Organisation for Economic Co-operation and Development

EDU/WKP(2022)20

Unclassified **English - Or. English**

16 February 2023

DIRECTORATE FOR EDUCATION AND SKILLS

Cancels & replaces the same document of 16 February 2023

THEORETICAL CONSIDERATIONS ON SCALING METHODOLOGY IN PISA

OECD Education Working Paper No. 282

Tomoya Okubo (OECD)

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Tomoya Okubo, Tomoya.OKUBO@oecd.org

JT03512465

OECD EDUCATION WORKING PAPERS SERIES

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Corrigendum

An early version of this report from January 2023 was revised:

Page 2: Insert description of OECD EDUCATION WORKING PAPERS SERIES and disclaimers, as above.

Page 3: Insert line at end of "Acknowledgement": "Jenny Baracaldo Fernández, Stephen Flynn and Rachel Linden contributed to the editing and production of this paper."

Page 6: Paragraph two, line three, change "Hopfenbeck, 2018" to "Hopfenbeck et al., 2018".

Page 6: Paragraph two, line three, change "Baird, 2016" to "Baird et al., 2016".

Page 25: Paragraph one, line nine, change "Chen, 2022" to "Chen et al., 2022".

Page 28 (References): Reference three, change "Baird, J. (2016)"to "Baird, J. et al., (2016)".

Page 28 (References): Reference 10, change "Chen, S. (2022)"to "Chen, S. et al. (2022)."

Page 29 (References): Reference 18, change "Hopfenbeck, T. (2018)" to "Hopfenbeck, T. et al. (2018)".

Page 30 (References): Reference 49. Change "van de Vijver, F. (2019)" to "van de Vijver, F. et al. (2019)".

Acknowledgement

The author would like to thank Dr Kentaro Kato[1] for his substantial comments and careful reviews, especially in mathematical parts. Further, thank you to Prof Shi-ich Mayekawa[2], who supported in elucidating the population modelling.

[1] Benesse Educational Research & Development Institute

[2] The National Center for University Entrance Examinations

Abstract

OECD Programme for International Student Assessment (PISA) scaling methodologies are reviewed from the mathematical perspective. In particular, the paper aims to elucidate the model structures and the model assumptions of item response theory used in the item scaling phase, the latent regression model employed in the student prior estimation phase, and the population modelling and Laplace approximation performed in the multiple imputations phase. It provides insights for maximising the impact of the innovations implemented in PISA while minimising the risk of impairing the reliability and the validity of the assessments, which is essential for maintaining technically sound international assessments. Based on the theoretical considerations, the scale robustness of PISA is confirmed against the growing number of countries/economies; nevertheless, the importance of assessing the model-data fit and investigating residual structures by sub-populations are indicated in this study. Furthermore, analysis procedures for nuisance factors such as testlet effect and locally dependent items are introduced in this paper. In summary, this study provides the theoretical considerations underpinning the stability and extensibility of the PISA scale.

Table of Contents

1. Introduction

1.1. PISA as an international benchmark

The OECD Programme for International Student Assessment (PISA) is an international large-scale assessment (ILSA) in which 15-year-old students' scholastic proficiencies are measured with cognitive items developed under the PISA analytical framework (OECD, 2019[1]) with the aim of assessing aspects of preparedness for adult life (OECD, 2000[2]). Together with the cognitive tests, PISA collects data on the students' background, perceptions, and attitudes in order to analyse relationships between student proficiency and student background variables. The cognitive items and the student questionnaires are designed to be comparable among countries/economies, and the instruments have been verified through field trials and the main studies of PISA over decades.

PISA has worked as an international benchmark of education systems, which maximises its value through publishing all the collected data (OECD, n.d.[3]). Numerous studies have been conducted based on the published data (Hopfenbeck, 2018[4]), and the extracted knowledge and expertise have been transferred into education policy and utilised in school practices (Lindblad, 2017[5]). Every time PISA publishes the results and data, they are featured by the media, and lively discussions are conducted among policymakers and educational practitioners (Baird, 2016[6]). Moreover, derivative instruments of PISA, such as PISA-based Test for Schools (PBTS) and PISA for Development (PISA-D), were developed under the same analytical framework and scaled on the same PISA scale. Therefore, the quality of the measure is of particular importance in PISA.

