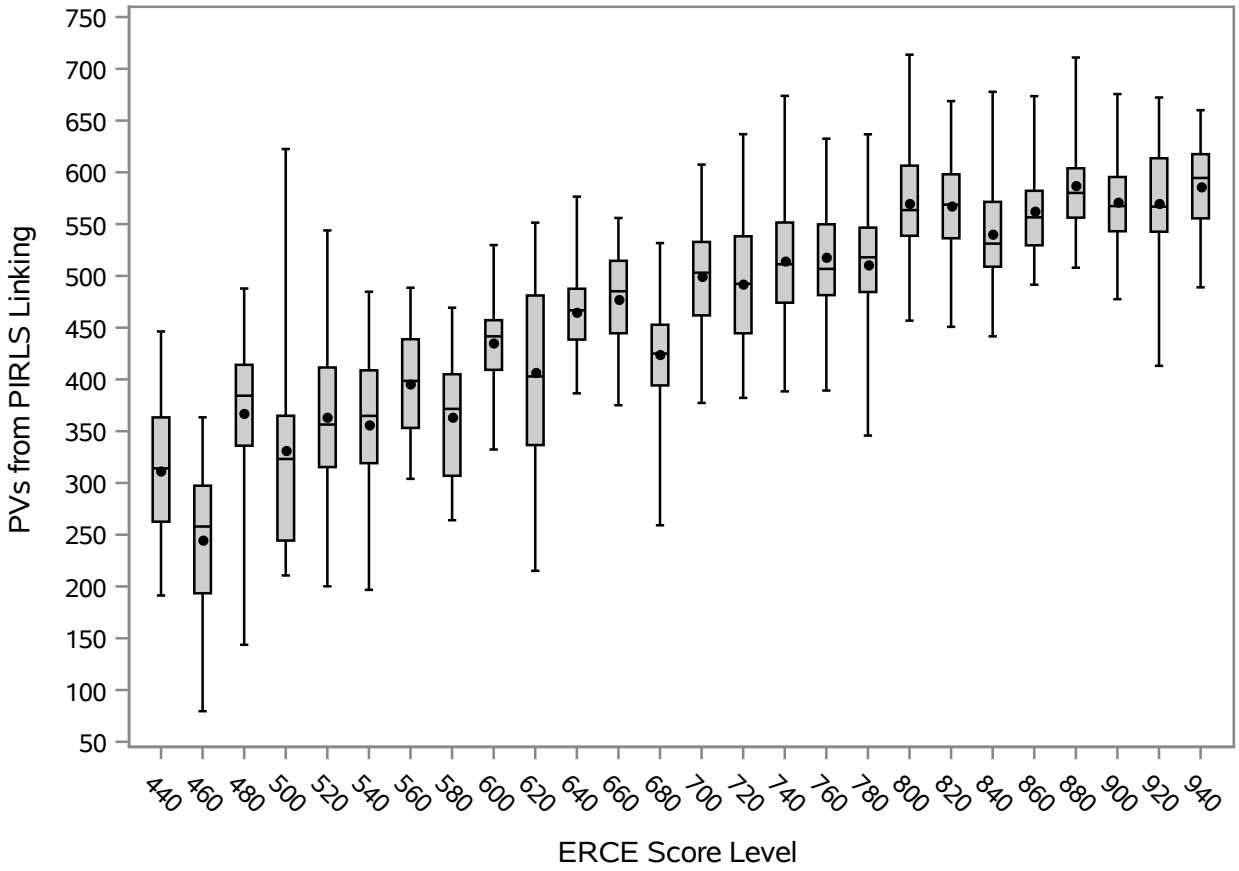


Exhibit 11.5: **Boxplots Plausible Values (PVs) from Selected Donors for Reading**



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 ERCE concordance scores beyond the range of 440 to 940, 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) ERCE score levels and the smoothed means were used to predict the conditional means of the likely posterior distributions at the concordance score levels 400, 420, 960, and 980. 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 ERCE mathematics and ERCE reading, respectively. The first column of each table shows the ERCE concordance score levels, either ERCE mathematics or ERCE 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 ERCE 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 ERCE 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 ERCE score level.

Exhibit 11.6: Concordance Table for ERCE Mathematics

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

Exhibit 11.7: Concordance Table for ERCE Reading

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

As an example of the usefulness of the concordance tables, the percentages of students in each country reaching the four TIMSS[2] and PIRLS [3] fourth grade benchmarks (Advanced: 625, High: 550, Intermediate: 475, Low: 400) for mathematics and reading were estimated and are illustrated in Exhibits 11.8a and 11.8b respectively.

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. Note, that the estimated percentages provided in Exhibit 11.8a and 11.8b are based on a sample of 6th graders (ERCE student population) while the TIMSS and PIRLS benchmarks are based on assessments at grade 4.

Exhibit 11.8a: Estimated Percentages of 6th-Grade ERCE Students Reaching the 4th-Grade TIMSS International Benchmarks

Estimated Percentages based on Rosetta Stone				
Country	**Advanced (625)**	**High (550)**	**Intermediate (475)**	**Low (400)**
Colombia	2.3 (0.5)	15.7 (1.4)	47.9 (2.4)	81.1 (1.7)
Guatemala	0.9 (0.3)	8.4 (0.9)	34.2 (1.6)	71.6 (1.7)
Average	**1.6 (0.3)**	**12.0 (0.8)**	**41.0 (1.4)**	**76.3 (1.2)**
Estimated Percentages based on Concordance				
Country	**Advanced (625)**	**High (550)**	**Intermediate (475)**	**Low (400)**
Colombia	2.7 (0.5)	16.4 (1.3)	48.2 (2.2)	81.3 (1.5)
Guatemala	1.1 (0.3)	8.1 (1.0)	30.4 (1.3)	66.6 (1.5)
Average	**1.9 (0.3)**	**12.3 (0.8)**	**39.3 (1.3)**	**73.9 (1.0)**

Note: Standard errors appear in parentheses.

