CHAPTER 11

TIMSS 2019 Scaling Methodology: Item Response Theory, Population Models, and Linking Across Modes

Matthias von Davier

Introduction

This chapter[1] describes the statistical and psychometric approaches underlying the analysis of the TIMSS 2019 data. The first part of the chapter reviews Item Response Theory (IRT), a methodology frequently used in educational measurement that is also increasingly 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 TIMSS 2019, half of the countries administered the computer-based version of TIMSS (known as eTIMSS) while the other half continued to assess the students using the paper-based version (paperTIMSS).

The second part of the chapter describes an extension of IRT that allows controlling for mode of administration effects on student performance and that produces a latent variable scale representing student proficiency that is comparable across paper- and computer-based assessment.

The third part of this chapter reviews the integration of achievement data from the TIMSS 2019 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. 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 TIMSS since the first assessment in 1995. All three parts provide references and information for further reading as well as information about where in other chapters of this volume these developments are being described in terms of actual application to TIMSS 2019 data.

1 The writeup of the psychometric methods presented in this chapter has many sources and the models presented here were developed by a variety of authors. The presentation as compiled here is focused on TIMSS 2019 and benefited greatly from conversations with, and reviews and proofreading by Michael O. Martin, Pierre Foy, Bethany Fishbein, and Liqun Yin.

Modern Test Theory: Item Response Theory

Item Response Theory (IRT; Lord & Novick, 1968) has become one of the most important tools of educational measurement as it provides a flexible framework for estimating proficiency scores from students' responses to test items. A Google search for the phrase "Item Response Theory" (IRT) produces 1,740,000 hits as of September 15, 2020.

TIMSS has been using IRT from the first round in 1995, initially in the form of the Rasch IRT model (Rasch, 1960; von Davier, 2016) and started to use more general IRT models (Lord & Novick, 1968) for the production of proficiency scores beginning with the 1999 cycle. An overview of recent applications of IRT in IEA studies was given by von Davier, Gonzalez, and Schulz (2020).

One of the major goals and design principles of TIMSS, but also other large-scale surveys of student achievement, is to provide valid comparisons across student populations based on broad coverage of the achievement domain. In mathematics as well as in science, this translates into several hundred achievement items, only a fraction of which can be administered to any one student given the available testing time (72 minutes at fourth grade, 90 minutes at eighth grade). Therefore, TIMSS uses an assessment design based on multi-matrix sampling or balanced incomplete block designs (e.g., Mislevy, Beaton, Kaplan, & Sheehan, 1992). As described in the TIMSS 2019 Assessment Design (Martin, Mullis, & Foy, 2017), these achievement items are arranged in blocks that are then assembled into student booklets (or booklet equivalents for eTIMSS) 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 sees and answers all tasks.

IRT is particularly well suited to handle such data collection design in which not all students are tested with all items. The assumptions made for enabling IRT methods to handle these types of designs, commonly known as balanced incomplete block designs (e.g. von Davier, Sinharay, Oranje & Beaton, 2006; von Davier & Sinharay, 2013) can be described and tested formally (e.g. Fischer, 1981; Zermelo, 1929).

In terms of mathematical notation used in this chapter, the item response variables on an assessment are denoted 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. For simplicity, we assume $x_{vi} = 1$ denotes a correct response and $x_{vi} = 0$ denotes an incorrect response.

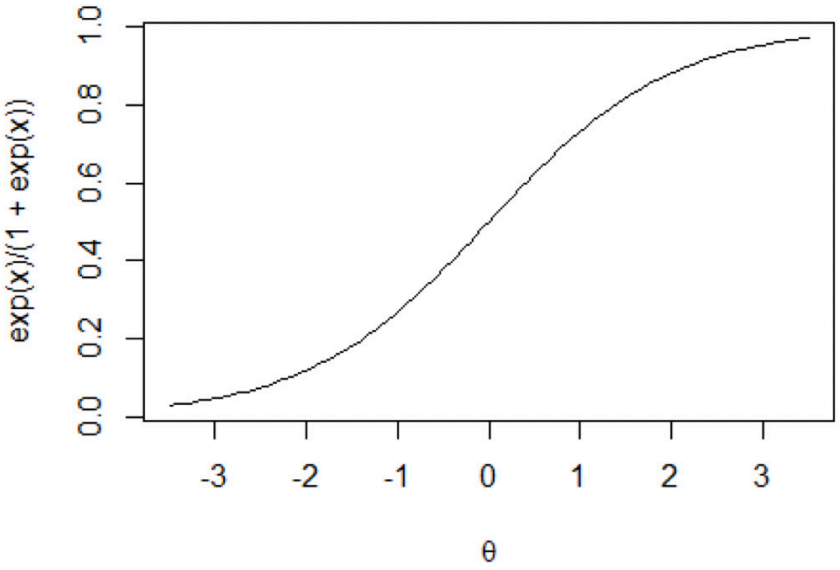
The achievement is assumed to be a function of an underlying latent proficiency variable, often in IRT denoted by θ_v, a real valued variable. Then, we can write

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

where $P(x_{vi} \mid \theta_v;\boldsymbol{\zeta}_i)$ represents the probability of an either correct or incorrect response of a respondent with ability θ_v and an item with a certain characteristic $\boldsymbol{\zeta}_i$. In IRT, these item specific effects are referred to as item parameters. Equation (11.1) is a statistical model describing the probability of a set of observed response given ability θ_v. This collective probability is the product of the individual item probabilities.

In TIMSS, 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, that describes 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 11.1.