PISA takes place every three years, and some items are repeatedly used (i.e., trend items) in order to provide comparable scores across the testing cycles. In PISA, item response theory (Lord, 1968[7]) is employed for scaling students' proficiencies and item characteristics; the scale comparability of PISA across testing cycles is ensured by the trend items and their item parameters, which define the PISA international scale. Since item response theory (IRT) provides us with student proficiency measures that take into account characteristics of the cognitive items, different sets of items can be assembled within the same test. The flexibility of test design enables us to assess the target population longitudinally under an appropriate test design; thus, the model has been widely used as the scaling model in ILSAs for decades.

PISA continuously evolves. Since 2012, PISA introduced the innovative domain assessments in order to provide the participating countries with a complex outlook on their students beyond their core literacies. For example, in the PISA 2022 creative thinking assessment, interactive item-types are incorporated into the test (Foster, 2022[8]). Furthermore, In PISA 2018, multi-stage adaptive testing (MSAT) was implemented in the reading domain, in which the item clusters to be presented to the students are selected according to the student's predicted proficiency at breakpoints in the assessment. MSAT is expected to mitigate the mismatch of the item difficulties for each student (Yamamoto, 2019[9]). Not only item-types evolution but also the number of participating countries in PISA is growing; 79 countries/economies joined in PISA 2018 where the new participating countries tend to have a higher proportion of low-performing students compared to OECD member countries. Obviously, the evolution of instruments and the changed circumstances as regards participation elicit a question on the impact of the changes on the reliability, validity, and international comparability of the assessments. The statistical modelling should be in accordance with the contents and the purpose of the research; therefore, in this paper, the stability and extensibility of the PISA scale are discussed from a statistical modelling point of view.

1.2. Scaling procedures

For the estimation of student proficiency, scaling by IRT is not the only procedure employed in PISA, whereas it is for many of the world's high-stakes examinations and qualification examinations. Therefore, the impact of introducing innovations into PISA should be considered in all the phases of the PISA scaling process. There are 3 phases up until generating the student-level datasets: namely, item scaling in Phase 1, student's prior distribution estimation in Phase 2, and multiple imputations of student-level proficiencies in Phase 3.

In Phase 1, item parameters are estimated using IRT. In order to keep comparability of the scores across testing cycles, the parameters of the trend items are fixed (OECD, 2017[10]). The remaining item parameters are imposed to be equal across countries/economies at the first step, with the aim to ensure score comparability across countries/economies (i.e., international parameters). In ILSAs, it is essential to examine item invariances across countries so as to keep international comparability of the scale and the constructs. In a second step, items that showed discrepancy from the international parameters are released from the parameter equality constraints and their own parameters (i.e., national parameters) are estimated. The model-data fit check and the parameter estimations are repeated until the model is optimised; thus, the item parameters that define the PISA scale are finalised in Phase 1.

In Phase 2, the multivariate prior distribution with regard to student proficiencies is estimated for each student by country/economy. The prior distributions are estimated using the latent regression model (LRM), in which student proficiencies are regressed on the covariates calculated from the responses to the student questionnaire (Mislevy, 1984[11]). The role of the prior distribution is to increase the reliability of the estimates of student proficiencies; moreover, a student's proficiency in a domain that has not been tested in the cognitive test due to the booklet design can be inferred based on the prior distribution, which is estimated based on the student's background information and the estimated parameters of the LRM. Lastly, the prior distribution ensures the convergence of the proficiency estimates of the students who got full marks or zero marks, while likelihood-based estimations, such as maximum-likelihood estimation (MLE), do not.

In Phase 3, multiple imputations with regard to the student proficiencies are performed. Together with the likelihood function calculated from the student responses to the cognitive items and their estimated parameters in Phase 1, the prior distribution obtained in Phase 2 constitutes the posterior distribution of each student in Phase 3. The random values drawn from the posterior distribution are called plausible values (PVs) in the context of ILSAs. Unlike the point estimates obtained from weighted likelihood estimation (Warm, 1989[12]) and MLE, the drawn PVs are random values representative of the posterior distribution of student proficiencies. The empirical distribution of the PVs over the students preserves unbiasedness and consistency with respect to the country/economy level statistics.