2 A description of the TIMSS 2015 grade 4 mathematics benchmarks can be found here: http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/mathematics/performance-at-international-benchmarks/item-map-and-summary-of-international-benchmarks/

3 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/

Exhibit 11.8b: Estimated Percentages of 6th-Grade ERCE Students Reaching the 4th-Grade PIRLS International Benchmarks

Estimated Percentages based on Rosetta Stone				
Country	**Advanced (625)**	**High (550)**	**Intermediate (475)**	**Low (400)**
Colombia	4.9 (0.7)	26.1 (1.9)	61.2 (1.9)	85.9 (1.2)
Guatemala	1.0 (0.2)	9.1 (0.7)	32.7 (1.4)	63.2 (1.9)
Average	**3.0 (0.4)**	**17.6 (1.0)**	**46.9 (1.2)**	**74.5 (1.1)**
Estimated Percentages based on Concordance				
Country	**Advanced (625)**	**High (550)**	**Intermediate (475)**	**Low (400)**
Colombia	4.3 (0.6)	23.2 (1.5)	55.7 (1.7)	83.6 (1.3)
Guatemala	1.9 (0.3)	11.6 (0.8)	36.5 (1.4)	67.3 (1.4)
Average	**3.1 (0.3)**	**17.4 (0.8)**	**46.1 (1.1)**	**75.5 (1.0)**

Note: Standard errors appear in parentheses.

Overall, Exhibits 11.8a and 11.8b show that while there is small 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 the two countries provide highly comparable results. When interpreting the percentages of students reaching the low (400) and intermediate (475) TIMSS and PIRLS benchmarks it is important to keep in mind that the international TIMSS and PIRLS benchmarks are based on assessments at grade 4 while the Rosetta Stone study was administered to ERCE students at grade 6.

For the use of the concordance tables provided above a few cautionary notes are in order. First, while the constructs that are measured with the Rosetta Stone assessment and the ERCE assessment are highly correlated (especially TIMSS and ERCE mathematics) as indicated by the latent correlations between the scales, they were developed based on different frameworks and by different assessment development teams – with no intent to produce a parallel form – and for different target grades (grade 4 versus grade 6). Second, the estimates are based on only two out of 18 target (i.e., ERCE) countries and the originally targeted sample sizes were reduced, especially in Colombia where about one-third of the students (1,375 out of 4,467 students) did not receive Rosetta Stone linking booklets. Therefore, not all students could be included in the population modeling and the sample sizes were reduced even more for the main analysis part used to construct the concordance tables (donor selection) as only students with responses in all 4 constructs (TIMSS linking, PIRLS linking, ERCE mathematics, ERCE reading) could be used (2,619 students in Colombia and 3,902 students in Guatemala). Third, the scaling approach does not account for potential linking error. 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, an ERCE mathematics score of 700 does not result in a TIMSS score of 465. 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 414 and 517 (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 362 to 569, 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 ERCE concordance table for mathematics and reading.

Besides making inferences about the likely score range on TIMSS or PIRLS scales given an ERCE 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 ERCE assessments, first, the posterior mean of the conditional distribution for each student from the ERCE population model needs to be obtained and transformed onto the ERCE reporting metric. Next, the posterior means are rounded to the nearest ERCE score levels as shown in the first column of Exhibits 11.6 and 11.7, so that the projected means and standard deviations can be assigned to individual students according to the rounded ERCE 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, a limited number of countries and 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' ERCE score distributions on the TIMSS and PIRLS scales. For countries that 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

ERCE (2021). *Executive Report of the Comparative and Explanatory Regional Study (ERCE 2019)*, prepared by the Latin American Laboratory for Evaluation of the Quality of Education (LLECE), of the Regional Office of Education for Latin America and the Caribbean (OREALC/UNESCO Santiago), UNESCO: Paris. https://en.unesco.org/sites/default/files/resumen-ejecutivo-informe-regional-logros-factores-erce2019.pdf_0.pdf

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, L. (2016). Scaling the TIMSS 2015 Achievement Data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 13.1–13.62). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timss.bc.edu/publications/timss/2015-methods/chapter-13.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

