CHAPTER 11

TIMSS Achievement Scaling Methodology: Item Response Theory and Population Models

Ummugul Bezirhan
Matthias von Davier

Introduction

This chapter describes the statistical and psychometric methods used to analyze the achievement data collected in TIMSS. The chapter begins by introducing item response theory (IRT), a widely-used methodology in educational measurement and psychometrics for analyzing item response data. IRT provides a model-based foundation for test development, item analysis, scale linking, computerized adaptive testing, and many other applications. Moreover, its utility can be extended to analyzing a wide range of human-response data, such as patient feedback, psychological evaluation, consumer choice, and data from other disciplines.

The second part of the chapter discusses the integration of achievement data from the TIMSS mathematics and science items with contextual data from student questionnaires (and parent questionnaires at the fourth grade) and describes the statistical imputation model used for this purpose, following essentially the approach used by TIMSS since its inception in 1995. 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. Then, plausible values of student achievement are drawn from the resulting proficiency distribution for use in the analysis. It is important to emphasize that plausible values are not intended to be used as individual test scores. Instead, 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.

This chapter provides references and information for further reading. Chapter 12 describes the application of the methods described here to TIMSS data for the 2023 assessment cycle.

Modern Test Theory: Item Response Theory

IRT, originally described by Lord and Novick (1968), has become widespread in educational measurement due to its flexible framework for estimating proficiency scores from students' responses to test items. Since its inception in 1995, TIMSS has implemented IRT, initially employing the Rasch model (Rasch, 1960; Adams, Wu, & Macaskill, 1997; von Davier, 2016), but later moving to more general IRT models (Lord & Novick, 1968; Yamamoto & Kulick, 2000) for the estimation of item parameters and proficiency scores. A comprehensive overview of recent modeling approaches and the application of IRT in IEA studies was given by von Davier et al. (2020).

One of the major goals and design principles of TIMSS, as well as other large-scale surveys of student achievement, is to provide valid comparisons across student populations based on broad coverage of the achievement domain. This typically requires ensuring a comprehensive coverage of the achievement domain through hundreds of items in the subject. However, given the limited testing time, only a portion of these items can be administered to any one student. In mathematics as well as in science, this translates into an assessment containing several hundred achievement items, while only a fraction can be administered to any one student given the available testing time (36 minutes per subject at fourth grade, 45 minutes per subject at eighth grade). To overcome this challenge, TIMSS uses an assessment design based on multi-matrix sampling or incomplete block designs (e.g., Mislevy et al. 1992). As described in TIMSS 2023 Assessment Design, these achievement items are arranged in blocks that are then assembled into student booklets that contain 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 answers all of the assessment items.

IRT is well suited for handling this type of data collection design where not all students are tested on all items. Data collected with item blocks presented using incomplete block designs can be linked through IRT (e.g., von Davier et al., 2006; von Davier & Sinharay, 2013) and the assumption needed to enable making many test forms comparable can be described and tested formally (e.g., Fischer, 1981; Zermelo, 1929).

The mathematical notation in this chapter represents the item response variables on an assessment as x_i, where the $i = 1, \ldots, I$ denote the item index. The set of responses to these items is $(\boldsymbol{x}_v) = (x_{v1}, \ldots, x_{vi})$ for student v. By convention, in the case of dichotomously scored items worth one score point, we assume $x_{vi} = 1$ 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, often in IRT denoted by θ_v, a real-valued variable. Then, the probability of the observed responses for a test taker of that proficiency is modeled as

$$P(\boldsymbol{x}_v|\theta_v) = \prod_{i=1}^{I} P(x_{vi}|\theta_v;\boldsymbol{\zeta}_i) \tag{11.1}$$