1.3. Objective and overview

The objective of this study is to revisit the statistical modelling employed for the scaling in PISA and to provide insights for maximising the expected effects of the innovations PISA introduces and minimising the risks of damaging the reliability and the validity of the assessments. Concretely, the following topics are discussed from psychometric point of view.

1. Scale robustness against the growing number of countries/economies.
2. Analysing impact of nuisance factors.

3. Population modelling and item assignment.

In this study, the statistical modelling employed in PISA will be reviewed thoroughly from the mathematical perspective; in particular, model definitions and assumptions will be confirmed so that the above-mentioned points can be discussed based on the proper understanding of the statistical models.

The remainder of this paper is structured as follows. In section 2, model definitions and assumptions of IRT are explained, and the response models of IRT employed in PISA are introduced. The aim of section 2 is to understand the model structure and the assumptions of the model in order to evaluate the impact of the nuisance factors. In section 3, the parameter estimation procedures of the model defined in the previous section are revisited to examine the factors that affect the model's likelihood function. Specifically, the influences of the country/economy proficiency levels on the item parameters are investigated. In section 4, the latent regression model, which estimates the students' prior distributions based on student background information, is shown in order to see how the model impacts the students' proficiencies inferences. In section 5, the methodology of constituting the posterior distribution of each student is explained, and the processes of the plausible draws are provided in detail. In section 6, the above-mentioned points are discussed based on the mathematical facts derived in the previous sections.

2. Modelling item response theory

2.1. Model identifications and assumptions

IRT is a latent variable modelling in which the probability of responding to category k $(= 0, \dots, K_j - 1)$ of item j $(= 1, \dots, J)$ is defined by the function of latent variables $\boldsymbol{\theta} = [\theta_1, \dots, \theta_m, \dots, \theta_M]$. The latent variables $\boldsymbol{\theta}$ are defined by the J items that measure M-dimensional latent traits on linear scales. In PISA, $\boldsymbol{\theta}$ are scholastic proficiencies (i.e., mathematics, reading, and science) as well as student's attitudes or perceptions (e.g., self-efficacy, attitudes towards immigrants, etc.) depending on the items that the domains compose.

In section 2.1, the model assumptions and the parameter identification issues of the model are formulated in order to provide information on the model features from a mathematical perspective and also to provide insights for analysing the impact of nuisance factors such as student's motivation toward the assessment and the effect of social desirability in the self-reporting items.

Consider the following latent variable model:

Equation 1

$$\boldsymbol{z} = \boldsymbol{\lambda\theta} + \boldsymbol{e}$$

where $\boldsymbol{z}$ is the size $J \times 1$ vector of latent variables, $\boldsymbol{\lambda}$ is the size $J \times M$ matrix of factor loadings, $\boldsymbol{\theta}$ are the M-dimensional latent variables, and $\boldsymbol{e}$ is the error vector of the model. Equation 1 is called a M-dimensional factor model. By setting the variances of $\boldsymbol{\theta}$ and $\boldsymbol{e}$ as $\boldsymbol{\Phi} = \mathrm{V}[\boldsymbol{\theta}]$ and $\boldsymbol{\Psi} = \mathrm{V}[\boldsymbol{e}]$, the covariance structure of Equation 1 is given by

Equation 2

$$\boldsymbol{\Xi} = \boldsymbol{\lambda\Phi\lambda}' + \boldsymbol{\Psi}$$

For $\boldsymbol{\theta}$ and $\boldsymbol{e}$, the following normal distributions are set.

Equation 3

$$\begin{bmatrix}\boldsymbol{\theta}\\ \boldsymbol{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{\mu}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Phi} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Psi}\end{bmatrix}\right)$$

Here, $\boldsymbol{\mu} = \mathbf{0}$ and $\text{diag}(\boldsymbol{\lambda\Phi\lambda}' + \boldsymbol{\Psi}) = \boldsymbol{I}$ are assumed. The probability density function of $\boldsymbol{z}$ is

Equation 4

$$p(\boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Phi}, \boldsymbol{\lambda}) = (2\pi)^{-\frac{m}{2}}|\boldsymbol{\Psi}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\lambda\theta})'\boldsymbol{\Psi}^{-1}(\boldsymbol{z} - \boldsymbol{\lambda\theta})\right)$$

