CHAPTER 10

PIRLS 2021 Achievement Scaling Methodology: Item Response Theory, Population Models, and Linking Across Modes

Ummugul Bezirhan
Pierre Foy
Matthias von Davier

Introduction

This chapter explains the statistical and psychometric methods used in the analysis of the data from PIRLS 2021. The first part of this chapter covers a review of item response theory (IRT), a methodology frequently used in educational measurement that is also common in other applications of quantitative analysis of human response data such as patient-reported outcomes, consumer choice, and other domains. Building on these foundations, the challenges introduced by a hybrid assessment database consisting of both computer-based and paper-based country data are addressed. In PIRLS 2021, about half of the countries administered the digital version (digitalPIRLS) and the other half administered the paper version (paperPIRLS) consistent with previous PIRLS assessments. digitalPIRLS countries also administered paper instruments with trend material called bridge booklets to a randomly selected equivalent group of students to control for possible mode of administration effects in PIRLS 2021.

The second part of the chapter describes the “randomly equivalent samples” design, a population-based linking approach that allows controlling for mode of administration effects on observed student response behavior and produces a latent variable scale representing student proficiency that is comparable across paper- and computer-based assessments.

The third part of this chapter reviews the integration of achievement data from the PIRLS 2021 items with contextual data from student and parent questionnaires and describes the statistical imputation model used for this purpose. This model is a combination of IRT approaches and a regression-based approach that utilizes the context data as predictors for the derivation of a prior distribution of proficiency and is essentially the approach adopted by PIRLS since the first assessment in 2001.

https://doi.org/10.6017/lse.tpisc.tr2103.kb3131

This chapter provides references and information for further reading as well as links to other chapters describing the application of these approaches to the PIRLS 2021 data.

Modern Test Theory: Item Response Theory

The use of item response theory (IRT; Lord & Novick, 1968) has become widespread in educational measurement, due to its flexible framework for estimating proficiency scores from students' responses to test items. PIRLS has been using general IRT models (Lord & Novick, 1968) since its inception in 2001 for the production of proficiency scores. Recent applications of IRT in IEA studies were reviewed by von Davier, Gonzalez, and Schulz (2020) and in TIMSS 2019 by von Davier (2020).

One of the major goals and design principles of PIRLS, but also of other large-scale surveys of student achievement, is to provide valid comparisons across student populations based on broad coverage of the achievement domain. This comprehensive coverage of the achievement domain typically requires hundreds of items in the subject domain. However, given the limited testing time, only a portion of these items can be administered to any one student. To overcome this challenge, PIRLS uses an assessment design based on multi-matrix sampling or incomplete block designs (e.g., Mislevy et al., 1992). As described in PIRLS 2021 Assessment Design (Martin et al., 2019), these achievement items are arranged in blocks that are then assembled into student booklets containing different (but systematically overlapping) sets of item blocks. Because each student receives only a fraction of the achievement items, statistical and psychometric methods are required to link these different booklets together so that student proficiency can be reported on a comparable numerical scale, even though no student sees and answers all items.

IRT is well suited for handling this type of data collection design where not all students are tested on all items. Incomplete block designs can be linked through IRT (e.g., von Davier et al., 2006; von Davier & Sinharay, 2013) and the assumptions can be described and tested formally (e.g., Fischer, 1981; Zermelo, 1929).

The mathematical notation used in this chapter denotes the item response variables on an assessment by x_i, for items $i = 1, \ldots, I$. The set of responses to these items is $(\boldsymbol{x}_v) = (x_{v1}, \ldots, x_{vi})$ for student v. By convention, $x_{vi} = 1$ denotes a correct response and $x_{vi} = 0$ denotes an incorrect response.

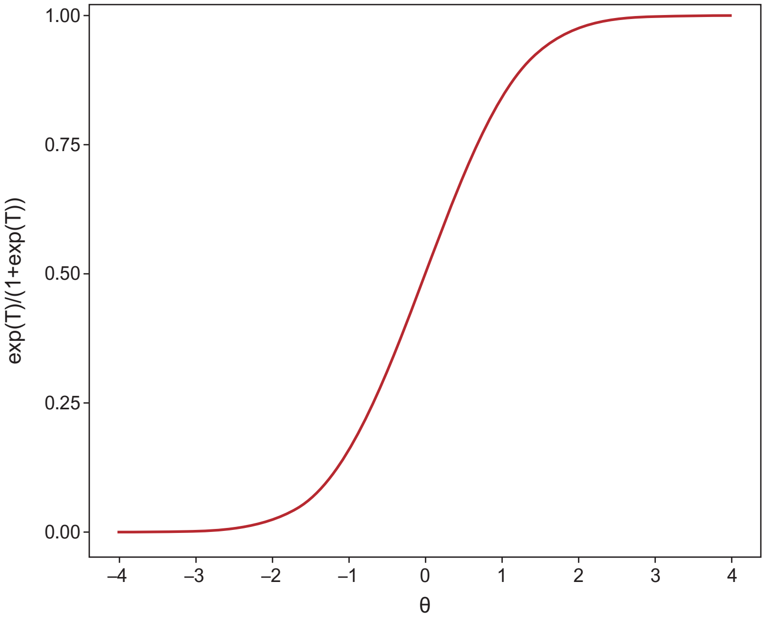
The achievement is assumed to be a function of an underlying latent proficiency variable, in IRT often denoted by θ_v, a real valued variable. Then, we can write