Exhibit 11.1: Sigmoid Function of the Rasch Model

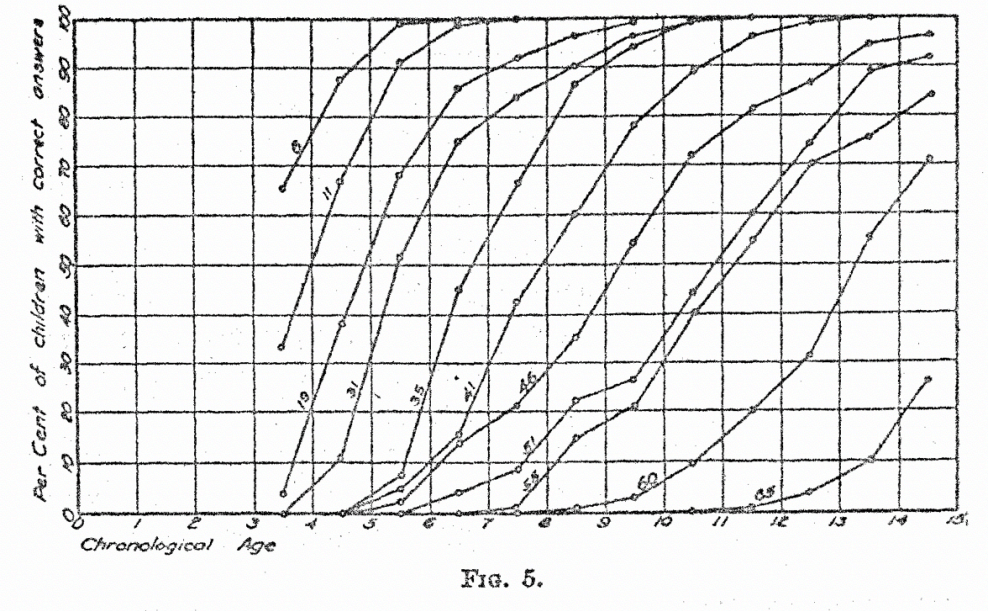


Sigmoid function of the Rasch model $P(x = 1) = exp(T)/[1 + exp(T)]$, *where* $T = a(\theta - b)$ *can be used to linearly adjust for item characteristics.*

Many IRT models used in educational measurement can be understood as relatively straightforward generalizations of the approach shown in Exhibit 11.1. For $a = 1$, where all assessment items contribute equally to the latent construct, this model is called the Rasch model (Rasch, 1960; von Davier, 2016). Why this and other more general approaches of IRT used in TIMSS are suitable choices for modeling assessment data can be seen in the following example.

When looking at test performance by age (a proxy of ability maturation along developmental stages), Thurstone (1925) found that the proportion of respondents who successfully master different tasks is monotonically related to age. Exhibit 11.2 shows this relationship.

Exhibit 11.2: Relationship between Age and Success on Tasks



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

The similarity to the sigmoid shown in Exhibit 11.1 is obvious. When, instead of developmental age, the total number of correct responses on a longer test is used, similar graphs are obtained (Lord, 1980). Natural choices for a parametric function that can fit these types of non-linear relationships with a lower and an upper asymptote of zero and one, respectively, are the probit and the logit (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. TIMSS used the Rasch model in 1995, and since 1999 uses the three-parameter logistic (3PL) IRT model (Lord & Novick, 1968) for multiple-choice items, the 2PL IRT model for constructed response items worth 1 score point, and the generalized partial credit model (Muraki, 1992) for constructed response items worth up to 2 score points (Yamamoto & Kulick, 2000).

The 3PL IRT model is given by

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

and is a popular choice for binary scored multiple-choice items. In equation (11.2), c_i denotes the pseudo guessing parameter—which, when set to 0.0, 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 2 points in TIMSS) is the generalized partial credit model (Muraki, 1992), given by

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

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

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

When there is more than one ability, for example mathematics and science, or content and cognitive process subscales of these, these 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, examples are main domains or subscales, are assumed to be measured by separate 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

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

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 show between-item multidimensionality (Adams, Wilson, & Wu, 1997; Adams & Wu, 2007).

Central Assumptions of IRT Models

This section reviews important assumptions of the IRT modeling approach that are central to the types of inferences to be made in TIMSS and other international large-scale assessments. When met, these assumptions allow users of the data to make valid inferences regarding student proficiency in subject

domains such as mathematics and science. They ensure that proficiency estimates are comparable across participating countries and over time, and generalizable within the assessment domains described in the framework beyond the limited sample of items each student received.

IRT models describe the probability of a correct response, given examinees proficiency and some item-specific parameters (such as the a_i, b_i described above). This, however, is not how IRT models are actually applied. Not only the item parameters but also the proficiency θ are unknowns that have to be estimated from the data, and all that analysts can rely on is a series of scored answers to a modest number of assessment items. What is needed, and what IRT provides for TIMSS, is a formal model that applies to an assessment domain as a whole, which is delineated in an assessment framework that describes the types of performances on topics viewed as representing the domain. The assumptions underlying IRT facilitate this goal in that they allow inferences about proficiency domains by providing a basis for proficiency estimates that depend on performance on assessment tasks in a well specified and scientifically testable way.

Unidimensionality

TIMSS assesses student achievement on several items students receive. Let I denote the number of items and let the response variables be denoted by $x = (x_1, \ldots, x_I)$. Unidimensionality means that a single quantity is sufficient to describe the probabilities of these responses to each of the items, and that this quantity is the same regardless of the selection of items a student received from within an assessment domain.

Denote P_{iv} and P_{jv} as the probability of person v scoring 1 on items i and j.

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

and

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

with the same real valued θ_v in each expression. Unidimensionality ensures that the same underlying proficiency is measured by all the test items in the domain. This of course holds only if the assessment development aims at producing a set of items that are indeed designed to assess the same assessment domain and that test developers diligently refer to the content specifications outlined in the assessment framework. Unidimensionality would (very likely) not hold, for example, if half of the items in a skills test consisted of multiplication problems, and the other half were assessing gross motor skills such as success on a soccer penalty kick practice. As these are two seemingly unrelated skills, one would likely need two proficiency scales: *Multiplication proficiency* and *Penalty kick proficiency*. However, if domains are closely

related, requiring for example different mathematical operations such as multiplication and addition, 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 probabilities of producing a correct response for a given level of proficiency are not dependent on the group to which a test taker belongs. In TIMSS, this independence is important for inferences across countries, but also within countries for inferences across different student groups. Formally population independence holds if

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

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

$$P(x_i, x_j \mid \theta) = P(x_i \mid x_j, \theta)P(x_j \mid \theta) = P(x_i \mid \theta)P(x_j \mid \theta). \quad (11.5)$$

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

$$P(\mathbf{X} = x_1, \ldots, x_I \mid \theta) = \prod_{i=1}^{I} P_i\,(X = 1 \mid \theta)^{x_i}[1 - P_i(X = 1 \mid \theta)]^{1-x_i}. \quad (11.6)$$

While this assumption appears to be a rather technical one, it can be made more understandable by the following considerations. The proficiency variable intended to be measured is not directly observable, so one can only make inferences about it from observable response behaviors that are assumed to relate to this variable. The assumption of population invariance and local independence facilitates these inferences, in that it is assumed that once a respondent's proficiency level is accounted for, responses become independent from each other, and also from other variables. That is, knowing whether or not a respondent taking a test has answered the previous question correctly does not help predicting the next response, if the respondent's proficiency level θ is known.