Let x_j be a polytomous response to cognitive item j, which is generated from z_j and threshold τ_{jk} as

$$x_j = k, \qquad \text{if } \tau_{jk} \le z_{ij} \le \tau_{j(k+1)}$$

where $\tau_{j0} = -\infty$, $\tau_{jK_j} = +\infty$, and each item has $K_j - 1$ thresholds. The probability density function of obtaining $x_j = k$ is described as

Equation 5

$$p_j(x_j = k|\boldsymbol{\theta}) = \int_{\tau_{jk}}^{\tau_{j(k+1)}} \frac{1}{\sqrt{2\pi}\Psi_{jj}} \exp -\frac{1}{2}\left(\frac{z_j - \boldsymbol{\lambda}_j\boldsymbol{\theta}}{\Psi_{jj}^{1/2}}\right)^2 \mathrm{d}z_j$$

where Ψ_{jj} is j-th diagonal element of $\boldsymbol{\Psi}$. Applying the variable transformation

$$v = f(z_j) = \left(\frac{z_j - \boldsymbol{\lambda}_j\boldsymbol{\theta}}{\Psi_{jj}^{1/2}}\right)$$

to Equation 5 derives

$$p_j(x_j = k|\boldsymbol{\theta}) = \int_{f(\tau_{jk})}^{f(\tau_{j(k+1)})} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) \mathrm{d}v$$

Based on the assumptions specified in Equation 3, and by setting $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$ as

Equation 6

$$\boldsymbol{a}_j = \boldsymbol{\lambda}_j\left(\boldsymbol{\Xi} - \boldsymbol{\lambda}_j\boldsymbol{\Phi}\boldsymbol{\lambda}_j'\right)^{-\frac{1}{2}}$$
$$\boldsymbol{b}_j = \boldsymbol{\tau}_j\left(\boldsymbol{\lambda}_j\mathbf{1}_M\right)^{-1}$$

the following model expression of the multi-dimensional normal ogive IRT model (Muraki, 1995[13]) is obtained; note $\mathbf{1}_M$ is the size $M \times 1$ vector composes 1.

Equation 7

$$p_j(x_j = k|\boldsymbol{\theta}) = \int_{-\boldsymbol{a}_j(\boldsymbol{\theta} - b_{jk}\mathbf{1}_M)}^{-\boldsymbol{a}_j(\boldsymbol{\theta} - b_{j(k+1)}\mathbf{1}_M)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) \mathrm{d}v$$

Equation 7 shows the mathematical relation whereby the IRT normal ogive model is equivalent to the factor analysis model (Takane, 1987[14]).

The marginal distribution of Equation 4 is

$$p(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Phi},\boldsymbol{\lambda}) = (2\pi)^{-\frac{m}{2}}|\boldsymbol{\Xi}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{z}'\boldsymbol{\Xi}^{-1}\boldsymbol{z}\right)$$

The probability of obtaining a response vector $\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iJ}]$ is formulated as

Equation 8

$$p(\boldsymbol{X} = \boldsymbol{x}_i) = \int_{\boldsymbol{\tau}} p(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Phi},\boldsymbol{\lambda})\,\mathrm{d}\boldsymbol{z}$$

where the intervals of the multiple integrals denoted by $\boldsymbol{\tau}$ are $[\tau_{jk}, \tau_{j(k+1)}]$ if $x_j = k$. The parameters to be estimated in this model are $\boldsymbol{\mu}$, $\boldsymbol{\Phi}$, $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$; however, either one of $\boldsymbol{\mu}$ or $\boldsymbol{\tau}$, and $\boldsymbol{\Phi}$ or $\boldsymbol{\lambda}$ need to be fixed in order to identify the model. This model identification issue of the latent variable modelling is called scale indeterminacy. When fixing $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$, only a set of parameters needs to be fixed to identify the model. Note $\boldsymbol{\Psi}$ can be calculated based on the estimated $\boldsymbol{\mu}$, $\boldsymbol{\Phi}$, $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$.