$$P(\boldsymbol{x}_v|\theta_v) = \prod_{i=1}^{I} P(x_{vi}|\theta_v; \boldsymbol{\zeta}_i) \qquad (10.1)$$

where $P(x_{vi} \mid \theta_v \, ; \zeta_i)$ represents the probability of an either correct or incorrect response of a respondent with ability θ_v, to an item with a certain characteristic ζ_i. In IRT, these item-specific effects are referred to as item parameters. Equation (10.1) is a statistical model describing the probability of a set of observed responses given the ability θ_v. This joint probability can be calculated as the product of the individual item probabilities, assuming local independence (described in later section) of responses for a given student ability θ_v.

The item-level probability model, $P(x_{vi} \mid \theta_v \, ; \zeta_i)$, is given by an IRT model that provides a formal mathematical description, an item function, describing how the probability of a correct response depends on the ability and the item parameters. One simple approach for an item function is the inverse of the logistic function, also sometimes called the sigmoid function, depicted in Exhibit 10.1.

Exhibit 10.1: Sigmoid Function of the Rasch Model

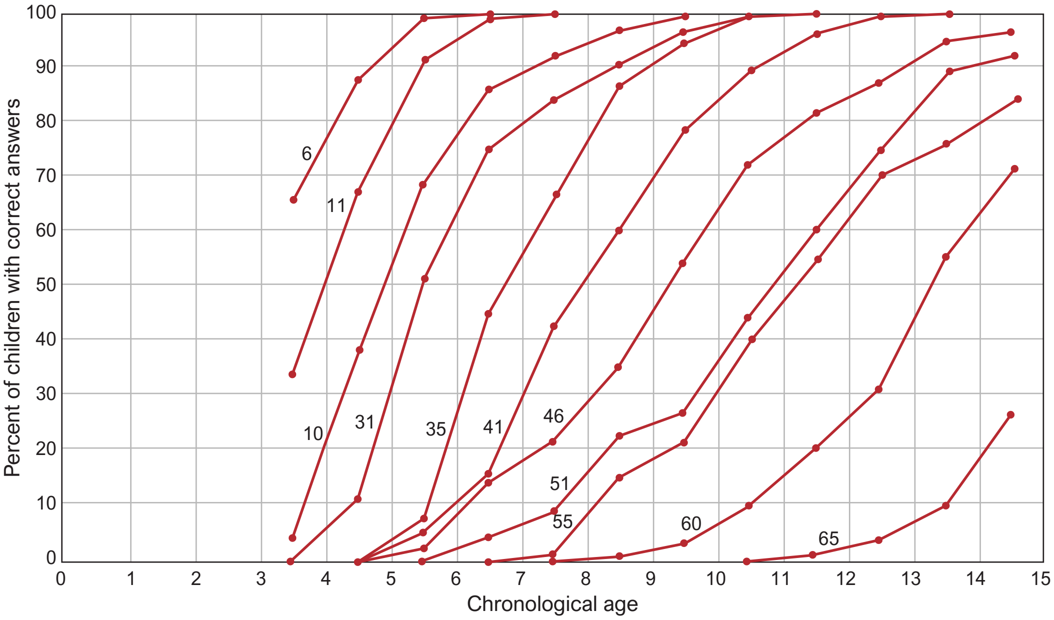


Sigmoid function: $P(x = 1) = exp(T)/[1 + exp(T)]$, where $T = a(\theta - b)$, can be used to linearly adjust for item characteristics.

Many IRT models used in educational measurement can be understood as relatively straightforward generalizations of the item function depicted in Exhibit 10.1. Among many IRT models, the Rasch model (Rasch, 1960; von Davier, 2016) is an important special case where all assessment items contribute equally to the latent construct $(a = 1)$. Why this and other more general approaches of IRT used in PIRLS are suitable choices for modeling assessment data can be seen as follows.

When analyzing test performance by age as an indicator of ability maturation along developmental stages, Thurstone (1925) discovered that the proportion of test-takers who can successfully perform different tasks is monotonically related to their age. This relationship is illustrated in Exhibit 10.2 and bears a strong resemblance to the sigmoidal shape displayed in Exhibit 10.1. Furthermore, when measuring the performance by the total number of correct responses on a longer test, a similar pattern can be observed (Lord, 1980).

Exhibit 10.2: Relationship between Age and Success on Tasks



Trace lines obtained by plotting percent correct against age from a series of tasks (Re-creation of Figure 5 in Thurstone, 1925).

Common parametric functions that fit these non-linear relationships with lower and upper asymptotes of zero and one, respectively, are the probit and logit models (e.g., Cramer, 2003).