According to the assumption of population invariance and local independence, if the model fits the data (and, for example, no learning occurs) and only one single proficiency is 'responsible' for the probability of giving correct responses, then no other variables (including language of the assessment,

citizenship, gender, and other contextual variables) are helpful in predicting a respondent's answer to the next item. In this sense, the assumption of local independence and population invariance encapsulate the goal that there is only one variable that needs to be considered, and that estimates of this variable will fully represent the available information about proficiency.

Monotonicity of Item-Proficiency Regressions

One important assumption of IRT models used for achievement data is the (strict) monotonicity of item functions. As seen in Exhibit 11.1, the Rasch model (but also the 2PL and 3PL IRT models) assumes that the probability of a correct response increases with an increasing proficiency. This is represented in the following 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 the proficiency 'orders' the success on the items the students receive, and implies that students with a higher level on the proficiency will also have a higher probability of success on each of the items in the achievement domain. By implication, there is also a strict monotonic relationship between the expected achievement scores and proficiency θ:

$$E(S|\theta_v) = \sum_{i=1}^{I} P(X_i = 1 \mid \theta_v) < E(S|\theta_w) = \sum_{i=1}^{I} P(X_i = 1 \mid \theta_w) \leftrightarrow \theta_v < \theta_w. \quad (11.7)$$

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

While the assumptions described above lay the foundation for IRT (and more generally, a large number of latent variable models), each of these assumptions can be relaxed to account for specific attributes 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 used for the analysis of questionnaire data (as described in Chapter 16).

Specialized variants of the IRT models described here are used for reporting on an achievement domain when many different test forms are used, as well as when additional factors have to be accounted for. One such example is the transition from paper- to computer-based assessment. In the context of TIMSS 2019, the move from paper-based to computer-based administration and the need to accommodate

both administration modes in estimating student proficiency requires statistically sound extensions of IRT models. The next section describes such psychometric tools that can be applied to enable the transition to computer-based testing.

Accounting for Mode of Administration Effects

The change from paper- to computer-based testing requires careful consideration, as students taking the assessment are faced with different types of response modalities (e.g., a keyboard and mouse or a touchpad or touchscreen, compared to a pencil and a paper sheet to record the answers). This section describes methods for linking the paper-based and the computer-based assessment data, utilizing appropriate extensions of IRT models to establish this link. Chapter 13 of this volume presents country-by-country data based on comparisons of the computer-based eTIMSS 2019 assessments and the paper-based bridge assessments. These comparisons focus on observed item statistics as well as estimates of expected proficiency scores.

Despite the advantages of computer-based assessments, the move from a paper- to a computer-based assessment mode poses challenges for the measurement of trend over time because the results of an assessment administered in different modes may not be directly comparable. One concern is that some assessment items may not function the same across modes and may differ in their difficulty, discrimination, or with respect to the composition of skills they tap into. Mode effects may manifest as differential item functioning (DIF) by (at least) some of the items when comparing equivalent groups across different assessment modes. This, in turn, can affect measurement invariance and may cause undesirable changes in comparability of proficiency scores.

The following section provides an overview of the types of violations of measurement invariance and presents extensions of the IRT models described above that can be used to examine mode effects. The approach presented here was used to select an appropriate adjustment for linking the proficiency scales across modes in TIMSS 2019.

Comparability and Measurement Invariance

There are different levels of measurement invariance (Meredith, 2003; Millsap, 2010) that have to be considered before comparing achievement from different groups or assessments across modes or over time. For valid comparisons, the assessments ideally should exhibit *scalar* or *strong* invariance for all items. This means that the same statistical quantities (IRT item parameters in this context) can be used to fit the items independent of the mode of administration. Weaker forms of invariance are *metric* invariance, where slope parameters are invariant across modes while intercepts are allowed to vary across modes or groups, and finally *configural* invariance, where the same loading pattern can be maintained.

When accounting for mode effects, scalar invariance is the gold standard, while metric invariance is a somewhat less desirable but still a manageable level of invariance as long as proper linking designs can be used to adjust for mode differences (von Davier, Khorramdel, He, Shin, & Chen, 2019). In international assessment, any two cycles are different due to item release and new item development. Therefore, as long as a large proportion of the items reach scalar or metric invariance across modes, it is quite appropriate to have a subset of items with weaker forms of invariance, while most items show strong invariance over time and across modes. Trends measured across modalities are expected to be comparable in order to assess change, and trend measures should provide consistent statistical associations across modes, particularly with external variables central to establishing validity. Ensuring that a large proportion of items show strong invariance properties is crucial for these comparisons. It should be noted that mode effects are just one possible source of violations of measurement invariance. Other sources such as translation errors, technical issues, and language differences are routinely examined and treated as well (e.g. Oliveri & von Davier, 2011; von Davier et al. 2006; von Davier & Sinharay, 2013) in fully paper-based as well as in computer-based assessment.

Assessment Design Requirements for Studying Mode Effects

To deal effectively with mode effects, the assessment design needs to involve items that are *by design comparable*. If only student groups are comparable and take completely different items in paper and computer-based assessments, little can be said about mode differences as items are not comparable. Paper-based assessment items converted for computer delivery so that they can be considered equivalent in terms of content, presentation, and response requirements are referred to here as *by design comparable* items, or *comparable items* for short. About 80 percent of the TIMSS 2019 trend items are in this category and provided a strong link across assessment modes (see Chapter 12: Implementing the TIMSS 2019 Scaling Methodology).

To evaluate the extent to which measurement invariance can be assumed when moving from a paper- to computer-based assessment, an appropriate *data collection design* is needed where the same items are administered in both modes to either the same test takers or equivalent groups of test takers. For operational efficiency, administering the assessment to each student in one mode only is often preferred, while randomly assigning students to modes so that groups taking the assessment in one or other mode are randomly equivalent and results can be compared. In this approach, the two modes of delivery can be understood as treatment assignments in an experiment, while the two randomly assigned (and hence equivalent) groups of students can be assumed to have the same proficiency distribution.

To be able to generalize from such a bridge study, a sufficiently large and representative sample for both modes is needed at the level at which inferences are planned. For example, if the level of inference is how items function in two modes on aggregate at the international level, the two samples must cover

the range of abilities that are assessed across countries. On the other hand, if the level of inference is the detection and potential treatment of mode effects at the individual country level, the samples for each country would have to be sufficiently large to enable stable estimates of item parameters at that level.