Equation 8 can be transformed into

Equation 9

$$\begin{aligned} p(\boldsymbol{X} = \boldsymbol{x}_i) &= \int_{\boldsymbol{\tau}} \int_{\boldsymbol{\theta}} \prod_{j=1}^{J} p_j\left(z_{ij}|\boldsymbol{\theta},\boldsymbol{\lambda}_j\right) h(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Phi})\,\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{z} \\ &= \int_{\boldsymbol{\theta}} \prod_{j=1}^{J} \int_{\boldsymbol{\tau}} p_j\left(z_{ij}|\boldsymbol{\theta},\boldsymbol{\lambda}_j\right)\mathrm{d}\boldsymbol{z}\; h(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Phi})\mathrm{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \prod_{j=1}^{J} p_j\left(x_{ij}|\boldsymbol{\theta},\boldsymbol{\lambda}_j,\boldsymbol{\tau}_j\right) h(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\Phi})\mathrm{d}\boldsymbol{\theta} \end{aligned}$$

$p_j\left(x_{ij}|\boldsymbol{\theta},\boldsymbol{\lambda}_j,\boldsymbol{\tau}_j\right)$ appearing in Equation 9 is called the item response model in the context of psychometrics. Details of the item response model are shown in section 2.3. Consequently, IRT is a latent variable model, which can be described within the modelling framework of structural equation modelling (SEM).

In educational assessments, the residual structure is assumed to be diagonal.

Equation 10

$$\boldsymbol{\Psi} = \mathrm{diag}\left(\Psi_{11}, \ldots, \Psi_{JJ}\right) = \begin{bmatrix} \Psi_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Psi_{JJ} \end{bmatrix}$$

This assumption is called weak local independence (Stout, 1990[15]), especially when each element of $\boldsymbol{\Psi}$ follows a normal distribution. Local independence is an assumption, not an identification condition; therefore, non-zero off-diagonal elements in $\boldsymbol{\Psi}$ are allowed as far as the model is identifiable. Misspecifying $\boldsymbol{\Psi}$ as $\mathrm{offdiag}(\boldsymbol{\Psi}) = \boldsymbol{0}$ in the model (Equation 2) adds the covariance components Ψ_{ij}, that are supposed to be specified in $\boldsymbol{\Psi}$, into $\boldsymbol{\lambda}\boldsymbol{\Phi}\boldsymbol{\lambda}'$, which results in overestimating $\boldsymbol{\lambda}\boldsymbol{\Phi}\boldsymbol{\lambda}'$ of the corresponding items. Thus, the violation of the local independence assumption introduces a bias for the estimates of $\boldsymbol{\lambda}$ and $\boldsymbol{\Phi}$. Furthermore, Equation 6 indicates that an overestimation of $\boldsymbol{\lambda}$ is conducive to an underestimation of $\boldsymbol{\tau}$ (Lord, 1980[16]). Note that the misspecification of the model and the residual structure decreases the precision of the estimate as well (Chen, 2007[17]).

Various causes of violating the local independence assumption are discussed in Hoskens and De Boeck (1997[18]), Ferrara, Huynh, and Michaels (1999[19]), and Yen (1993[20]). From

a modelling perspective, it occurs when the parameters are not appropriately specified in the model, pointedly when $\boldsymbol{\lambda\Phi\lambda}'$ is not structured appropriately to cover $\boldsymbol{\Xi}$. A possible inappropriate model is a model with fewer factors for the data than actual. Suppose $\boldsymbol{\Xi}$ composes a M-dimensional data structure but only m ($m < M$) dimensions are modelled in $\boldsymbol{\lambda}$ and $\boldsymbol{\Phi}$; this leads to violation of local independence. A possible cause of violating the assumption is the influence of students' response behaviours to the particular sets of items, which may be due to the effect of the mismatch of item difficulties to the students in the cognitive tests (e.g. losing motivation, etc.) or the influences of item orders and positions (Sijtsma, 1996[21]). In any case, none of those nuisance factors are harmful themselves, but the discrepancy between the model and the data is the issue. Consequently, appropriate model setting is crucial, especially in ILSAs.