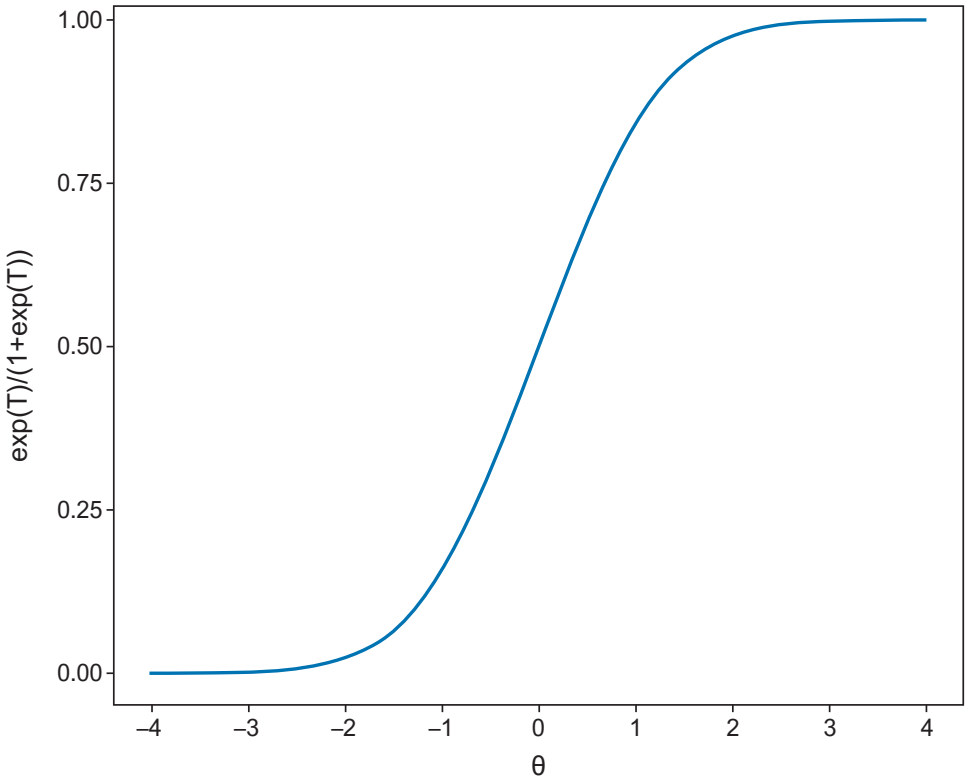
where $P(x_{vi} \mid \theta_v ; \boldsymbol{\zeta}_i)$ represents the probability of either correct or incorrect responses of a respondent with ability θ_v to an item i with item parameters $\boldsymbol{\zeta}_i$. To allow a more compact notation, we use

$$P(x_{vi} \mid \theta_v; \boldsymbol{\zeta}_i) = P(X = 1 \mid \theta_v; \zeta_i)^{x_{vi}}[1 - P(X = 1 \mid \theta_v; \zeta_i)]^{1-x_{vi}}.$$

Equation (11.1) is a statistical model describing the probability of a set of observed responses $(\boldsymbol{x}_v) = (x_{v1}, \ldots, x_{vi})$ as independent, conditional on the ability θ_v. This joint probability can be calculated as the product of the individual item probabilities, assuming local independence (described in a later section) of responses for a given student ability θ_v as the responses are assumed to depend only on a test taker's proficiency, and no other variables.

The item-level probability model, $P(x_{vi} \mid \theta_v ; \boldsymbol{\zeta}_i)$, 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 example of an item function is the inverse of the logistic function depicted in Exhibit 11.1.

Exhibit 11.1: The Logistic Function, a Fundamental Building Block of IRT



The function depicted in Exhibit 11.1 is given by $f(\mathrm{T}) = exp(T)/[1 + exp(T)] = logit^{-1}(T)$. With $\boldsymbol{\zeta}_i=(a_i,b_i)$ and $T_{vi} = a_i(\theta_v - b_i)$ we can define

$$P(x_{vi} = 1 \mid \theta_v ; \boldsymbol{\zeta}_i) = logit^{-1}(a_i(\theta_v - b^i)).$$

Then, it can be shown that

$$P(x_{vi}|\theta_v, \zeta_i) = \frac{exp(x_{vi}[a_i(\theta_v - b_i)])}{1 + exp\big(a_i(\theta_v - b_i)\big)}$$