In TIMSS, this would at least require two large samples of at least the size of the TIMSS national sample (150 schools or more, and 1-2 classrooms of approximately 30 students), one for eTIMSS and the other for paperTIMSS. Because this was not possible due to limited resources being available at the national level, TIMSS 2019 opted for a bridge design to link modes at the international level, with eTIMSS countries selecting a full student sample for eTIMSS together with a smaller, randomly equivalent bridge sample for paperTIMSS. The bridge sample of 1,500 students provided about 375 responses per item per country and was sufficient to evaluate items for mode effects at the international level (i.e., aggregated across all samples). However, these sample sizes are not large enough to provide stable item parameter estimates for individual countries, and hence country-level studies in international contexts require a careful consideration of the limitations of the sample. An example of feasible analyses at the country level is given in Chapter 13 of this volume.

Once the data is collected in both modes, a bridge data set that provides comparable data on the previous mode of assessment, and a new mode data set, statistical analysis and psychometric modeling can commence.

Analysis of Mode Effects Using Graphical Model Checks

As an initial comparison prior to any psychometric modeling approaches or IRT-based analysis, *graphical model checks* (e.g. Khorramdel & von Davier, 2016; Rasch, 1960) can provide important insights. These checks reveal whether the rank order of item parameters and the relations between item parameters agree strongly (as they should) in the eTIMSS and paperTIMSS samples, ensuring that invariance assumptions implemented in subsequent statistical and psychometric modeling are tenable. For this analysis, item parameters for comparable items were estimated separately for the eTIMSS and paperTIMSS samples but pooled across countries to focus on mode comparisons only and to ensure sufficient sample sizes for accurate calibrations.

Exhibits 11.3 through 11.6 show examples of location parameter comparisons between modes using the eTIMSS and paperTIMSS data (including bridge) for each grade and subject assessed by TIMSS. It should be noted that eTIMSS bridge samples responded to trend items only, so comparisons for "new" items are less informative for mode comparisons due to different countries taking the new items in different modes.

Exhibit 11.3: Location Parameter Comparison between eTIMSS and paperTIMSS 2019—Grade 4 Mathematics

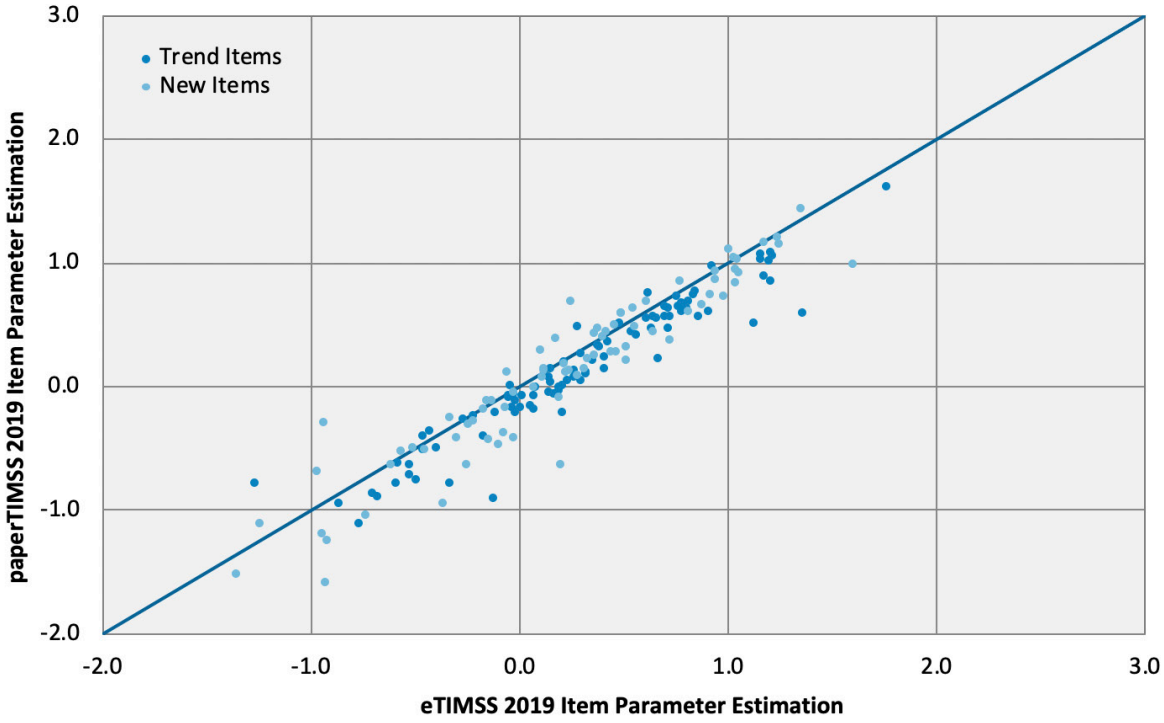


Exhibit 11.4: Location Parameter Comparison between eTIMSS and paperTIMSS 2019—Grade 4 Science