While the Rasch model specifies a single item parameter b_i in the form of a negative intercept, more general IRT models can be defined that allow for variation of the trace lines in terms of slopes and asymptotes. PIRLS uses the three-parameter logistic (3PL) IRT model (Lord & Novick, 1968) for selected-response items, the two-parameter logistic (2PL) IRT model for constructed response items worth 1 score point, and the generalized partial credit model (GPCM; Muraki, 1992) for constructed response items worth up to 2 or 3 score points (Yamamoto & Kulick, 2000).

The 3PL IRT model (Birnbaum, 1968) is given by

$$P(x=1|\theta_v;\boldsymbol{\zeta}_i)=c_i+(1-c_i)\frac{exp\left(a_i(\theta_v-b_i)\right)}{1+exp\left(a_i(\theta_v-b_i)\right)} \tag{10.2}$$

and is a popular choice for binary scored selected-response items. In (10.2), c_i denotes the pseudo guessing parameter (which, when set to 0, yields the 2PL for 1-point constructed response items), b_i denotes the item difficulty parameter, and a_i is the slope parameter.

A model frequently used for binary and polytomous ordinal items (items worth up to 3 points in PIRLS) is the GPCM (Muraki, 1992), given by

$$P_i(x|\theta_v) = \frac{exp\left(a_i(x\theta_v - b_{ix})\right)}{1 + \sum_{z=1}^{m_i} exp\left(a_i(z\theta_v - b_{iz})\right)} \qquad (10.3)$$

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

The proficiency variable θ_v is sometimes assumed to be normally distributed, that is, $\theta_v \sim N(\mu,\sigma)$. This can be a useful assumption, but is not a requirement, and may not be an appropriate assumption if a population consists of multiple subpopulations with distinct achievement distributions. In operational scaling applied in national and international large-scale assessments, assuming a common normal distribution for all countries is often not appropriate. Countries differ not only in average and variability of achievement, but also in the shape of their achievement distributions: Student populations may consist of distinct subpopulations, which leads to asymmetric shapes or heavy tails that are not well represented by a normal distribution. The normality constraint is needed for latent regression models (von Davier et al., 2006), but for item calibration it can be relaxed and other types of distributions may be utilized (Haberman et al., 2008; von Davier & Sinharay, 2013; von Davier et al., 2006; von Davier & Yamamoto, 2004; Xu & Jia, 2011; Xu & von Davier, 2008). In PIRLS 2021, the latent distribution was estimated using the empirical histogram method (Bock & Aitkin, 1981; Mislevy, 1984; Woods, 2007) just like it is done in TIMSS, NAEP, PISA, and PIAAC (e.g., von Davier & Sinharay, 2013; Xu & Jia, 2011).

The samples of students that participate in the PIRLS assessment in each cycle come from diverse populations with varying characteristics. Consequently, the calibration procedure needs to account for the possibility of systematic variations in ability distributions from different subpopulations while assuming that the items are comparable across participating countries. To conduct the item calibration, a multiple-group IRT model using country-by-cycle groups is employed. The item parameters are constrained to be equal across groups, with flexibility to allow a unique ability distribution in each country that participated in PIRLS. Minimizing constraints on ability distributions is grounded in best practices used in NAEP, TIMSS, PIRLS, PISA, PIAAC, and other large-scale assessment programs, and the assumption of multiple populations facilitates using the equivalent samples linking approach (described later in this chapter) simultaneously across trend countries. In this approach, respondents from the same digitalPIRLS country assigned to either the digital or the paper-based bridge assessment are sampled from the same population.

When more than one ability is reported, such as reading purpose and comprehension process subscales of overall reading, they are represented in a d-dimensional vector $\boldsymbol{\theta}_v = (\theta_{v1}, \ldots, \theta_{vd})$. In this case, one may assume a multivariate normal distribution, $\boldsymbol{\theta}_v \sim N(\boldsymbol{\mu}, \Sigma)$. For the IRT models used in PIRLS, these d-dimensions are assumed to be measured by separate sets of items, so that

$$\boldsymbol{x}_v = \left((x_{v11}, \ldots, x_{vI_1 1}), \ldots, (x_{v1d}, \ldots, x_{vI_d d}) \right)$$

represents d sets of I_1 to I_d responses, respectively. A d-dimensional version of the model in (10.1) is given by

$$P(\boldsymbol{x}_v | \boldsymbol{\theta}_v) = \prod_{k=1}^{d} \prod_{i=1}^{I_k} P(x_{vik} | \theta_{vk}; \boldsymbol{\zeta}_{ik}) \qquad (10.4)$$

with item-level IRT models (10.2) or (10.3) plugged in for $P(x_{vik} \mid \theta_{vk}; \boldsymbol{\zeta}_{ik})$ as appropriate. The model given in (10.4) is a multidimensional IRT model for items that show between-item multidimensionality (Adams et al., 1997; Adams & Wu, 2007).