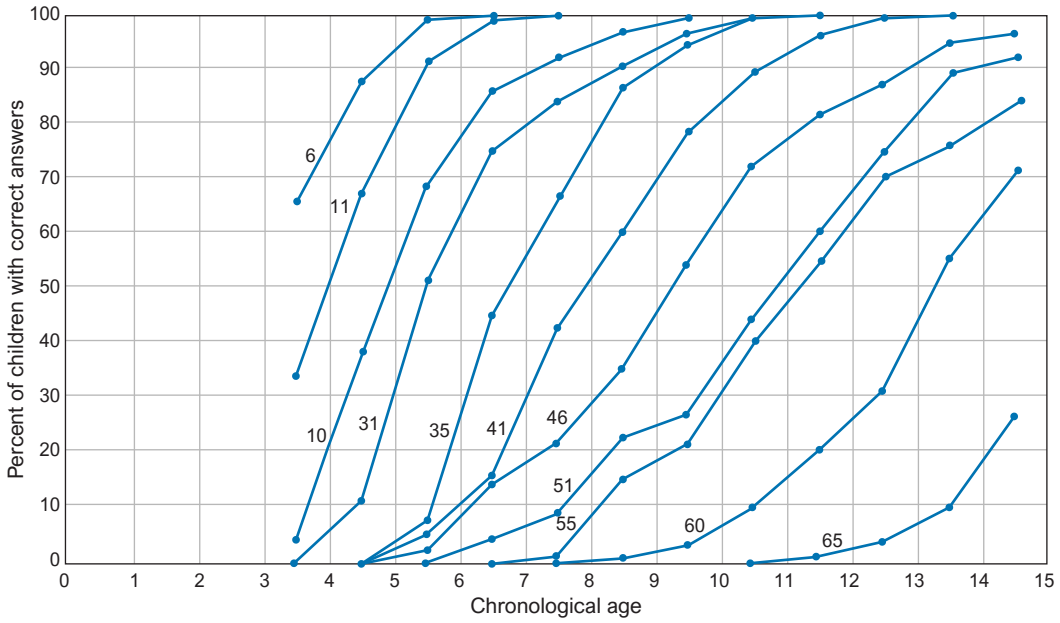
which is the model equation of the two-parameter logistic (2PL) IRT model. The two parameters are the a_i, parameters that quantify the slope of the item function, and the b_i, which parameterize the position of the item function relative to the person proficiency parameter θ_v. The 2PL IRT model is a very common model for analyzing binary response data, and it is one of the models used in TIMSS, as explained below.

Many IRT models used in educational measurement can be understood as relatively straightforward variants of the item function depicted in Exhibit 11.1. Among the most popular IRT models, the Rasch model (Rasch, 1960; von Davier, 2016) is an important special case, also called the one-parameter logistic (1PL) IRT model, where all assessment items are considered to contribute equally to the latent construct. In the Rasch model, all item slope parameters are assumed to be the same, so there is only one parameter per item, the b_i. Why this and other more general approaches of IRT used in TIMSS are suitable choices for modeling assessment data can be seen as follows:

Thurstone (1925) discovered that the proportion of test-takers who can successfully perform different tasks is monotonically related to their age when analyzing test performance by age as an indicator of ability maturation along developmental stages. This relationship is illustrated in Exhibit 11.2 and closely resembles the inverse logistic function displayed in Exhibit 11.1. Furthermore, a similar pattern can be observed when measuring the performance by the total number of correct responses on a longer test (Lord, 1980). This led to the choice of the inverse logistic function as the basis for the item functions in IRT.

The probit and logit models are common parametric functions that fit these non-linear relationships with lower and upper asymptotes of zero and one, respectively (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.

Exhibit 11.2: Relationship between Age and Success on Tasks that Inspired IRT Development



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

TIMSS generally employs the two-parameter logistic (2PL) IRT model for items worth one score point and the generalized partial credit model (GPCM; Muraki, 1992) for items worth up to two score points. The three-parameter logistic (3PL) IRT model (Lord & Novick, 1968) is used for all multiple-choice items (or "single-selection" items) introduced in TIMSS 2019 or earlier cycles and continues to be applied for all multiple-choice items that cannot be fitted with the 2PL model.

The 3PL model adds a third parameter to the item function, which acts as a lower asymptote. This lower asymptote is denoted by c_i and quantifies the theoretical probability of a correct response for respondents with the lowest possible proficiency levels.

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

$$P(x_{vi} = 1|\theta_v; \boldsymbol{\zeta}_i) = c_i + (1 - c_i)\frac{exp\left(a_i(\theta_v - b_i)\right)}{1 + exp\left(a_i(\theta_v - b_i)\right)} \tag{11.2}$$

where c_i denotes the pseudo guessing parameter—which, when set to 0, yields the 2PL, as before, b_i denotes the item difficulty parameter, and a_i is the slope parameter.