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 ERCE Mathematics

ERCE 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
400	290	64	246	202	334	249	280	351	299	385	219	261	353	197	313	217	176	247	313	380	432	312
420	304	63	223	201	249	167	201	360	298	326	358	320	386	349	250	258	351	224	308	241	297	330
440	319	62	324	369	336	235	235	300	266	252	274	455	276	359	366	393	247	311	283	289	362	229
460	318	63	268	328	436	353	206	371	446	253	333	359	285	254	240	256	375	348	575	253	311	330
480	322	62	317	311	340	441	262	222	309	426	392	184	226	326	385	235	295	341	258	235	353	225
500	326	62	385	301	262	283	242	251	263	347	215	421	323	391	339	246	238	415	324	217	310	362
520	334	62	291	372	386	394	387	313	310	430	274	333	420	400	314	449	250	303	325	264	314	422
540	342	64	396	305	254	248	321	271	445	443	346	432	381	376	297	382	322	290	422	305	325	469
560	357	63	353	379	349	286	537	284	375	336	227	300	376	473	429	419	368	353	410	359	314	407
580	371	62	296	197	390	442	333	417	370	353	435	256	403	481	529	422	345	363	322	271	370	295
600	389	61	360	243	366	447	356	421	469	450	392	292	332	313	363	462	361	410	406	354	377	352
620	403	61	314	441	535	436	390	368	364	437	438	386	420	463	590	396	410	436	283	320	460	470
640	420	58	534	359	570	448	384	425	416	430	346	352	321	461	472	438	380	359	284	443	336	455
660	432	57	499	461	454	432	498	441	419	521	412	472	393	464	413	419	421	553	435	476	452	403
680	449	53	355	497	506	525	429	458	401	401	313	458	459	404	434	331	481	436	459	459	487	462
700	465	52	463	413	457	505	426	452	481	458	393	378	434	418	528	410	486	451	429	419	446	503
720	481	51	391	476	503	483	417	536	512	477	366	477	506	382	525	550	438	528	448	427	475	374
740	497	49	571	537	462	507	456	634	463	525	406	470	488	466	485	460	466	490	517	471	454	467
760	515	50	553	570	378	588	477	487	499	586	499	568	505	494	465	432	490	454	559	485	560	537
780	531	50	544	541	503	454	470	568	605	456	516	459	524	507	452	492	502	587	531	574	519	455
800	548	48	554	504	570	648	499	609	540	498	599	584	519	628	533	508	473	618	526	606	479	558
820	563	46	603	522	548	479	546	559	610	570	520	527	543	592	553	502	540	563	514	565	605	569
840	576	46	640	551	631	575	617	634	556	542	611	504	552	639	524	547	653	585	597	467	618	576
860	590	45	639	609	566	621	643	527	610	643	607	550	592	595	583	554	580	643	551	628	675	606
880	599	46	607	626	595	563	629	601	509	554	529	763	741	620	536	563	633	699	633	624	613	522
900	608	46	563	592	651	666	684	574	592	619	559	514	631	544	679	724	647	586	696	628	655	652
920	617	48	617	574	663	608	612	582	627	658	607	598	657	665	663	607	599	610	577	645	590	727
940	624	51	699	701	641	653	598	658	657	684	627	550	577	639	677	638	628	690	620	650	621	618
960	638	50	626	732	640	569	727	640	588	667	670	620	631	649	592	542	659	677	705	634	748	686
980	653	49	618	612	694	698	602	715	737	633	633	716	648	682	652	640	644	699	642	645	553	640

ERCE 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
400	290	64	355	273	273	298	276	327	347	254	208	309	342	257	317	261	285	206	298	196	354	208
420	304	63	450	324	429	238	360	255	144	373	318	371	314	339	191	261	200	307	355	214	274	358
440	319	62	219	285	320	228	347	294	404	309	365	414	292	144	193	242	371	319	246	301	380	315
460	318	63	176	280	253	398	310	316	315	296	297	341	372	448	382	258	251	295	278	467	363	430
480	322	62	349	210	420	320	291	351	318	331	242	253	251	299	182	334	268	242	283	230	366	396
500	326	62	310	288	343	370	452	308	243	398	355	280	279	336	380	360	372	259	356	278	298	333
520	334	62	345	343	319	345	368	289	333	316	341	352	327	448	387	327	389	295	295	383	249	411
540	342	64	406	243	375	401	362	416	297	201	412	386	258	385	401	366	447	492	416	350	279	437
560	357	63	327	412	405	303	321	322	373	279	500	300	454	386	361	296	397	236	296	280	354	439
580	371	62	361	406	279	402	377	426	429	376	478	423	362	287	456	396	423	360	370	470	340	333
600	389	61	360	418	413	338	390	307	312	405	333	378	300	405	372	372	304	473	383	383	329	307
620	403	61	510	422	344	436	331	375	391	459	319	488	333	362	341	363	506	415	417	350	472	470
640	420	58	362	447	418	426	373	468	398	392	369	489	421	472	396	396	370	377	431	406	477	357
660	432	57	422	327	367	429	461	426	337	459	430	408	452	434	361	364	401	437	483	397	419	377
680	449	53	392	407	499	499	392	373	417	369	428	443	571	507	441	449	417	475	363	402	454	369
700	465	52	358	470	445	439	527	404	408	469	432	547	408	564	476	483	458	495	443	448	511	502
720	481	51	538	449	460	459	424	438	544	513	473	394	484	532	577	458	526	462	408	474	468	463
740	497	49	545	443	465	451	521	535	489	463	540	497	437	517	488	489	497	557	573	444	542	554
760	515	50	631	450	470	475	487	502	499	509	541	542	544	393	545	558	492	549	543	493	634	565
780	531	50	488	500	588	588	523	490	535	563	505	503	567	485	497	593	516	487	578	535	580	635
800	548	48	512	497	592	487	500	478	582	497	582	558	464	610	539	575	496	440	587	644	539	581
820	563	46	574	614	584	530	564	522	546	623	549	599	711	611	549	594	489	493	500	591	661	542
840	576	46	556	529	510	523	483	506	548	565	663	526	561	550	577	556	655	651	565	566	603	535
860	590	45	565	620	568	634	526	516	590	526	608	568	597	630	657	606	651	644	538	558	535	597
880	599	46	607	544	691	680	703	538	706	697	507	538	623	639	632	603	557	577	590	620	645	598
900	608	46	570	547	545	625	626	584	480	674	700	543	650	682	675	575	655	685	589	551	498	559
920	617	48	631	717	579	640	600	581	629	578	641	633	613	627	651	586	677	639	625	636	521	573
940	624	51	659	645	687	646	749	722	658	687	654	626	721	514	617	528	666	633	621	527	551	648
960	638	50	650	600	562	673	667	688	688	675	635	635	629	727	692	695	502	574	654	668	608	647
980	653	49	653	640	621	562	565	599	766	673	651	690	655	703	557	574	592	645	673	720	643	636