Central Assumptions of IRT Models

The IRT modeling approach has several important assumptions that are crucial for making valid inferences in PIRLS and other international large-scale assessments. Meeting these assumptions ensures that proficiency estimates are comparable across participating countries and over time and are generalizable to the broader assessment domains outlined in the assessment framework beyond the limited task sample each student received.

IRT models describe the probability of a correct response, given examinees' proficiency θ and some item-specific parameters (such as a_i, b_i, … described above). However, in actual practice, both proficiency and item parameters are unknown and must be estimated from the data, which is limited to a series of scored answers to a number of assessment items. What is needed, and what IRT provides for PIRLS, is a formal model that applies to an assessment domain as a whole, which is delineated in an assessment framework describing the types of performances on topics viewed as representing the domain. The assumptions underlying IRT support this endeavor by allowing the estimation of proficiency based on performance on assessment tasks within the specified domain in a well-defined and scientifically testable way.

Unidimensionality

PIRLS measures student achievement through a set of items students receive. Let I denote the number of items and let the observed response variables be denoted by $x = (x_1, \ldots, x_I)$. *Unidimensionality* refers to the idea that a single underlying proficiency is measured by all items in an assessment domain, so that the item-level response probabilities can be described by a single quantity, regardless of the specific items a student receives from the entire assessment domain.

Let P_{iv} and P_{jv} denote the probability of person v scoring 1 on items i and j. If the assumption of unidimensionality holds, this can be expressed as

$$P_{iv} = P_i(X = 1 \mid \theta_v)$$

and

$$P_{jv} = P_j(X = 1 \mid \theta_v)$$

with the same real valued θ_v in each expression. Unidimensionality only holds if all the test items are designed to measure the same assessment domain and if test developers follow the assessment framework's content specifications. If the items assess seemingly unrelated skills, such as a gross motor skill and reading comprehension, two proficiency scales may be necessary. But if the domains are closely related, such as reading for literary experience and reading to acquire and use information in PIRLS, it is typically possible to report these appropriately using only one underlying proficiency variable.

Local Independence and Population Independence

The assumption of *population independence* states that the probability of a student answering an item correctly does not depend on their membership to a particular group or demographic. In PIRLS, this assumption is critical for making valid inferences not just across different countries but also within countries for various student groups. Formally, population independence holds if

$$P(X_i = x_i \mid \theta, g) = P(X_i = x_i \mid \theta)$$

for any context variable g. Additionally, this independence also holds for groups defined by performance on x_j, that is, 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 yields

$$P\left(x_i, x_j \mid \theta\right) = P\left(x_i \mid x_j, \theta\right)P\left(x_j \mid \theta\right) = P(x_i \mid \theta)P\left(x_j \mid \theta\right).$$

Based on this independence assumption, the joint probability of observing a series of responses, given an examinee's proficiency level θ, can be written as the product of the individual item-level probabilities. This is known as the *local independence* assumption and takes the form

$$P(\mathbf{X} = x_1, \dots, x_I \mid \theta) = \prod_{i=1}^{I} P_i\,(X = 1 \mid \theta)^{x_i}[1 - P_i(X = 1 \mid \theta)]^{1-x_i}.$$

The concept of local independence appears as a technical assumption, but it can be better understood with the following considerations. The proficiency variable intended to be measured cannot be directly observed, and must instead be inferred from observable responses that are assumed to relate to this variable. The assumptions of population independence and local independence facilitate these inferences by postulating that once a respondent's proficiency level is known, their responses will be independent from each other and from other variables. This means that knowing whether a respondent answered the previous question correctly does not help predicting their next response, provided the respondent's proficiency level θ is known.

According to this assumption, if the model fits the data (and, for example, no learning occurs) and only one proficiency is deemed "responsible" for the probability of giving correct responses, then no other variables, such as language of the assessment, citizenship, gender, or other contextual variables, will play a role in predicting a respondent's answer to the item. The assumptions of local independence and population independence 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.

Monotonicity of Item-Proficiency Regressions

The (strict) *monotonicity* of item functions is a crucial assumption in IRT models used for achievement data. As seen in Exhibit 10.1, the Rasch model as well as the 2PL and 3PL IRT models assume that the probability of a correct response increases with increasing proficiency. This is represented by the inequality

$$P(X_i = 1 \mid \theta_v) < P(X_i = 1 \mid \theta_w) \leftrightarrow \theta_v < \theta_w$$

for all items i. This assumption ensures that proficiency systematically affects the probability of success on the items the students receive. Higher proficiency levels lead to a higher probability of success on each of the items in the achievement domain. This also is reflected in the strict monotonic relationship between the expected achievement scores and proficiency θ:

$$E(S|\theta_v) = \sum_{i=1}^{I} P(X_i = 1 \mid \theta_v) < E(S|\theta_w) = \sum_{i=1}^{I} P(X_i = 1 \mid \theta_w) \leftrightarrow \theta_v < \theta_w. \qquad (10.5)$$

Equation (10.5) shows that a person with a higher skill level θ_w, compared to a person with a lower skill level θ_v, will obtain, in terms of expected score $E(S|\theta_w)$, on average a larger number of correct responses. This monotonicity ensures that the items and test takers are ranked in a systematic manner, where a higher proficiency level is 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.