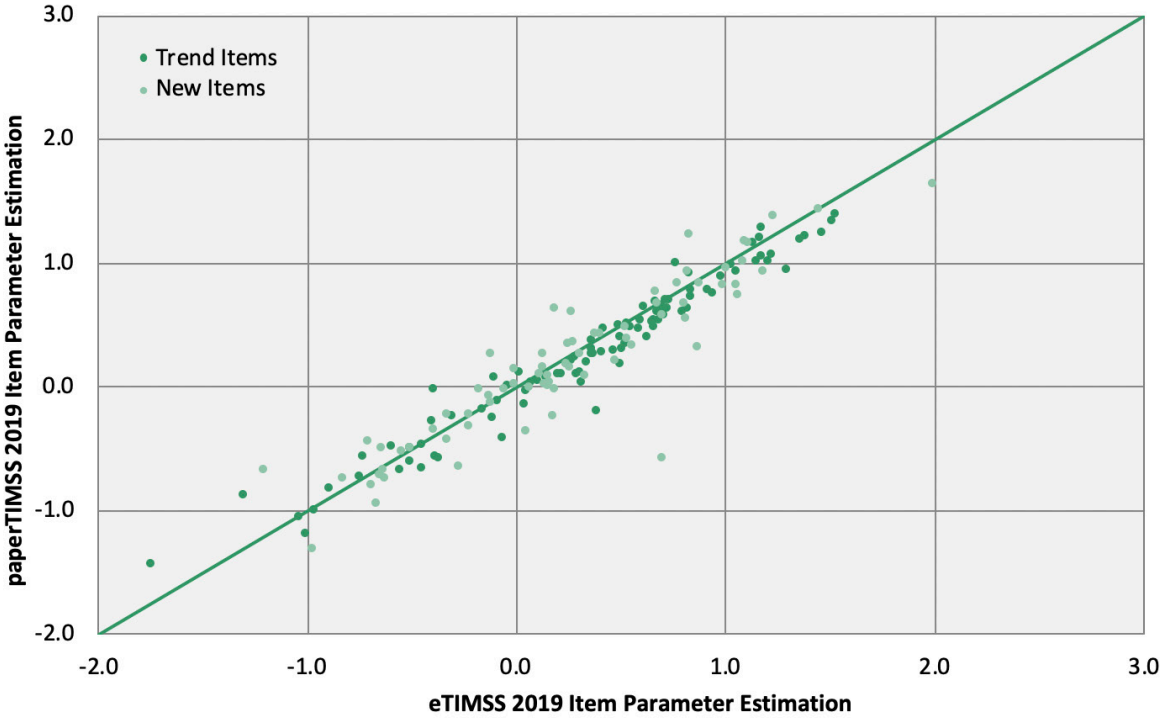


Exhibit 11.5: Location Parameter Comparison between eTIMSS and paperTIMSS 2019—Grade 8 Mathematics

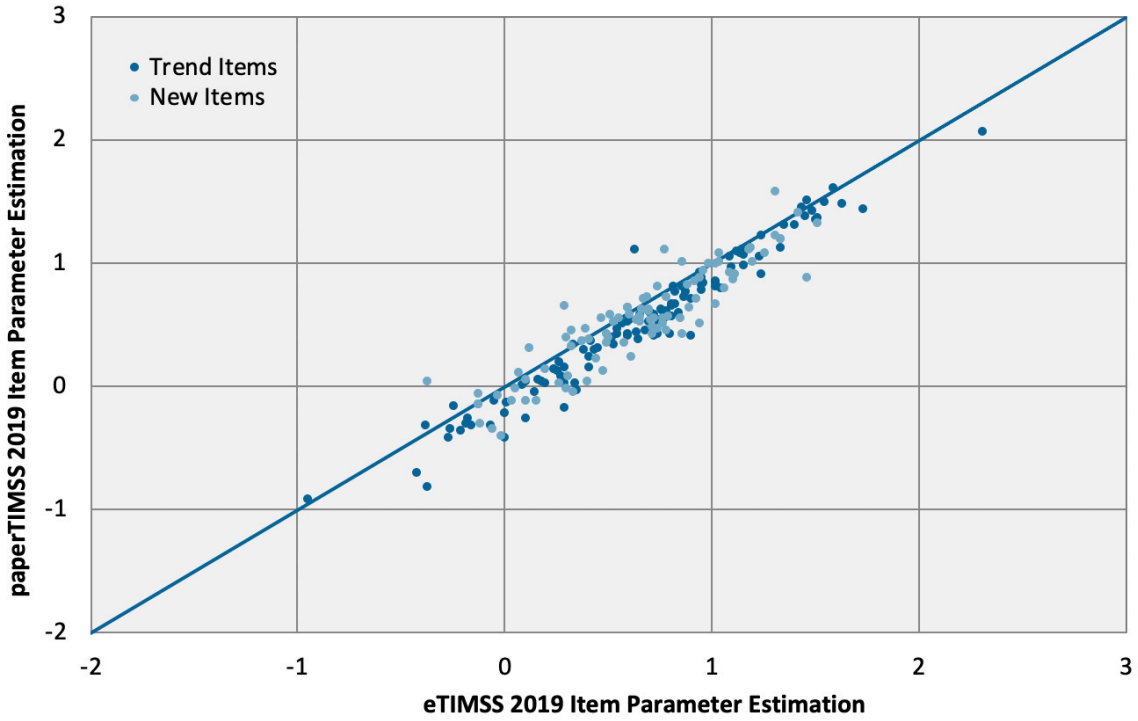
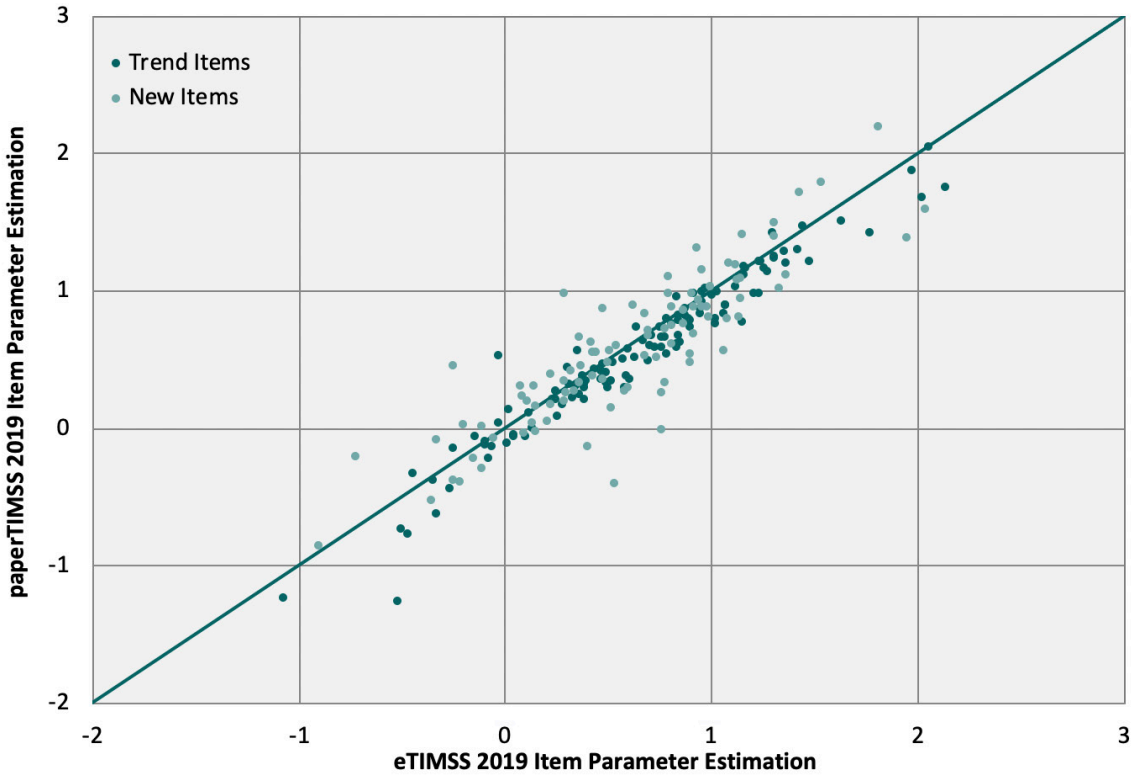