ERCE 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
400	290	64	346	412	373	371	330	320	296	282	283	278	295	388	334	239	242	359	291	234	279	368
420	304	63	329	302	308	356	184	288	314	324	241	289	361	230	416	227	260	314	382	293	329	332
440	319	62	265	322	370	268	453	231	283	446	345	408	345	203	344	425	310	405	292	297	269	342
460	318	63	273	374	336	437	258	309	387	227	412	327	277	307	318	416	358	350	277	397	241	430
480	322	62	235	406	341	373	420	311	294	235	306	275	304	201	178	226	269	355	295	232	277	384
500	326	62	377	290	385	331	370	340	279	330	252	237	274	320	255	368	328	394	383	284	306	318
520	334	62	305	334	278	370	354	416	373	407	259	425	326	326	241	246	305	296	329	363	261	286
540	342	64	340	216	323	357	380	358	409	373	302	310	353	337	374	476	452	422	273	346	250	244
560	357	63	360	301	366	343	311	350	299	409	427	301	333	303	336	502	352	383	386	392	350	521
580	371	62	291	336	293	431	400	383	318	445	407	347	244	248	420	320	352	411	383	434	423	434
600	389	61	489	397	333	353	394	386	478	461	324	402	356	445	338	366	438	469	434	386	354	410
620	403	61	473	368	445	362	450	409	487	342	393	427	375	436	469	334	390	393	447	380	442	356
640	420	58	474	480	443	451	419	369	520	418	453	436	461	316	387	338	455	465	370	461	426	395
660	432	57	463	425	429	473	448	428	451	438	410	433	432	438	511	359	385	342	416	400	439	366
680	449	53	515	477	409	390	462	404	445	363	371	456	370	517	447	382	390	487	435	482	476	365
700	465	52	415	516	433	474	511	490	465	444	500	365	492	449	370	417	512	416	478	388	406	487
720	481	51	517	555	467	528	499	478	550	500	462	549	435	458	433	434	521	437	425	445	353	431
740	497	49	590	553	493	517	455	447	452	594	531	429	470	553	454	484	487	538	453	522	508	491
760	515	50	478	514	549	464	497	514	470	509	520	588	411	588	486	450	555	443	593	530	539	577
780	531	50	546	476	548	499	504	531	508	628	626	542	558	534	511	516	429	536	541	590	604	487
800	548	48	556	624	422	578	514	504	564	568	530	537	494	518	514	559	559	497	606	562	631	528
820	563	46	488	589	591	543	592	595	554	537	563	561	595	505	556	495	608	569	527	520	511	573
840	576	46	637	594	567	600	527	584	620	512	516	598	536	548	470	600	567	561	531	596	530	527
860	590	45	510	545	518	671	660	641	555	524	599	647	637	578	552	539	563	623	590	623	598	558
880	599	46	647	606	656	565	552	663	624	567	579	564	584	574	668	591	587	578	554	538	517	590
900	608	46	610	583	649	579	615	519	638	661	610	573	586	604	517	618	581	584	603	561	522	627
920	617	48	672	631	651	655	656	678	626	583	547	606	709	665	521	697	641	674	591	601	575	672
940	624	51	657	646	718	609	633	548	632	624	676	569	521	628	646	618	543	688	648	684	632	598
960	638	50	621	587	627	672	749	673	636	603	591	650	634	638	617	570	581	561	732	605	567	593
980	653	49	753	730	642	625	623	627	631	652	603	626	610	667	587	581	670	658	595	740	646	606

ERCE 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
400	290	64	394	265	255	217	207	166	335	213	335	267	342	295	321	316	187	306	341	425	293	306
420	304	63	243	244	168	322	282	308	445	406	200	189	241	371	253	299	305	289	345	236	198	344
440	319	62	342	254	431	299	257	364	282	238	302	303	309	379	334	301	340	215	286	306	189	295
460	318	63	345	361	224	325	241	275	306	263	361	293	302	306	389	333	392	325	335	166	315	352
480	322	62	366	326	368	303	377	380	292	281	389	310	300	438	316	274	349	344	316	172	340	260
500	326	62	393	376	249	367	401	275	330	365	275	362	282	301	295	463	454	296	327	295	379	328
520	334	62	283	211	352	238	380	277	398	261	379	357	323	269	324	223	374	367	344	342	310	369
540	342	64	493	348	329	313	334	314	272	302	321	349	346	489	273	434	315	246	421	319	268	334
560	357	63	293	437	350	381	350	223	382	394	347	480	326	363	344	384	250	378	417	358	279	424
580	371	62	281	355	306	320	425	284	305	300	348	410	298	304	365	404	371	310	499	328	377	297
600	389	61	336	366	371	513	406	369	352	365	425	216	427	305	308	342	318	480	363	277	375	363
620	403	61	422	393	419	375	319	369	421	405	363	324	487	422	345	467	468	390	377	398	255	447
640	420	58	313	392	322	342	496	322	537	489	510	381	448	325	455	452	425	318	469	409	424	440
660	432	57	415	463	368	363	426	364	531	547	466	417	383	477	460	334	437	529	415	515	345	481
680	449	53	523	428	383	405	471	471	505	489	435	404	487	477	493	512	450	380	493	352	471	391
700	465	52	477	510	376	464	522	456	523	514	362	468	441	443	519	434	489	493	389	473	399	418
720	481	51	482	558	500	539	508	508	537	570	499	412	499	487	496	468	564	555	410	559	519	515
740	497	49	542	568	472	545	476	518	547	488	513	444	484	503	451	488	538	487	539	490	493	526
760	515	50	544	534	511	557	484	411	589	420	418	482	471	539	590	558	499	559	598	489	595	500
780	531	50	514	621	549	534	526	591	598	606	521	482	527	531	543	520	498	507	468	597	604	566
800	548	48	523	667	555	515	513	496	522	534	584	616	462	594	524	572	558	420	493	617	500	571
820	563	46	538	619	513	604	564	594	548	527	599	562	608	558	643	552	521	618	579	603	604	568
840	576	46	529	536	545	493	506	541	612	582	524	612	558	573	623	589	514	546	615	551	609	600
860	590	45	627	530	551	565	560	581	634	619	543	701	545	588	529	633	587	543	558	613	587	596
880	599	46	526	569	592	605	734	574	643	584	622	599	581	635	564	581	579	685	675	579	533	586
900	608	46	551	614	624	654	618	618	601	691	640	637	554	576	580	585	575	571	560	561	600	653
920	617	48	608	699	637	622	567	545	600	658	564	592	642	683	634	604	630	600	601	624	635	537
940	624	51	634	612	632	610	556	623	635	615	607	571	643	565	664	570	604	545	644	655	551	579
960	638	50	725	630	617	572	618	582	671	633	618	605	581	604	679	607	697	671	574	516	593	639
980	653	49	616	723	734	540	556	676	750	705	671	607	613	695	716	567	700	677	647	671	697	556