A model frequently used for polytomous ordinal items (items worth up to two points) is the GPCM (Muraki, 1992), given by

$$P_i(X = x|\theta_v) = \frac{exp\left(a_i(x\theta_v - b_{ix})\right)}{1 + \sum_{z=1}^{m_i} exp\left(a_i(z\theta_v - b_{iz})\right)} \tag{11.3}$$

assuming a response variable with $m_i + 1$ ordered categories. The threshold parameters are often split into a location and normalized step parameters, $b_{ix} = \delta_i - \tau_{ix}$, with $\sum \tau_{ix} = 0$ (e.g. Muraki, 1992).

The proficiency variable θ_v is sometimes assumed to be normally distributed, that is, $\theta_v \sim N(\mu, \sigma)$ for convenience. 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 joint normal distribution for all countries is often inappropriate. 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 used (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 TIMSS, the latent distribution is estimated using the empirical histogram method (Bock & Aitkin, 1981; Mislevy, 1984; Woods, 2007), just like it is done in other IEA studies including PIRLS, as well as NAEP, PISA, and PIAAC (e.g., von Davier & Sinharay, 2013; Xu & Jia, 2011).

The samples of students who participate in each cycle of TIMSS come from diverse populations with varying characteristics. Consequently, the calibration procedure must account for the possibility of systematic variations in ability distributions from different subpopulations while assuming that the items are comparable across participating countries. A multiple-group IRT model using country groups is employed to conduct the item calibration. The item parameters are constrained to be equal across groups, with flexibility to allow a unique ability distribution in each country. Minimizing constraints on ability distributions is grounded in best practices used in large-scale assessment programs.

When more than one ability is reported, for example, mathematics and science, or content and cognitive subscales of these overall domains, 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 TIMSS, these d-dimensions are assumed to be measured by non-overlapping sets of items so that

$$\boldsymbol{x}_v = \left((x_{v11}, \ldots, x_{vI_11}), \ldots, (x_{v1d}, \ldots, x_{vI_dd}) \right)$$

represents *d* sets of I_1 to I_d responses, respectively. A d-dimensional version of the model in (11.1) is given by

$$P(\boldsymbol{x}_v | \boldsymbol{\theta}_v) = \prod_{k=1}^{d} \prod_{i=1}^{I_k} P(x_{vik} | \theta_{vk}; \boldsymbol{\zeta}_{ik}) \qquad (11.4)$$

with item-level IRT models (11.2) or (11.3) plugged in for $P(x_{vik} \mid \theta_{vk}; \boldsymbol{\zeta}_{ik})$ as appropriate. The model given in (11.4) is a multidimensional IRT model for items that each measure one dimension, but across subtests show between-item multidimensionality (Adams, Wilson, & Wu, 1997; Adams & Wu, 2007).

Central Assumptions of IRT Models

As a mathematical model, IRT depends on a set of assumptions about the data to which the model applies. These assumptions specify the relationships between observable and non-observable (latent) constructs within the model, establishing the foundational conditions for its application. Meeting these assumptions ensures that proficiency estimates are comparable across different assessment instruments and participating countries over time and are generalizable to the broader domains outlined in the assessment frameworks beyond the limited tasks each student received.

IRT models describe the probability of a correct response, given examinees' proficiency θ and several item-specific parameters such as the discrimination (a), difficulty (b), and other characteristics described above. However, proficiency and item parameters are unknown in actual practice and must be estimated from the data, which typically consists of scored responses to a limited number of assessment items.

For large-scale assessments like TIMSS, IRT provides a structured model that applies to the entire assessment domain, delineated in assessment frameworks that describe the types of performances on topics viewed as representing the domain. The underlying assumptions of IRT support this endeavor by allowing for the estimation of proficiency levels on assessment tasks within the specified domain in a well-defined and scientifically testable way.

Unidimensionality

TIMSS 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 probabilities of responses to each item 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 any two of the items, for example, items i and j. If the assumption of unidimensionality holds, there is a single real-valued available θ_v with

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

and

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

for any pair of items i, j on the test. These identities imply that the response probabilities only depend on the person's ability θ_v, which can be expressed as a single real-valued variable, and no other characteristics of a person. However, unidimensionality only holds if the test items are designed to assess the same assessment domain and if test developers follow the assessment framework's content specifications. If the items assess seemingly unrelated skills, such as gross motor skills, reading, or science in one set of items, and mathematics in another set of items, two or more proficiency scales may be necessary. However, a unidimensional proficiency may suffice if the domains are closely related and require knowledge, for example, of different subdomains within overall mathematics, such as algebra and geometry.