Exhibit 11.6: Location Parameter Comparison between eTIMSS and paperTIMSS 2019—Grade 8 Science



Exhibits 11.3 through 11.6 show that item location parameters (item difficulties) are highly correlated across the eTIMSS and paperTIMSS modes. A similar level of agreement was found for other parameter types. These results suggest there is excellent agreement between paper- and computer-based items for items that were deemed comparable based on design and response similarity across modes. The presence of some outliers, however, suggests that some items differ between modes and may require separate parameter estimates. Note that there is always estimation error in item parameter estimates and, therefore, parameter estimates from two finite samples are never perfectly correlated, even for two independent samples taking the same assessment in the same mode. However, the cross-mode correlations between item parameters of paper- and computer-based items for TIMSS 2019 are very high, suggesting that a strong link can be established so that computer- and paper-based results across countries can be reported on the same scale.

Before such a link can be established, the extent to which some items may exhibit mode effects, and may require separate estimates, has to be carefully examined during IRT scaling. The next section provides an overview of IRT model extensions that facilitate the examination of mode effects and for linking across assessment modes by testing for, and if present, utilizing the invariance of item parameters across modes.

Mode Effect Models

While graphical model checks provide a useful starting point for examining overall agreement between item parameters from different samples and for exploring potential drivers of these differences, they do not provide the most rigorous way to account for mode effects in proficiency estimation (e.g. von Davier & von Davier, 2007). Extensions of IRT models such as the ones described subsequently can be used to analyze mode differences with a high level of statistical rigor in order to obtain unbiased proficiency estimates by utilizing the equivalency of the bridge and eTIMSS samples in the analysis.

IRT models have been extended to include various types of mode effect parameters in order to provide information about whether the mode effect is best described by an overall difference between assessment modes (i.e., the difference between modes is changing the difficulty of all comparable assessment items by a constant), whether it is a person- or group-specific effect that may have an impact differentially on different groups (i.e., some test takers are more affected by mode differences than others), or whether it is an item-specific effect that is only impacting a subset of tasks.

These different hypotheses about mode differences can be checked by formalizing these within a general latent variable model (von Davier, 2008; von Davier, Xu, & Carstensen, 2011) and applying these models to the eTIMSS and bridge data. Taking the two-parameter logistic model (Birnbaum, 1968) as the base model, von Davier et al. (2019) introduced additional model parameters to formalize various assumptions of how mode effects may impact item functioning. Let

$$P(x = 1|\theta, \alpha_i, \beta_i) = \frac{\exp(\alpha_i\theta + \beta_i)}{1 + \exp(\alpha_i\theta + \beta_i)} \tag{11.8}$$

denote the probability of a correct response by a respondent with proficiency θ for an item i with parameters α_i, β_i. The notation used in (11.2) can be transformed to the customary notation by letting $a = \alpha/1.7$ and $b = -\beta/\alpha$.

Mode Effects on the Item Level

The most parsimonious mode effect assumption is that all items show strong invariance and need to be "shifted" by a certain amount with respect to their difficulty when comparing groups taking the assessment in one mode of administration with another. This could be because, for example, reading any item stem or stimulus is generally harder or easier (by the same amount for all items) on the computer, or responding using the keyboard or a mouse is more tedious or simpler than bubbling in a response on an answer sheet. Here, the mode is a "treatment" that changes the apparent average proficiency between groups, which needs to be corrected for using the equivalency of randomly assigned groups taking the bridge and the eTIMSS assessment, respectively. A mode treatment effect of this type that applies homogeneously to all comparable items can be controlled for by adding the same constant to each of the item difficulty parameters. This general mode effect parameter δ_m quantifies how much more difficult (or easy) all the comparable items appear when presented in a mode other than the reference mode (11.9). In terms of standard IRT linking designs, this general mode effect shift is similar to a non-equivalent groups design with anchor test (NEAT). However, the groups were randomly assigned, so the non-equivalence is really caused by the treatment (mode) that has an overall effect, which can be controlled for through the δ_m that reflects treatment differences.

Formally, for items presented in the "new" mode, we assume that

$$P(X = 1|\theta, \alpha_i, \beta_i, \delta_m) = \frac{\exp\left(\alpha_i\theta + \beta_i - 1_{\{I+1,\ldots,2I\}}(i)\delta_m\right)}{1 + \exp\left(\alpha_i\theta + \beta_i - 1_{\{I+1,\ldots,2I\}}(i)\delta_m\right)}. \tag{11.9}$$

This can be thought of as a model for twice the number of items. The indicator function $1_{\{I+1,\ldots,2I\}}(i)$ equals 1 if the item index is in the second half, that is, the range $I + 1, \ldots, 2I$. The first $1, \ldots, I$ items are the paper-based items without mode effect, and the items in the new mode are indexed by $I + 1, \ldots, 2I$. In this notation it is assumed that item i and item $i + I$ are the same but administered in different modes. This leads to a model with $2I$ items (instead of I items for each delivery mode) in which the difficulty parameters for items presented in one mode (say, paper) are assumed to be β_i for $i = 1, \ldots, I$ and the item parameters for the other mode (say, computer) are appended as parameters β_j for $j = I + 1, \ldots, 2I$ and arranged in the same order and constrained to follow $\beta_j = \beta_i - \delta_m$.

In the bridge design, each test taker receives a subset of items from either the paper-based items, indexed by $i = 1, \ldots, I$, or the computer-based items indexed by $i = I + 1, \ldots, 2I$. The two assignments are based on randomly equivalent respondents that only differ in the treatment they received, the mode of assessment. Note that this form of adjustment is equivalent to assuming the item parameters for comparable items to be strongly invariant and adjusting only for the overall mean differences, between bridge and eTIMSS sample. This model can be estimated by assuming one ability distribution across groups assessed in different modes and adding the mode parameter δ_m as an explanatory effect to the items administered in the new mode.