ERCE 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
400	290	64	376	330	250	385	217	326	287	419	220	285	325	122	270	281	364	225	346	352	292	346
420	304	63	258	278	274	290	280	348	324	217	379	366	401	228	250	270	297	364	315	273	293	332
440	319	62	444	189	315	291	260	225	301	398	346	331	360	283	275	168	292	279	298	369	306	428
460	318	63	263	377	296	192	224	283	392	363	380	386	393	208	337	386	362	265	384	292	322	287
480	322	62	386	367	312	429	379	307	406	316	292	169	330	281	296	234	267	419	274	327	306	309
500	326	62	440	331	232	287	287	322	429	283	343	344	291	356	404	434	314	264	303	412	324	293
520	334	62	213	263	431	317	421	289	431	357	291	315	280	433	304	298	404	280	326	254	330	401
540	342	64	201	379	371	351	235	293	267	377	358	387	207	491	223	329	430	283	326	328	343	358
560	357	63	242	412	346	367	302	313	370	180	267	297	318	337	370	278	422	431	311	318	352	380
580	371	62	330	266	310	310	373	382	327	347	290	415	341	263	301	304	408	405	256	373	357	373
600	389	61	388	342	408	380	388	443	304	327	404	364	401	195	492	369	406	413	423	360	374	416
620	403	61	428	359	405	499	398	405	372	305	416	313	432	435	416	423	512	373	509	347	403	397
640	420	58	424	404	481	407	356	426	513	414	453	394	368	480	430	394	346	368	446	342	412	328
660	432	57	469	363	481	497	420	430	419	367	419	300	397	409	407	493	458	486	441	398	426	420
680	449	53	456	490	500	349	529	530	385	491	522	450	430	360	425	488	590	482	421	478	440	427
700	465	52	363	453	399	497	499	439	381	359	431	508	456	448	393	435	449	535	386	524	449	334
720	481	51	445	435	459	561	443	481	559	466	463	451	350	469	488	402	493	487	513	585	476	494
740	497	49	485	487	438	477	535	549	463	542	450	505	456	433	495	500	539	483	450	514	493	519
760	515	50	517	599	503	587	465	597	409	605	513	543	546	596	453	578	662	493	604	379	514	473
780	531	50	522	507	561	481	484	587	497	588	547	581	553	474	493	483	513	578	560	524	527	582
800	548	48	600	521	599	602	552	466	596	593	643	464	608	578	454	490	495	521	546	578	540	546
820	563	46	631	632	508	584	539	601	621	511	576	546	518	562	539	517	465	567	530	450	556	569
840	576	46	636	567	544	562	551	538	549	615	577	532	630	651	569	576	616	582	482	547	563	562
860	590	45	563	577	634	661	542	564	618	650	538	560	567	633	518	546	521	582	587	523	582	643
880	599	46	602	562	653	516	628	671	576	608	668	616	581	633	670	563	567	623	684	671	600	522
900	608	46	624	555	604	640	635	556	508	584	663	528	661	680	629	703	620	639	609	607	601	679
920	617	48	610	620	571	541	610	598	638	615	646	543	622	630	651	557	576	578	682	684	614	561
940	624	51	574	654	586	591	630	589	644	642	584	678	683	618	623	608	657	589	644	637	621	580
960	638	50	630	724	616	637	592	656	585	636	652	600	698	661	645	685	576	670	611	619	630	617
980	653	49	682	678	616	715	728	670	642	633	757	661	590	609	693	667	593	639	685	631	645	694