The foundations for IRT and other latent variable models are based on the aforementioned assumptions. However, it is worth noting that these assumptions can be relaxed to accommodate specific characteristics of the data collection or assessment design. Models that have been described in this chapter are suitable for achievement data, and the same or variations of these models are also used for the analysis of questionnaire data (as described in Chapter 15).

Population-Based Linking Across Digital and Paper PIRLS 2021

This section describes methods for linking the paper-based and the computer-based assessment data utilizing an equivalent samples linking approach. Chapter 12 of this volume presents comparisons of the digitalPIRLS 2021 assessment and the paper-based bridge assessment, focusing on observed item statistics as well as estimates of expected proficiency by country.

The "mode effect" in international assessments has been the subject of investigation by several researchers (e.g., Fishbein et al., 2019; Jerrim, 2018; von Davier et al., 2019). Their findings suggest that any differences observed in how items function across paper-based and computer-based assessments have been generally small or confined to a subset of items. These small differences can be adjusted for by a linear transformation that preserves the properties of measurement invariance at the item level (e.g., Millsap, 2011). However, in order to link assessments and assume invariance of items across modes, it is necessary to identify "link items" that remain invariant across paper-based, computer-based, and previous administrations of the assessment. It should be kept in mind that, any type of linking is a form of constraint imposed on the parameter space (M. von Davier & A. von Davier, 2007). In the case of item invariance linking, if parameters are estimated jointly across mode-based or other treatment-based groups, it takes the form of constrained estimation (e.g., Aitchison & Silvey, 1958; M. von Davier & A. von Davier, 2007) that assumes parameters to be the same across groups. In the case of mode effect models that assume invariant link items, this amounts to a large number of item parameters that are assumed to be the same.

The PIRLS 2021 data collection design was developed to collect comparable samples from the same populations, allowing for the exploration of multiple approaches to linking paper-based and computer-based scales. While a measurement invariance approach using link items was considered for PIRLS 2021, an equivalent samples linking approach ultimately was utilized that required fewer assumptions: for example, the assumption of a set of invariant items that were directly comparable. This population-based linking approach was supported by the design of PIRLS 2021, which incorporates equivalent samples taking the paper-based bridge and the digitalPIRLS assessment, respectively. This randomly equivalent samples approach to linking modes does not assume a common mode effect and allows for each item administered in the new digital mode to have a set of parameters that differ from the paper-based assessment.

In particular, this methodology is made feasible through the selection of an additional, equivalent, paper-based sample in countries that have transitioned to the computer-based assessment. The paper-based sample, called a bridge sample, is often collected from the same schools, but different classrooms, as the digital assessment, enabling the linkage of the two assessments. This design feature facilitates the use of a well-established and widely implemented linking approach, commonly referred to as a "randomly equivalent samples" design (Haberman, 2015; Kim & Lu, 2018; Kolen & Brennan, 2004; Livingston & Kim, 2010; van der Linden & Barrett, 2016; A. von Davier et al., 2004; M. von Davier & A. von Davier, 2007).

The randomly equivalent samples design has been previously utilized in both TIMSS and PIRLS assessments. In PIRLS 2001, a colorized booklet called the PIRLS Reader was linked to the main PIRLS 2001 assessment, which used black and white booklets. This was accomplished with a comparable random sample of students who responded to each type of booklet (Gonzalez, 2003). In TIMSS 2007, a bridge study was conducted due a substantial change to the assessment design to link two different modes of administration (the 2003 design variant and the design used in all other paper based TIMSS cycles). The bridge study relied on a design in which a group of students came from the same population (Foy et al., 2008).

The efficacy of the linking design relies on the assurance that both samples are drawn from the same population. In international assessments, these are two samples from each of the many populations that moved from a paper-based to a computer-based assessment. Of the PIRLS 2021 populations that are included in trend scaling, there are 16 populations that had samples drawn for both the paper-based bridge assessment and the digitalPIRLS assessment. Both samples are presumed to be equivalent in terms of the quality of their sampling design and the quantifiable coverage of their respective populations. This allows the linking of the paper-based and computer-based data on a common scale, albeit not supporting direct comparison of modes at the item level because the items are not assumed to be the same across modes.

The population-based linking approach is particularly relevant for PIRLS 2021, given the need to consider the varying effects of the pandemic on different populations. While most participating countries collected data in the originally targeted testing window, the COVID-19 pandemic necessitated delaying data collection in a number of participating countries, resulting in data collection occurring over a two-year period. In response to school closures and substantial disruptions to normal instruction across participating countries, the pandemic spurred the need to develop online schooling and innumerable activities, including national and local programs launched to use digital tools for remote learning and online schooling (e.g., Barron Rodriguez et al., 2022; Szili et al., 2022). The diverse contexts and approaches taken across education systems to supplement normal schooling during the COVID-19 pandemic likely had differential effects on what was challenging and what was second nature for students as they navigated digital environments.