Population Independence and Local Independence

The assumption of *population independence* states that the likelihood of a student answering an item correctly does not depend on their membership to a particular group or demographic. In TIMSS, this assumption is critical for making valid inferences across different countries and 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 contextual variable g. Additionally, this independence 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 also be considered a grouping variable, as it splits the sample into those that produced a correct response and those that did not, in the simplest case. Applying the assumption of population independence yields

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

Based on this population independence assumption, the joint probability of observing a series of responses, given the examinees' 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

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

Local independence is a technical assumption, but it can be better understood when considering the following: 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 and local independence facilitate these inferences by postulating that once a respondent's proficiency level is known, their responses will be independent of each other and from other variables. That is, knowing whether a respondent answered the previous question correctly does not help predict their next response, provided the respondent's proficiency level θ is known.

According to this assumption, if the model fits the data and only one proficiency is deemed "responsible" for the probability of giving correct responses, then no other variables, such as the language of the assessment or those related to other student attributes, will play a role in predicting a respondent's answer to the next item. The assumptions of local and population independence encapsulate the goal that only one variable 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 the achievement data. As shown in Exhibit 11.1, the Rasch and the 2PL and GPCM 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 affects the probability of success on the items the students receive, whereas higher proficiency levels lead to a higher probability of success on each item in the achievement domain. This is also reflected in the strict monotonic relationship between the expected achievement scores and proficiency θ:

$$E(S|\theta_v) = \sum_{i=1}^{I} P(X_i = 1 \mid \theta_v) < E(S|\theta_w) = \sum_{i=1}^{I} P(X_i = 1 \mid \theta_w) \leftrightarrow \theta_v < \theta_w. \quad (11.5)$$

Equation (11.5) 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 systematically, 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 (e.g., Thissen et al., 1989). Models that have been described in this chapter are suitable for achievement data, and the same or variations of these models are used for the analysis of questionnaire data (as described in Chapter 15).

Population Models Integrating Achievement Data and Context Information

TIMSS employs a population model to estimate distributions of proficiencies based on the likelihood function of an IRT model, as introduced in the previous 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 the estimation of 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. In the case of TIMSS, multiple imputations–five plausible values (PVs)–representing the proficiency variable are drawn from the resulting posterior distribution for each respondent in each domain. It should be noted that PVs are not individual test scores in the traditional sense. They 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 high-dimensional imputation models that incorporate an extensive set of contextual variables in the latent regression to ensure the inclusion of any essential information collected with the context questionnaires (von Davier et al., 2006; von Davier et al., 2009; von Davier & Sinharay, 2013). Before estimating the latent regression model, a principal component analysis (PCA) is conducted on the student context variables to create orthogonal variables and, therefore, eliminate collinearity and then identify a smaller number of principal components that account for most of the variation. To avoid overspecification of the conditioning model, TIMSS selects principal components for each country such that 90% of the common variance is explained or the number of components is no more than 5% of the unweighted student sample size, whichever leads to fewer principal components.

Estimating proficiency involves combining data from the context questionnaires with the responses obtained from the achievement items. For each respondent n, the complete observed data is expressed as $d_n = (x_{n1}, \ldots, x_{nI}, g_n, z_{n1}, \ldots, z_{nB})$, where $z_{n1}, \ldots, z_{nB}$ represents the context information in the form of principal components; the $x_{n1}, \ldots, x_{nI}$ 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 use 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

$$\mu_n = \sum_{b=1}^{B} \beta_{g(n)b}\, z_{nb} + \beta_{g(n)0}. \tag{11.6}$$

This expectation utilizes available information on how context variables relate to proficiency. The distribution of the proficiency variable is assumed to be normal 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 achievement items. In simpler terms, the model assumes that the posterior distribution of proficiency depends on the observed responses to the achievement items as well as the context variables. Given 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 context variables are selected so that correlations with proficiency are likely, one obtains a distribution around the expected value (11.6) that is noticeably more accurate than a country-level distribution of proficiency.

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 regression estimation is done separately for each country, as context information cannot be assumed to have the same regression effects across participating countries. Parents' highest level of education, for example, is well known as a strong predictor of student performance. Still, this association can be moderated by other factors at the level of educational systems, so 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 TIMSS, 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 item parameters from the previous IRT estimations as fixed quantities. Several chapters and articles have discussed this methodology in detail (e.g., Mislevy & Sheehan, 1987; Thomas, 1993; von Davier et al., 2006; von Davier & Sinharay, 2013).