APPENDIX B

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

ERCE 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
400	284	84	329	291	367	273	245	279	295	356	172	327	368	303	362	299	352	218	290	317	348	312
420	296	83	440	317	305	327	230	336	163	233	327	222	332	282	215	231	400	289	246	337	294	267
440	309	82	472	235	446	310	352	323	250	185	222	197	408	356	497	229	191	346	139	287	202	199
460	321	81	270	456	260	458	342	257	374	180	325	335	201	422	527	433	371	425	251	330	325	276
480	329	80	247	296	250	338	284	221	508	295	292	373	303	302	408	221	359	364	258	237	340	403
500	342	77	311	288	406	243	331	412	515	445	261	234	374	413	482	349	415	329	214	279	314	445
520	347	76	410	266	321	347	428	400	313	347	292	345	352	320	198	363	329	412	364	438	331	282
540	373	74	300	312	383	269	325	388	267	475	395	365	344	368	471	438	356	474	298	306	244	460
560	378	74	414	337	444	402	420	357	523	440	470	383	304	432	299	239	525	413	420	350	309	211
580	395	68	262	385	264	336	335	357	417	440	466	224	374	431	371	272	390	372	326	404	479	420
600	410	65	335	493	387	329	500	316	572	503	297	308	434	465	347	352	262	463	532	423	490	468
620	421	62	373	287	508	402	445	407	323	382	430	459	374	409	375	328	390	367	480	363	395	389
640	436	62	500	452	476	372	529	428	449	492	529	406	453	296	438	484	329	517	457	468	435	399
660	455	63	446	413	433	423	481	448	267	613	453	499	503	470	397	529	441	583	569	512	426	441
680	467	64	374	539	558	408	443	550	542	495	387	510	536	445	506	471	360	375	435	468	465	440
700	482	60	420	446	468	372	472	494	490	506	469	455	559	458	546	428	450	588	570	470	451	465
720	488	63	333	505	511	516	511	546	440	502	505	356	458	532	420	489	511	468	317	373	475	573
740	501	65	510	547	633	649	534	519	585	499	497	355	478	500	504	530	505	552	479	547	491	462
760	522	64	535	539	478	574	405	610	619	569	544	614	478	580	346	499	614	584	514	521	464	578
780	526	63	523	596	536	567	426	498	559	504	510	462	546	460	452	477	465	509	570	505	506	600
800	536	61	550	507	422	560	534	556	556	521	517	424	517	616	429	560	613	523	501	550	477	644
820	546	60	683	456	433	513	493	591	556	554	505	627	507	667	508	526	527	539	605	563	530	632
840	555	60	571	628	471	564	655	633	602	656	617	595	576	527	601	667	513	594	594	662	494	561
860	565	59	524	568	560	477	583	569	629	614	602	558	671	576	469	560	629	568	559	604	633	572
880	568	56	542	549	530	593	558	532	550	558	538	644	496	586	609	653	593	530	628	742	599	474
900	576	56	620	483	524	491	495	596	509	591	561	634	484	537	607	525	552	601	559	640	582	659
920	588	58	477	556	693	632	581	559	644	581	656	619	598	461	551	717	558	579	532	598	507	562
940	596	59	543	523	680	558	549	611	740	623	581	691	555	570	582	612	630	585	669	570	687	522
960	608	58	620	631	653	544	565	711	587	567	716	561	547	630	629	570	594	577	595	604	570	535
980	620	57	738	603	641	657	698	647	588	690	664	655	565	677	584	587	604	680	698	574	532	711

ERCE 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
400	284	84	217	225	211	324	328	405	177	309	371	249	250	138	187	184	330	342	349	258	356	200
420	296	83	173	143	181	357	377	501	292	326	328	402	289	435	458	219	296	365	207	130	406	409
440	309	82	334	167	303	280	278	351	328	228	285	460	358	251	327	447	206	344	212	359	331	307
460	321	81	193	294	280	372	287	378	386	338	323	355	299	170	210	263	406	279	181	234	368	374
480	329	80	488	410	348	264	265	490	339	201	376	289	470	467	264	396	413	339	360	369	272	309
500	342	77	374	305	411	369	304	386	482	339	461	430	427	364	371	433	221	351	473	422	257	406
520	347	76	269	337	293	418	321	349	365	438	337	294	330	262	481	399	437	358	371	218	298	474
540	373	74	285	334	378	316	302	305	225	248	320	268	421	401	325	238	260	352	455	550	358	347
560	378	74	406	315	355	349	257	422	396	490	429	595	386	334	385	380	300	428	315	95	375	375
580	395	68	478	318	277	432	436	419	540	448	436	420	462	343	463	472	479	356	466	441	379	337
600	410	65	369	370	429	391	332	358	338	416	380	446	311	353	389	458	341	468	420	480	547	390
620	421	62	375	390	404	423	389	467	424	520	399	328	336	472	344	419	524	391	386	501	358	461
640	436	62	561	363	517	522	484	381	393	550	467	505	379	343	417	442	471	479	428	319	407	434
660	455	63	364	484	478	420	425	548	407	430	533	612	466	530	443	503	444	506	547	369	428	328
680	467	64	402	421	362	515	406	460	543	502	545	537	499	477	428	554	395	509	454	404	464	274
700	482	60	409	613	559	477	532	574	433	517	513	537	461	459	384	418	606	590	529	565	449	420
720	488	63	443	457	498	440	456	504	439	393	403	446	444	517	515	533	538	450	536	524	627	519
740	501	65	550	463	492	581	552	592	426	500	494	456	597	661	416	530	483	565	478	554	580	438
760	522	64	498	503	468	535	429	532	480	525	558	519	581	552	549	504	610	458	576	536	311	639
780	526	63	541	410	475	604	496	616	599	480	519	511	386	575	560	468	538	509	430	520	540	512
800	536	61	563	513	386	532	529	504	555	477	557	484	546	558	604	436	527	516	520	639	462	567
820	546	60	574	614	458	552	639	649	519	555	653	465	631	564	423	614	595	504	587	536	586	566
840	555	60	518	434	527	533	476	590	540	477	543	647	555	597	547	623	586	547	580	523	534	574
860	565	59	688	491	622	548	565	573	578	628	492	633	510	599	557	602	528	575	588	578	663	650
880	568	56	494	563	518	512	555	567	515	554	608	512	444	507	585	481	495	577	557	491	571	539
900	576	56	598	643	677	617	683	489	622	604	617	520	563	471	491	560	478	567	604	517	576	610
920	588	58	603	533	736	720	662	580	679	530	676	652	557	602	614	564	515	654	670	595	538	464
940	596	59	598	621	430	602	601	608	533	665	523	683	642	620	533	567	681	547	431	633	586	671
960	608	58	633	593	598	595	662	551	555	505	713	611	602	588	612	630	619	547	595	557	585	618
980	620	57	640	592	629	718	625	569	600	535	610	635	612	648	546	653	635	575	738	614	661	575