Because students in the participating countries had differing levels of exposure to digital reading, potentially interacting with time of testing, an approach was needed that did not assume items to be invariant across paper-based assessments in 2016 and 2021 on the one hand and digital assessment in 2021 on the other hand. While it was still assumed that the paper-based items were comparable across 2016 and 2021, the assumption of equivalent items was not tenable for the digital items. Therefore, the PIRLS 2021 linking instead relied on equivalent samples, a central design feature of the digital and bridge data collection.

The population-based linking approach does not remedy all potential confounding factors, such as delayed testing, but it removes one strong invariance assumption while fully capitalizing on the digital and bridge data collection design of PIRLS 2021. Nevertheless, the approach for scaling the PIRLS 2021 data goes beyond what was done, or possible, in previous cycles of PIRLS. The analysis incorporates a multiple population model that provides more robust evaluation of the equivalent samples assumption across the digitalPIRLS countries in PIRLS 2021. Chapter 11 of this volume gives a detailed explanation of the procedures applied to the PIRLS 2021 data, and Chapter 12 provides a comprehensive investigation of the equivalence of digital and bridge samples by examining and comparing the two sampling outcomes and outcomes for key background variables in each country.

Population Models Integrating Achievement Data and Context Information

PIRLS employs a population model to estimate distributions of proficiencies based on the likelihood function of an IRT model, as introduced in the first section of this chapter, and a latent regression of the proficiency on contextual data (e.g., Mislevy, 1984; Mislevy & Sheehan, 1987; von Davier et al., 2006; von Davier et al., 2009). This model is designed to impute the unobserved proficiency distribution, aiming to obtain unbiased group-level proficiency distributions. To achieve this, the model requires estimating an IRT measurement model, which provides information on how responses to assessment items depend on the latent proficiency variable. The latent regression component provides information on how background information is related to achievement and is used to improve estimates by borrowing information through similarities of test takers with respect to contextual variables and the way these relate to achievement. The population model is estimated separately for each country, and in PIRLS 2021 five plausible values (PVs) representing the proficiency variable are drawn from the resulting posterior distribution for each respondent. It should be noted that PVs are not individual test scores and should only be used for analyses at the group level using the procedures described in this report and available, for example, through the IEA IDB Analyzer.

Population models are examples of high-dimensional imputation models that incorporate an extensive set of contextual variables in the latent regression to prevent omission of any essential information gathered in the questionnaires (von Davier et al., 2006; von Davier et al., 2009; von Davier & Sinharay, 2013). Prior to estimating the latent regression model, a principal component analysis is conducted on the student context variables to eliminate collinearity by identifying a smaller number of orthogonal predictors that account for most of the variation in the background variables.

The procedure for estimating proficiency involves combining data from the context questionnaires with the responses obtained from the achievement items. For each individual n, the complete observed data is expressed as $d_n = (x_{n1}, \ldots, x_{nI}, g_n, z_{n1}, \ldots, z_{nB})$, where $z_{n1}, \ldots, z_{nB}$ represent the context information; $x_{n1}, \ldots, x_{nI}$ represent the answers to the achievement items; and g_n represents the country or population the respondent was sampled from.

Proficiency estimation using IRT models can make use of proficiency distributions in the population of interest. By incorporating contextual data, a population model can specify 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

$$\mu_n = \sum_{b=1}^{B} \beta_{g(n)b}\, z_{nb} + \beta_{g(n)0}. \qquad (10.6)$$

This expectation utilizes available information on how context variables relate to proficiency. The proficiency variable is assumed to be normally distributed around this conditional expectation, namely $\theta_n \sim N(\mu_n, \sigma)_.$

Together with the likelihood of the responses expressed by the IRT model, this provides a model for the expected distribution of proficiency given the context data $z_{n1}, \ldots, z_{nB}$ and the responses to the PIRLS items. In simpler terms, the model assumes that the posterior distribution of proficiency depends not only on the observed responses to the PIRLS items but also on the context variables. Given that the amount of contextual data is much larger than the number of countries typically participating in an assessment, the added value of using a model that includes contextual information for every test taker is considerable. Therefore, if background variables are selected so that correlations with proficiency are likely, one obtains a distribution around the expected value given in (10.6) that is noticeably more accurate than a country-level distribution of proficiency.

Formally, this approach can be described as a multiple (latent) regression model that regresses the latent proficiency variable on background data collected in context questionnaires. The estimation of the regression is done separately within countries, as it cannot be assumed that context information has the same regression effects across different participating countries. Mothers' highest level of education, for example, is well known as a strong predictor of student performance, but this association can be moderated by other factors at the level of educational systems, so that in some countries it may be stronger than in others.