In contrast to the assumptions of a general mode effect parameter, δ_m, one could argue that not all items are affected when moving from paper to computer: Some could be more difficult, some could be at the same difficulty level, and some could even get easier. This leads to a model with weaker invariance that adds an item-specific effect δ_{mi} to the difficulty parameter. This can be written as a DIF parameter, quantifying item-specific changes from paper presentation, namely

$$P(X = 1|\theta, \alpha_i, \beta_i, \delta_m) = \frac{\exp\left(\alpha_i\theta + \beta_i - 1_{\{I+1,\ldots,2I\}}(i)\delta_{mi}\right)}{1 + \exp\left(\alpha_i\theta + \beta_i - 1_{\{I+1,\ldots,2I\}}(i)\delta_{mi}\right)}. \tag{11.10}$$

The difference in comparison to the model of metric (or "weak") factorial invariance (Meredith, 1993) is that the computer-based item difficulties relative to the paper-based difficulties are decomposed into two components, that is $\beta_{i+I} = \beta_i - \delta_{mi}$, while continuing to assume that $\alpha_{i+I} = \alpha_i$ for the slope parameters. This decomposition indicates that the difficulties are shifted by some (item or item feature)-dependent amount, the shift being applied to one mode on an item-by-item basis—one that is being considered the reference mode with no shift. Assuming ability equivalence, the average treatment (mode) effect can be assessed by calculating $\frac{1}{I}\sum_i \delta_{mi} = \bar{\delta}_m$. This average effect can be compared against the estimated average effect from model (11.9).

The model in equation (11.10) with constraints across both modes on slope parameters, as well as potential constraints on the DIF parameters, establishes weak (also sometimes called metric) invariance (e.g., Meredith, 1993) IRT model, whereas model (11.9), which TIMSS 2019 was able to use, establishes strong invariance. The average mode effect is equivalent to a shift in group means when the item parameters are invariant in model (11.9), whereas model (11.10) allows individual items to deviate from this average mode shift. The larger the number of constraints of the type $\delta_{mi} = c$ for some constant adjustment can be assumed, the more we approach a model with strong factorial invariance, that adjusts only for overall mode treatment differences. Note that an overall adjustment as used in TIMSS 2019 retains the equality of means and variances of the latent variable in both modes as both groups were randomly assigned to modes but selected from a single population.

Mode Effects on the Respondent or Proficiency Level

For completeness of discourse, if it cannot be assumed that the mode effect is a constant (even if item dependent) shift for all respondents, then an additional proficiency may be required to accurately model response probabilities for the new mode. This leads to a multidimensional model with a second latent variable that is added to the item function for items administered in the new mode. The expression $\alpha_{mi}\vartheta$ in the model (11.11) below indicates that there is a second slope parameter α_{mi} for items ($i = I + 1, \ldots, 2I$) administered in computer mode and that the effect of the mode is person dependent and quantified through a second latent variable ϑ. We obtain

$$P(X = 1|\theta, \alpha_i, \alpha_{mi}, \beta_i, \vartheta) = \frac{\exp(\alpha_i\theta + \beta_i - \alpha_{mi}\vartheta)}{1 + \exp(\alpha_i\theta + \beta_i - \alpha_{mi}\vartheta)}. \quad (11.11)$$

Note that the common slope parameters, α_i, and item difficulties, β_i, are, as before in models (11.9) and (11.10), equal across modes. However, an additional "mode-slope" parameter α_{mi}, for $i = I + 1, \ldots, 2I$, is estimated, with constant $\alpha_{mi} = 0$ for $i \leq I$ for the reference items that are not affected by mode changes. For the joint distribution $f(\theta, \vartheta)$ one assumes uncorrelated latent variables, $\text{cov}(\theta, \vartheta) = 0$, to ensure identifiability in the bridge design.

In equation (11.11) it is assumed that the effect of the person "mode" variable varies across items, which may be the more plausible variant, but a model with item-invariant effects $\alpha_m\vartheta$ (a Rasch variant of a random mode effect) also is feasible. However, an item-specific model is more likely to provide better model data fit. As in model (11.10), the link between modes can be viewed as increasingly more invariant as more slope parameters can be assumed to be $\alpha_{mi} = 0$ for items in the new mode. Each constraint $\alpha_{mi} = 0$ makes the respective item response functions for items i and $i + I$ identical across modes.

Application of Mode Effect Models to TIMSS 2019

The models presented above were available to accommodate a range of mode effects and item invariances across the two TIMSS assessment modes. However, based on the very good agreement between bridge and eTIMSS sample estimates (see Exhibits 11.3–11.6) of item parameters for the TIMSS 2019 comparable items, it was concluded that only a small overall mode adjustment constant was necessary (see Chapter 12). This adjustment was estimated separately for mathematics and science at the fourth and eighth grades. Additional analyses with standard IRT linking methods (Haebera, 1980; Marco, 1977; von Davier & von Davier 2007; Xu & von Davier, 2008b) were in agreement with the results obtained from model (11.9) as well as with the graphical model checks, so that this convergence of results supported the use of an overall mode adjustment.

Using a single adjustment of parameter for each subject/grade combination based on the randomly equivalent samples from the bridge and eTIMSS samples keeps the scaling methods in line with prior

TIMSS trend scaling methods, and enables country-level mode effect analyses as presented in Chapter 13. The eTIMSS sample and bridge sample were of central importance for linking through model (11.9) because, as randomly equivalent groups with a large set of comparable items as anchors, they form the basis for estimating the adjustment using the proficiency distribution estimates in the two modes.

After establishing the size of the item parameter adjustment required by model (11.9), this adjustment was applied to each of the comparable items in scaling the eTIMSS data, resulting in eTIMSS data on the same scale as the bridge data. The effect of the adjustment was verified in terms of item fit and scaling outcomes using country adjustment compared to the separate scaling of items in equivalent groups designs (see Chapter 12).

The major outcome of the foregoing procedure was that the eTIMSS 2019 proficiency data were successfully linked to the existing TIMSS proficiency scales so that results from the paper- and the computer-based assessments can be directly compared without any further adjustment. Very high levels of comparability of item parameters across the two administration modes were established, so that the mode-adjusted item parameters can be used in the population model described in the following section. This population model is used to generate plausible values for estimation of group level results and to examine the relation between student proficiency and other contextual variables. The strong link of paperTIMSS and eTIMSS across modes based on comparable items and equivalent groups design enabled reporting TIMSS 2019 on the same scale for all participating countries. It also formed the basis of an important and final step that provides the proficiency database by means of a country specific population modeling approach as described in the next section.