ERCE 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
400	284	84	116	376	156	267	340	227	282	352	469	300	255	176	311	335	210	310	354	178	262	233
420	296	83	276	228	313	234	217	354	469	445	241	320	251	360	365	352	356	515	312	329	355	227
440	309	82	507	418	273	319	321	297	344	330	405	94	297	335	295	362	336	336	279	340	328	441
460	321	81	226	482	351	341	344	475	354	284	368	372	274	264	320	378	334	419	358	291	323	313
480	329	80	318	328	330	311	291	361	267	329	441	229	267	172	202	399	235	307	273	410	397	551
500	342	77	375	428	308	347	372	315	416	314	357	389	386	242	324	339	385	254	349	394	385	408
520	347	76	432	374	269	365	328	293	319	420	329	334	499	430	333	364	469	330	241	345	291	303
540	373	74	337	297	424	337	356	327	282	346	226	398	450	458	541	343	428	412	268	461	365	541
560	378	74	399	425	301	467	301	338	287	426	253	397	391	423	377	337	430	371	315	527	458	291
580	395	68	335	397	399	207	473	415	399	307	346	331	430	457	343	489	471	374	421	354	367	482
600	410	65	337	376	295	540	246	476	583	469	526	410	400	516	472	392	178	547	447	451	354	412
620	421	62	382	425	387	415	348	367	390	388	381	516	460	411	450	449	400	423	464	407	447	329
640	436	62	461	394	452	435	647	529	420	354	401	468	407	324	448	470	587	388	383	435	329	444
660	455	63	492	509	402	495	487	441	528	370	374	422	415	457	464	469	504	462	470	430	595	452
680	467	64	382	412	398	477	512	394	419	528	518	511	524	443	491	486	409	601	446	570	450	539
700	482	60	439	477	432	473	490	466	514	479	445	483	513	418	515	462	447	452	549	540	550	490
720	488	63	437	491	487	538	713	434	479	418	443	496	458	503	497	503	439	593	560	507	488	459
740	501	65	499	473	554	466	472	560	436	453	462	559	526	572	567	426	531	423	478	408	392	494
760	522	64	454	544	547	550	499	419	566	514	549	377	511	499	603	626	581	480	497	520	518	540
780	526	63	399	544	561	434	317	504	535	547	486	569	481	510	413	466	453	581	488	568	499	414
800	536	61	485	621	473	542	608	595	555	431	464	525	596	559	486	564	660	469	429	502	402	463
820	546	60	461	525	488	466	549	430	528	520	557	601	446	579	545	544	561	489	550	532	455	468
840	555	60	582	618	505	519	562	522	638	432	451	525	487	490	516	514	498	616	616	540	512	488
860	565	59	553	478	579	663	638	516	528	569	562	601	560	449	450	551	615	650	587	508	466	635
880	568	56	611	613	566	522	570	433	656	518	609	614	520	592	594	596	602	604	605	499	655	605
900	576	56	572	630	619	595	619	610	547	639	543	597	596	596	580	563	546	618	611	626	581	472
920	588	58	566	579	461	594	590	519	551	582	535	605	523	592	566	602	608	620	529	538	654	502
940	596	59	575	619	670	554	653	693	509	610	554	513	650	662	587	590	567	481	515	591	618	728
960	608	58	686	690	645	638	553	620	645	528	515	651	685	627	693	530	636	619	598	631	597	722
980	620	57	632	623	707	706	585	667	546	663	639	606	626	620	580	629	708	609	652	657	654	656

ERCE 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
400	284	84	355	329	226	246	290	168	351	328	162	311	371	307	192	203	292	308	387	385	186	323
420	296	83	277	223	285	318	286	400	395	271	327	322	255	255	399	309	179	342	289	300	357	451
440	309	82	307	492	365	287	323	388	331	323	299	247	241	327	417	344	368	250	316	342	271	30
460	321	81	414	286	221	368	265	363	407	304	335	385	322	274	371	287	320	326	261	324	188	381
480	329	80	425	383	357	297	303	317	351	434	292	276	336	363	251	229	269	244	322	305	344	181
500	342	77	141	454	354	272	375	344	340	401	272	377	373	259	239	323	363	254	390	414	236	446
520	347	76	413	508	291	396	383	401	424	363	327	253	289	427	281	314	491	379	375	314	299	256
540	373	74	482	418	497	384	320	494	397	342	307	362	532	165	400	377	275	305	290	564	303	371
560	378	74	398	284	392	283	231	297	225	359	394	376	376	280	370	271	364	382	299	399	266	264
580	395	68	363	464	403	350	308	479	431	579	423	456	437	451	367	471	426	433	336	405	401	366
600	410	65	467	421	395	405	373	411	255	424	409	401	390	412	446	396	430	416	330	388	335	369
620	421	62	427	454	387	423	396	305	423	469	437	436	398	398	494	403	532	369	466	440	515	430
640	436	62	451	596	404	482	468	430	423	391	351	541	476	430	545	401	468	438	343	483	440	499
660	455	63	467	433	386	369	429	445	417	576	386	427	384	410	404	485	465	388	570	541	569	371
680	467	64	439	389	602	490	577	416	444	468	456	481	419	458	430	505	421	537	455	432	617	492
700	482	60	423	584	469	531	485	356	497	460	485	444	537	449	564	541	440	512	519	547	545	395
720	488	63	512	566	571	419	574	432	427	447	482	519	405	580	530	489	384	563	474	568	565	521
740	501	65	525	562	552	523	424	448	384	499	576	550	516	462	580	444	489	451	653	526	533	502
760	522	64	496	495	532	562	604	424	582	514	448	555	399	537	659	449	586	586	532	560	552	606
780	526	63	554	444	585	429	381	500	480	565	516	514	539	523	461	582	619	507	549	696	648	549
800	536	61	558	687	614	484	508	623	571	575	552	584	608	504	525	552	504	584	519	579	470	641
820	546	60	537	592	482	555	470	524	608	542	549	589	445	611	474	547	581	444	512	491	543	646
840	555	60	576	535	514	614	651	531	563	533	548	588	456	593	531	551	508	519	490	549	472	598
860	565	59	491	586	581	562	558	587	666	625	584	561	513	599	610	504	532	606	611	564	556	633
880	568	56	512	590	608	540	542	663	598	508	678	538	564	514	610	521	632	594	553	551	601	607
900	576	56	588	547	546	539	569	589	645	507	620	565	598	630	597	513	677	616	545	547	581	571
920	588	58	593	604	642	581	518	661	618	593	587	619	642	458	468	614	551	587	708	552	514	615
940	596	59	566	522	648	509	542	577	546	623	668	587	681	661	536	615	586	564	584	625	539	556
960	608	58	625	587	772	601	627	503	592	564	637	746	646	641	547	590	559	527	545	558	522	626
980	620	57	605	612	737	542	618	762	592	618	560	665	632	594	605	732	606	527	569	623	673	604