Multiple approaches can be used to estimate the latent regression parameters. In large-scale assessments like PIRLS, the latent trait (proficiency) is determined through the IRT models estimated across countries in a previous step. Then the (latent) regression model is estimated treating the mode-specific item parameters from the previous IRT estimations as fixed quantities. This has been implemented using estimated paper and digital item parameters in the respective countries according to which mode of administration they participated. This methodology has been discussed in several studies (e.g., Mislevy, 1991; Thomas, 1993; von Davier et al., 2006; von Davier & Sinharay, 2013).

Group-Level Proficiency Distributions and Plausible Values

The objective of the psychometric methods outlined earlier is to generate a database that provides dependable, valid, and comparable information for reporting student proficiency and for those who use the PIRLS assessment data for secondary analysis. This information takes the shape of plausible values for all respondents based on their responses to the assessment items and their answers to the context questionnaires. Integrating the IRT model described in the first part of this chapter with the regression model introduced in the previous section, we can estimate the probability of the responses, conditional on contextual information, as

$$P_g(\mathbf{x}_n \mid \mathbf{z}_n) = \int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni} \mid \theta)\, \phi\left(\theta; \sum_{b=1}^{B} \beta_{gb}\, z_{nb} + \beta_{g0}, \sigma\right) d\theta\,. \tag{10.7}$$

Equation (10.7) provides the basis for drawing the imputations of proficiency that are commonly known as plausible values (Mislevy, 1991). To allow a more compact notation, we use

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

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

$$P_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n) = \frac{\prod_{i=1}^{I} P_{ig}(x_{ni} \mid \theta) \phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right)}{\int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni} \mid \theta) \phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right) d\theta}.$$

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

$$E_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n) = \int_\theta \theta \, P_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n) d\theta.$$

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

$$V_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n) = E_g(\theta^2 \mid \mathbf{x}_n, \mathbf{z}_n) - \left[E_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n)\right]^2.$$

Estimates of the mean and variance are used to define a posterior proficiency distribution. A set of plausible values (Mislevy, 1991) is then drawn from this distribution for each student. Plausible values are the basis for all reporting of proficiency data in PIRLS, allowing reliable group level comparisons because they are based not only on students' answers to the PIRLS items but also reflect how contextual information is related to achievement.

It should be emphasized that in each country, the correlation between context and proficiency is estimated separately to avoid bias or inaccurate attribution that could have an impact on the results. Although the expected value of country-level proficiency remains the same with or without context information, incorporating such information becomes advantageous when conducting group-level comparisons. Research has shown that including contextual information in a population model substantially reduces potential biases in group level comparisons through both analytical and simulation approaches (von Davier et al., 2009).

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

$$\tilde{\theta}_{ng} \sim N\left(E_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n), \sqrt{V_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n)}\right)$$

that are based on response data x_n as well as context information z_n estimated using a group-specific model for each country g. The inclusion of context information allows for a more accurate estimation of student proficiency and helps to eliminate bias in group-level comparison distributions (e.g., Little & Rubin, 1987; Mislevy 1991; Mislevy & Sheehan, 1987; von Davier et al., 2009). However, it is worth noting that two respondents with the same item responses but different context information will receive a different predicted distribution of their corresponding latent trait. While this may seem counterintuitive, it is important to keep in mind that plausible values are not intended to be used as individual test scores. Rather, they are a tool for producing a useful database of valid and reliable information for reporting aggregated student proficiency and for secondary users of the assessment data.

Linear Transformations of Proficiency Scores

In order to produce the PIRLS 2021 assessment results on the existing PIRLS achievement scale, the 2021 plausible values had to be transformed to the PIRLS reporting metric. This process involved performing a series of linear transformations that are determined using data across all countries contributing to the trend scaling. The first linear transformation is a component of concurrent calibration for paper data which aligns the re-estimated PIRLS 2016 ability distribution with the published PIRLS 2016 ability distribution. Using the first transformation, all paper results including the bridge can be put on the PIRLS trend scale. The second linear transformation aligns the digital ability distribution from the separate digital item calibration with the transformed bridge ability distribution resulting from the paper calibration. This second transformation sets the digital results on the PIRLS trend scale.

These linear transformations are given by

$$PV_{ik}^{*} = A_{ik} + B_{ik} \times PV_{ik}$$

where PV_{ik} is the plausible value i of scale k prior to transformation, PV_{ik}^{*} is the plausible value i of scale k after transformation, and A_{ik} and B_{ik} are the linear transformation constants.

PIRLS 2021 yielded two sets of transformation constants. The first set was obtained for trend scaling and was derived from the published and re-estimated PIRLS 2016 ability distributions. The second set of scale transformation constants for digital scaling was obtained from the transformed bridge and untransformed digital ability distributions.

Transformation constants were obtained by first computing the international means (μ_{ik}) and standard deviations (σ_{ik}) of the plausible values for the overall reading scale using the published 2016 plausible values based on the 2016 item calibration (or using the bridge plausible values based on the 2021 paper calibration). Next, the means (μ_{ik}^{*}) and standard deviations (σ_{ik}^{*}) were calculated using the re-scaled 2016 plausible values based on the 2021 paper calibration model (or using the digital plausible values based on the 2021 digital calibration). From these calculations, the linear transformation constants were defined as:

$$B_{ik} = \sigma_{ik} \Big/ \sigma_{ik}^{*} \tag{10.8}$$

and

$$A_{ik} = \mu_{ik} - B_{ik} \cdot \mu_{ik}^{*} . \tag{10.9}$$

The transformation constants in (10.8) and (10.9) were applied for overall reading as well as for the respective reading purpose and comprehension process subscales. Additional information about the PIRLS 2021 achievement scaling procedures and outcomes is available in Chapter 11.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47-76.