Population Models Integrating Achievement Data and Context Information

TIMSS uses a latent regression (or 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 (von Davier, Gonzalez, & Mislevy, 2009; von Davier et al., 2006). This approach can be viewed as an imputation model for the unobserved proficiency distribution that aims at obtaining unbiased group-level proficiency distributions. The approach requires the estimation of an IRT measurement model, which provides information about how responses to the assessment items depend on the latent proficiency variable. In addition, the latent regression, which provides information about the extent to which background information is related to achievement, is used to improve estimates by borrowing information through similarities of test takers with respect to context variables and the way these relate to achievement. The population model is estimated separately for each country and in TIMSS 2019 five plausible values (PVs) representing the proficiency variable are drawn

from the resulting posterior distribution for each respondent in each cognitive domain. It is important to note 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 IDB analyzer.

Population models are examples of high dimensional imputation models, and utilize a large number of context variables in the latent regression to avoid omission of any useful information collected 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 (PCA) of the student context variables is used to eliminate collinearity by identifying a smaller number of orthogonal predictors that account for most of the variation in the background variables (90% in the case of TIMSS 2019).

In order to fully describe the proficiency estimation procedure, the data from the context questionnaires are combined with the responses obtained from the achievement items. The complete observed data for a person n can be expressed as $d_n = (x_{n1}, \ldots, x_{nI}, g_n, z_{n1}, \ldots, z_{nB})$, 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.

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

$$\mu_n = \sum_{b=1}^{B} \beta_{g(n)b}\, z_{nb} + \beta_{g(n)0}. \tag{11.12}$$

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

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 TIMSS items. In other words, the model implements the assumption that the posterior distribution of proficiency depends on the context data as well as on the observed achievement. 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 background variables are selected so that correlations with proficiency are likely, one obtains a distribution around the expected value given in (11.12) 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 addressed separately within countries. The regression is country specific since 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.

There are several ways to address the estimation of the latent regression parameters. In TIMSS and other large-scale assessments, the latent trait (proficiency) is determined by the IRT model estimated across countries in a previous step. Then the (latent) regression model is estimated treating the item parameters from the previous IRT estimation as fixed quantities. This ensures that the invariance properties that were determined through IRT estimation and potential mode effect adjustments across countries are applied equally to each national dataset (see for example, Mislevy & Sheehan, 1992; Thomas, 1993; von Davier et al., 2006; von Davier & Sinharay, 2013).

Group-Level Proficiency Distributions and Plausible Values

The goal of the psychometric methods described above is to produce a useful database that contains comparable, valid, and reliable information for reporting student proficiency and for secondary users of the TIMSS assessment data. This information comes in the form of likely proficiency estimates for all respondents given 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 context information, as

$$P_g(\mathbf{x}_n \mid \mathbf{z}_n) = \int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni} \mid \theta)\, \phi\left(\theta; \sum_{b=1}^{B} \beta_{gb}\, z_{nb} + \beta_{g0}, \sigma\right) d\theta. \qquad (11.13)$$

This equation provides the basis for the imputation of proficiency estimates 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 TIMSS assessment items $x_1, \ldots, x_I$ and the context information $z_1, \ldots, z_B$. The posterior distribution of the proficiency given the observed data can be written as

$$P_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n) = \frac{\prod_{i=1}^{I} P_{ig}(x_{ni} \mid \theta)\phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right)}{\int_\theta \prod_{i=1}^{I} P_{ig}(x_{ni} \mid \theta)\phi\left(\theta; \sum_{b=1}^{B} \beta_{gb} z_{nb} + \beta_{g0}, \sigma\right) d\theta}. \quad (11.14)$$

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

$$E_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n) = \int_\theta \theta\, P_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n) d\theta. \quad (11.15)$$

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

$$V_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n) = E_g(\theta^2 \mid \mathbf{x}_n, \mathbf{z}_n) - \left[E_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n)\right]^2. \quad (11.16)$$

Using these two estimates (the mean and variance) to define the posterior proficiency distribution, it is possible to draw a set of plausible values (Mislevy, 1991) 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.

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

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

$$\widetilde{\theta}_{ng} \sim N\left(E_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n), \sqrt{V_g(\theta \mid \mathbf{x}_n, \mathbf{z}_n)}\right) \quad (11.17)$$

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

In order to provide a more detailed picture of the analytic methods, this chapter focused on the rationale behind the methodologies used in TIMSS 2019, ranging from IRT, to mode effects, to population modeling for unbiased reporting of group level proficiency distributions. Additional information is available in the chapter on scaling outcomes (Foy, Fishbein, von Davier, & Yin, 2020) and the chapter on examining country-level mode related quantities (von Davier, Foy, Martin, & Mullis, 2020).

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.

Birnbaum, A. (1968). *On the estimation of mental ability* (Series Report No. 15). Randolph Air Force Base, TX: USAF School of Aviation Medicine.

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

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

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

Haberman, S. J., von Davier, M., & Lee, Y.-H. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous ability distributions.* Educational Testing Service RR-08-45. Princeton, NJ: Educational Testing Service.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144–149.

Khorramdel, L., & von Davier, M. (2016). Item response theory as a framework for test construction. In K. Schweizer & C. Distefano (Eds.), *Principles and methods of test construction: standards and recent advancements* (pp. 52-80). Göttingen, Germany: Hogrefe.

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.

Marco, G. L. (1977). Item characteristic curves solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Martin, M. O., Mullis, I. V. S., and Foy, P. (2017). TIMSS 2019 assessment design. In I. V. S. Mullis & M. O. Martin (Eds.), *TIMSS 2019 assessment frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2019/frameworks/

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*, 5–9. doi:10.1111/j.1750-8606.2009.00109

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.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, 53(3) 315-333.

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

San Martín, E., González, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika, 80*(2), 450-467.

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. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, Vol. 61, No. 2. (November), pp. 287-307. https://doi.org/10.1348/000711007X193957

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspectives*, 7(2), 110-114, doi: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., Foy, P., Martin, M. O., & Mullis, I. V. S. (2020). Examining eTIMSS country differences between eTIMSS data and bridge data: A look at country-level mode of administration effects. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 13.1-13.24). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/methods/chapter-13.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 pschometric 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., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76,* 318-336.

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

Xu, X., & von Davier, M. (2008a). *Fitting the structured general diagnostic model to NAEP data* (ETS Research Report, RR-08-27). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008b). *Linking with the general diagnostic model* (ETS Research Report, RR-08-08). Princteton, NJ: Educational Testing Service. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.2008.tb02094.x/full

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.

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