2.2. Scale indeterminacy and linear transformation

An important feature of the latent variable model is that it involves scale indeterminacy. In order to avoid scale indeterminacy, at least either one of $\boldsymbol{\Phi}$ or $\boldsymbol{\lambda}$, and one of $\boldsymbol{\mu}$ or $\boldsymbol{\tau}$ need to be fixed. In PISA, $\boldsymbol{\Phi} = \boldsymbol{I}$ and $\boldsymbol{\mu} = \boldsymbol{0}$ were set for the student responses of the OECD countries in the first cycle. In addition, PISA employs a unidimensional IRT model; therefore, $\boldsymbol{\Phi}$ was the scalar value of 1. Another way to avoid scale indeterminacy is to fix $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$. In PBTS, $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$ of anchor items are fixed at the values of the parameter estimates defined in PISA 2015, instead of fixing $\boldsymbol{\Phi} = \boldsymbol{I}$ and $\boldsymbol{\mu} = \boldsymbol{0}$ since the populations of PBTS were different from PISA 2015. Since PISA 2015, PISA fixes $\boldsymbol{\lambda}$ and $\boldsymbol{\tau}$ of the trend items to avoid scale indeterminacy and estimate the scores on the same scale as the past testing cycles.

Consider the following linear transformation with respect to parameters $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{\theta}$.

Equation 11

$$\begin{aligned}\boldsymbol{a}_j^* &= \boldsymbol{K}^{-1}\boldsymbol{a}_j \\ \boldsymbol{b}_j^* &= \boldsymbol{K}\boldsymbol{b}_j + \boldsymbol{l} \\ \boldsymbol{\theta}_i^* &= \boldsymbol{K}\boldsymbol{\theta}_i + \boldsymbol{l}\end{aligned}$$

The transformation coefficients $\boldsymbol{K}$ and $\boldsymbol{l}$ are the size $M \times M$ diagonal matrix and the size $M \times 1$ vector, respectively.

$$\boldsymbol{K} = \begin{bmatrix} K_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K_M \end{bmatrix}$$

$$\boldsymbol{l} = \begin{bmatrix} l_1 \\ \vdots \\ l_M \end{bmatrix}$$

This linear transformation (Equation 11) holds the same probability density function as Equation 7. This mathematical feature is utilised in various situations in educational assessments, such as tests linking and differential item functioning (DIF) analysis (Asparouhov, 2014[22]). In PISA, K_m and l_m are applied for all m in the multiple imputations phase so that all the PISA scores are represented on an $N(500, 100)$ scale.

2.3. Item response models

Here, the item response model presented in Equation 9 is discussed. In PISA, the item category response functions (ICRF) for category k of item j are computed according to the following unidimensional logistic model.

Equation 12

$$p_j(x_j = k|\theta) = p_{jk}(\theta) = \frac{\exp\left(\sum_{k'=0}^{k} l_{jk'}(\theta)\right)}{\sum_{k'=0}^{K_j-1}\left(\exp\left(\sum_{k''=0}^{k'} l_{jk''}(\theta)\right)\right)}$$

where

Equation 13

$$\begin{aligned} l_{jk}(\theta) &= Da_j(\theta - b_j + d_{jk}) \\ &= Da_j(\theta - b_{jk}) \end{aligned}$$

When $k = 0$, $l_{j0}(\theta)$ is fixed at a constant (e.g., 0). a_j and b_j are the slope parameter and the location parameter of item j, respectively. d_{jk} is the step parameter of category k. Note that $d_{j1} = 0$ when $K_j = 2$. D is the scale constant which is set to 1.7 in PISA cognitive items. The item response model (Equation 12) is called the generalised partial credit model (Muraki, 1992[23]). A special case of the generalised partial credit model (GPCM) is called the partial credit model (Masters, 1982[24]), in which a_j is fixed to 1.0. In the PISA technical reports, GPCM for binary data is called the 2-parameter model, while the partial credit model (PCM) for binary data is called the Rasch model (Rasch, 1960[25]). The above parametrisation is used in the item scaling phase in PISA. Together with the nominal categories model (Bock, 1972[26]) in which subscript k is attached to a_j (i.e., a_{jk}), the models are called divide-by-total models (Thissen, 1986[27]). While the normal ogive model was introduced in the previous sections, the same logic can be applied for the logistic model.