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logic model: A generalized form the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 57-75). New York, NY: Springer Science + Business Media, LLC.

Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *The Annals of Mathematical Statistics, 29*, 813-828. https://doi.org/10.1214/aoms/1177706538

Barron Rodríguez, M., Cobo, C., Munoz-Najar, A., & Sánchez Ciarrusta, I. (2020). *Remote learning during the global school lockdown: Multi-country lessons*. Washington, D.C.: World Bank Group. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/668741627975171644/remotelearning-during-the-global-school-lockdown-multi-country-lessons

Birnbaum, A. (1968). Some latent trait models. 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.

Cramer, C. (2003). *Advanced quantitative data analysis*. New York, NY: McGraw-Hill.

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

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education, 6*(1), 1-23. https://doi.org/10.1186/s40536-018-0064-z

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2007/techreport.html

Gonzalez, E. J. (2003). Scaling the PIRLS reading assessment data. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2001 technical report*. Chestnut Hill, MA: Boston College. https://timssandpirls.bc.edu/pirls2001i/PIRLS2001_Pubs_TR.html

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40*(3), 254–273. https://doi.org/10.3102/1076998615574772

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.

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it?. *Oxford Review of Education, 44*(4), 476-493. https://doi.org/10.1080/03054985.2018.1430025

Kim, S., & Lu, R. (2018). The pseudo-equivalent groups approach as an alternative to common-item equating. ETS Research Report Series, 2018: 1-13. https://doi.org/10.1002/ets2.12195

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer Science+Business Media.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: J. Wiley & Sons.

Livingston, S. A., & Kim, S. (2010). Random-groups equating with samples of 50 to 400 test takers. *Journal of Educational Measurement, 47*, 175–185.

Lord, F. M. (1980). *Applications of items response theory to practical testing problems*. Hillsdales, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Martin, M. O., von Davier, M., Foy, P., & Mullis, I. V. S. (2019). PIRLS 2021 assessment design. In I. V. S. Mullis & M.O. Martin (Eds.), *PIRLS 2021 assessment frameworks*. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2021/frameworks/

Millsap, R. E. (2011). Statistical approaches to measurement invariance. New York, NY: Routledge/Taylor & Francis Group.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359-381.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*, 177-196.

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.

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.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–176.

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

Szili, K., Kiss, R., Csapó, B., & Molnár, G. (2022). Computer-based development of reading skills to reduce dropout in uncertain times. *Journal of Intelligence, 10*(4), 89. http://dx.doi.org/10.3390/jintelligence10040089

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics, 2*, 309 322.

Thurstone, L. L. (1925). A method of psychological and educational tests. *Journal of Educational Psychology, 16*(7), 433-451. https://doi.org/10.1037/h0073357 van der Linden, W. J., & Barrett, M. D. (2016). Linking item response model parameters. *Psychometrika, 81*, 650–673. https://doi.org/10.1007/s11336-015-9469-6

von Davier, A. A., Holland, P.W., & Thayer, D.T. (2004). *The kernel method of test equating*. New York: Springer-Verlag.

von Davier, M. (2016). The Rasch model. In W. J. van der Linden (Ed.), *Handbook of item response theory* (2nd ed., Vol. 1, pp. 31-48). Boca Raton, FL: CRC Press.

von Davier, M. (2020). TIMSS 2019 scaling methodology: Item response theory, population models, and linking across modes. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 11.1-11.25). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-11.html

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., Gonzalez, E., & Schulz, W. (2020). Ensuring validity in international comparisons using state-of-the-art psychometric methodologies. In H. Wagemaker (Ed.), *Reliability and validity of international large-scale assessments* (Vol. 10, pp. 187 219). International Association for the evaluation of Educational Achievement. https://doi.org/10.1007/978-3-030-53081-5_11

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*(6), 671–705.

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.

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.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences, 3*(3), 115-124.

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

Woods, C. M. (2007). Empirical histograms in item response theory with ordinal data. *Educational and Psychological Measurement, 67*(1), 73-87.

Xu, X., & Jia, Y. (2011). *The sensitivity of parameter estimates to the latent ability distribution*. ETS Research Report Series, 2011: i-17. https://doi.org/10.1002/j.2333-8504.2011.tb02276.x

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.

Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report*. Chestnut Hill, MA: Boston College. https://timssandpirls.bc.edu/timss1999i/tech_report.html

Zermelo, E. (1929). The calculation of tournament results as a maximum-likelihood problem [German]. *Mathematische Zeitschrift, 29*, 436-460.