Group-Level Proficiency Distributions and Plausible Values

The psychometric methods outlined earlier aim to generate a database that provides reliable and comparable information for reporting student proficiency and for those who use the TIMSS assessment data for secondary analysis. This information takes the shape of proficiency estimates in the form 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 above, we can estimate the probability of the responses, conditional on context information, as

$$P_g(\mathbf{x}_n \mid \mathbf{z}_n) = \int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni} \mid \theta)\, \phi\left(\theta; \sum_{b=1}^{B} \beta_{gb}\, z_{nb} + \beta_{g0}, \sigma\right) d\theta. \tag{11.7}$$

Equation (11.7) provides the basis for drawing the imputations of proficiency commonly known as plausible values (Mislevy, 1991).

The model given in (11.7) enables inferences about the posterior distribution of the proficiency θ, given both the 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 is then drawn from this distribution for each student. Plausible values are the basis for all reporting of proficiency data in TIMSS, allowing reliable group-level comparisons because they are based not only on students' answers to the TIMSS items but also reflect how contextual information is related to achievement.

It should be emphasized that in each country, the correlation between contextual information and proficiency is estimated separately to avoid bias or inaccurate attribution that could have an impact on the results. Although the expected value of average 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).

The plausible values used in TIMSS 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 and context information z_n estimated using a group-specific model for each country g. Including context information allows a more accurate estimation of student proficiency and helps eliminate bias in group-level comparisons for those grouping variables included in the model (e.g. Little & Rubin, 1987; Mislevy 1991; Mislevy & Sheehan, 1987; von Davier et al., 2009).

One consequence of this approach is worth noting: Two respondents with the same item responses but different context information will likely receive a different predicted distribution of their corresponding latent trait and, thus, a different set of imputed proficiency estimates. 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. They facilitate group-level comparisons, which is the main goal of internationally and nationally comparable student surveys.

Linear Transformations of Proficiency Scores

To produce the TIMSS assessment results for a given cycle on the existing TIMSS achievement scale, the plausible values need to be transformed onto the TIMSS reporting metric. This process involves performing a series of linear transformations determined using data across all trend countries contributing to the scaling, which incorporates data from the current cycle and the data from the previous cycle (for example, TIMSS 2023 and TIMSS 2019). The first linear transformation is a component of concurrent calibration, which aligns the re-estimated ability distribution of the previous TIMSS cycle with the published ability distribution of the previous TIMSS cycle. Using this transformation, all results for a given TIMSS cycle can be put on the TIMSS 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 before transformation, PV^{*}_{ik} is the plausible value i of scale k after transformation, and A_{ik} and B_{ik} are the linear transformation constants.

Transformation constants are obtained by first computing the international means (μ_{ik}) and standard deviations (σ_{ik}) of the plausible values for the overall mathematics and science scales using the published plausible values of the previous cycle based on the previous cycle item calibration. Next, the means (μ^{*}_{ik}) and standard deviations (σ^{*}_{ik}) are calculated using the rescaled plausible values of the previous cycle based on the current cycle calibration model. From these calculations, the linear transformation constants are defined as:

$$B_{ik} = \frac{\sigma_{ik}}{\sigma^{*}_{ik}} \tag{11.8}$$

and

$$A_{ik} = \mu_{ik} - B_{ik} \cdot \mu^{*}_{ik}. \tag{11.9}$$

The transformation constants in (11.8) and (11.9) are applied separately for overall mathematics and science at each grade. The same transformations used for the overall subjects are applied for their respective content and cognitive subscales.

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.

Adams, R. J., Wu, M. L., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematics and science scales. In M. O. Martin & D. L. Kelly (Eds.), *TIMSS Technical Report Volume II: Implementation and Analysis (Primary and Middle School Years)* (pp. 111-145). Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss1995i/TechVol2.html

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

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.

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. Retrieved from http://dx.doi.org/10.1007/BF02293919

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* (Research Report No. RR-08-45). Princeton, NJ: Educational Testing Service.

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

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.

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).

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247-260.

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

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. (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., 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). https://ierinstitute.org/fileadmin/Documents/IERI_Monograph/Volume_2/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., & 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., & 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

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

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* (pp. 237-263). Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. https://timss.bc.edu/timss1999i/tech_report.html

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.

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