ERCE 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
400	284	84	206	364	160	258	141	336	123	202	204	197	555	61	263	429	299	393	285	304	279	225
420	296	83	352	309	332	332	416	199	173	409	406	198	316	437	309	407	216	479	278	341	311	417
440	309	82	395	249	178	292	229	388	274	370	385	245	167	184	249	357	229	274	254	135	302	264
460	321	81	268	462	496	354	247	272	237	271	328	401	237	372	249	230	160	384	356	316	321	354
480	329	80	236	298	300	412	485	308	338	301	306	326	253	158	316	463	268	378	448	419	325	296
500	342	77	350	326	399	402	373	357	440	381	228	478	414	437	308	280	245	183	250	304	349	516
520	347	76	256	391	273	439	337	242	441	287	430	440	252	326	341	228	331	375	327	400	347	407
540	373	74	365	554	341	134	433	397	399	472	485	222	493	329	340	277	329	380	315	451	363	380
560	378	74	453	332	449	289	443	389	342	286	320	419	397	482	324	481	475	404	533	372	367	369
580	395	68	432	479	314	400	439	331	367	314	396	350	417	407	394	471	466	387	420	370	395	389
600	410	65	320	459	442	279	368	491	360	378	379	416	421	363	383	497	485	359	377	416	402	430
620	421	62	436	434	406	328	421	310	392	415	432	576	365	322	474	399	340	490	474	249	408	400
640	436	62	311	498	433	470	473	510	422	503	416	463	438	465	351	467	479	392	500	485	442	388
660	455	63	353	461	400	389	470	529	440	305	424	429	411	400	388	350	387	451	466	435	448	419
680	467	64	486	509	438	586	366	606	568	443	492	449	394	507	483	426	456	494	455	477	467	370
700	482	60	391	514	490	474	451	377	543	446	433	425	488	332	499	421	538	519	426	494	478	478
720	488	63	438	466	538	484	507	412	485	527	465	560	412	368	476	521	594	509	436	437	482	489
740	501	65	529	517	447	547	473	642	600	457	549	570	529	487	529	513	537	525	544	578	509	456
760	522	64	582	517	560	634	539	521	487	505	519	631	474	565	536	523	564	409	371	620	522	510
780	526	63	576	572	556	434	530	614	596	510	528	527	561	566	632	507	528	535	514	565	514	530
800	536	61	494	649	623	569	529	502	531	491	528	515	492	598	484	576	619	514	587	644	533	463
820	546	60	564	571	505	535	488	629	481	597	621	614	529	640	460	511	549	518	552	531	538	586
840	555	60	558	528	612	518	536	487	591	572	618	608	541	594	545	543	526	586	559	578	550	539
860	565	59	542	556	625	559	658	528	536	551	533	559	532	688	660	473	544	509	573	534	567	479
880	568	56	673	557	582	543	522	589	627	568	521	533	578	631	537	627	597	467	580	619	562	566
900	576	56	577	409	562	568	562	594	619	581	635	578	663	676	503	543	652	605	530	600	572	521
920	588	58	593	560	558	502	668	523	550	527	621	627	563	669	593	573	676	624	619	570	581	527
940	596	59	570	638	491	527	607	644	624	574	514	523	635	715	622	680	537	622	580	555	588	518
960	608	58	701	603	582	692	539	637	654	766	633	675	544	555	574	657	576	668	547	609	604	584
980	620	57	684	658	568	699	669	651	620	548	611	603	674	631	685	662	650	591	483	550	624	667

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 ERCE sample for each student in the domain of interest, either ERCE mathematics or ERCE reading, to obtain an approximate posterior mean on the ERCE scale for each studen

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 ERCE 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:

$$t_0 = \frac{1}{5}\sum_{m=1}^{5} t_m \qquad \text{(C.1)}$$

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 $\frac{6}{5}$:

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

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:

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

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

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

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

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

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 2 other countries, not including the present country. Therefore, there are sources of error variance and bias that are not reflected in the projections.

unesco
Institute for Statistics

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

Rosetta Stone Analysis Report: **Establishing a Concordance between ERCE and TIMSS/PIRLS**