Equation 13 can be replaced with the following parametrisation during the item scaling phase

Equation 14

$$l_{jk}(\theta) = \alpha_j\theta - \beta_{jk}$$

where $\beta_{j0} = 0$. α_j and β_{jk} are called factor loadings and thresholds, respectively. Note that both parameterisations (Equation 13 and Equation 14) are mathematically equivalent and can be converted from one to the other. In this paper, factor loadings or slope parameters are represented by α, and thresholds or location parameters and step parameters are denoted by β.

Another family of IRT models is called difference models, where the ICRF is defined with the difference of the adjacent boundary response function $p^*_{jk}(\theta)$, namely:

$$p_{jk}(\theta) = p^*_{jk}(\theta) - p^*_{j(k+1)}(\theta)$$

where

$$\begin{aligned} p^*_{jk}(\theta) &= \int_{-\infty}^{a_j(\theta - b_{jk})} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) \mathrm{d}v \\ &\cong \frac{1}{1 + \exp\left(Da_j(\theta - b_{jk})\right)} \end{aligned}$$

The boundary response function $p^*_{jk}(\theta)$ describes a probability of selecting category k or higher categories than k. Therefore, $b_{j0} = -\infty$ and $b_{jK_j} = +\infty$; hence, $p^*_{j0}(\theta) = 1$ and $p^*_{jK_j}(\theta) = 0$. In case of $K_j \geq 3$, the model is called graded response model (Samejima,

1969[28]). The merit of using the graded response model (GRM) is that b_{jk} is independently defined from $b_{jk'}$. Therefore, GRM is able to equate the items with a different number of categories, while GPCM is not. This model feature is ideal, especially for Likert scale items.

2.4. Parameter constraints

In ILSAs, a subscript $g(=1,\ldots,G)$ is attached to α_j and β_j. Mathematically, this implies that the item parameters of a country (i.e., α_{gj} and β_{gj}) can be different from those of another country (i.e., $\alpha_{g'j}$ and $\beta_{g'j}$); on the other hand, the item parameters should be on the same scale to ensure comparability across countries/economies. From a psychometric perspective, the comparability of a latent variable θ across countries/economies is ensured by the equivalent parameters $\boldsymbol{\alpha}_g$, $\boldsymbol{\beta}_g$ and $\boldsymbol{\Phi}_g$ across g.

The equivalence of factors has 4 hierarchical levels (Chen, 2007[29]);

1. Configural invariance: the lowest level of the equivalence in which only $\boldsymbol{\alpha}_g \neq \mathbf{0}$ is held. In general, $\boldsymbol{\alpha}_g > \mathbf{0}$ is assumed in educational assessments.
2. Metric invariance: only $\boldsymbol{\alpha}_g = \boldsymbol{\alpha}_{g'}$ is assumed across g.
3. Scalar invariance: is an assumption that holds $\boldsymbol{\alpha}_g = \boldsymbol{\alpha}_{g'}$ and $\boldsymbol{\beta}_g = \boldsymbol{\beta}_{g'}$. It entails the same item functioning among the countries.

In addition to the above three equivalence levels, the following most strict level of item equivalence can be set.

4. Residual invariance: it assumes $\boldsymbol{\alpha}_g = \boldsymbol{\alpha}_{g'}$, $\boldsymbol{\beta}_g = \boldsymbol{\beta}_{g'}$, and $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}_{g'}$, which is also called strict invariance.

Technically, the structural invariance with respect to $\boldsymbol{\theta}$, $\boldsymbol{\Phi}_g = \boldsymbol{\Phi}_{g'}$ and $\boldsymbol{\mu}_g = \boldsymbol{\mu}_{g'}$, can also be examined.

In PISA, scalar invariance is set for the criteria of the item-level equivalence since this ensures comparability of the latent variable θ (Widaman, 1997[30]). The ideal situation in an ILSA is that all items hold the scalar invariance (or above level); however, it is not practical to hold this assumption since there should be some factors that make item functioning different for some of countries/economies, such as linguistic quality control or the curriculum of the domain. In ILSAs, partial invariance is assumed instead of postulating full invariance.

The partial invariance assumption accepts some of the items to be variant across g. Full invariance is not necessarily required as long as the vast majority of the items are invariant and retain the content validity of the factor via the remaining items (Meredith, 1993[31]; Reise, 1993[32]) For the items that do not satisfy the invariance assumption, the equal parameter restrictions are released, and the item parameters of the country/economy are estimated independently from the other countries (i.e., national parameters). The items are excluded from the analysis if the national parameters do not satisfy the configural invariance assumption. The statistical standards for the invariance judgements in PISA are provided in section 3.3. The methodologies for detecting invariant items are discussed in van de Vijver et al. (2019[33]).

3. Model likelihood and parameter estimation

3.1. Model likelihood

In PISA, a student is considered as a random variable drawn from the population that follows the normal distribution with parameters $\boldsymbol{\varphi}_g = [\mu_g, \sigma_g^2]$ that is specific to a country/economy. The item parameters of country/economy g that need to be estimated are expressed as $\boldsymbol{\Lambda}_g = [\boldsymbol{\alpha}_g, \boldsymbol{\beta}_g]$. In section 3, the subscript for a domain m is excluded since the IRT model employed in PISA is a unidimensional IRT model.

Let $\boldsymbol{\Lambda}^* = [\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*]$ be the fixed item parameters of the trend items that define the PISA scale. Note that $\boldsymbol{\Lambda}^*$ can be international parameters, unique parameters, or national parameters, which have been estimated in PISA 2015 (OECD, 2017[10]). By calibrating $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\varphi}_g$ on the scale defined by $\boldsymbol{\Lambda}^*$, the indeterminacy between $\boldsymbol{\varphi}_g$ and $\boldsymbol{\Lambda}_g$ can be avoided since $\boldsymbol{\Lambda}^*$ plays a role of anchor that puts both $\boldsymbol{\mu}$ and $\boldsymbol{\Phi}$ on the scale.

Under the given item parameters $\boldsymbol{\Lambda}^*$, the likelihood function to be maximised in the model[3] is described as

Equation 15

$$L(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}, \boldsymbol{\varphi} \mid \boldsymbol{u}) = \prod_{g=1}^{G} \prod_{i=1}^{N_g} \int_{\theta} f(\boldsymbol{u}_{gi}|\theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g) h(\theta|\boldsymbol{\varphi}_g) \, \mathrm{d}\theta$$

where $f(\boldsymbol{u}_{gi}|\theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g) h(\theta|\boldsymbol{\varphi}_g)$ is the joint distribution of the observed data $\boldsymbol{u}$ and the missing data θ. Here, local independence of the student responses is assumed. u_{ijk} is the dummy coded response of cognitive item j by student i, which is defined as

$$u_{ijk} = \begin{cases} 1, & \text{if } x_{ij} = k \\ 0, & \text{otherwise} \end{cases}$$

Here, the fixed item parameters $\boldsymbol{\Lambda}^*$ are not the parameters to be estimated; they are presented in order to show explicitly that the likelihood depends on $\boldsymbol{\Lambda}^*$. The observed data part can be described as

Equation 16

$$f(\boldsymbol{u}_{gi}|\theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g) = \prod_{j=1}^{J} \prod_{k=0}^{K_j - 1} p_{gjk}(\theta)^{u_{ijk}}$$

Note that subscript j is nested within a domain m; hence, the likelihood function is independent of domains. In addition, subscript i is nested within a country/economy g.

3.2. Parameter estimation

Equation 15 cannot be maximised directly; therefore, an iterative procedure, EM-algorithm (Dempster, 1977[34]), is employed to maximise the (log)likelihood function in which $\boldsymbol{\varphi}_g$ and $\boldsymbol{\Lambda}_g$ are updated iteratively by replacing the missing data with the expectation based on the provisional parameters $\boldsymbol{\varphi}_g^{(t)}$ and $\boldsymbol{\Lambda}_g^{(t)}$ in the t-th cycle of expectation (E) and maximisation (M) iteration.

[3] In this paper, student weights are not considered.

Here, the parameter optimisation method, the marginalised maximum-likelihood estimation (MMLE) via the EM-algorithm (Bock, 1981[35]), is introduced. The marginalised maximum-likelihood estimator of IRT modelling has consistency, while the joint maximum-likelihood estimation (Birnbaum, 1968[36]; Albert, 1992[37]) does not (Hambleton, 1985[38]; Harwell, 1991[39]). In ILSAs, new items are developed every cycle and inserted into the booklets in addition to the existing items. The psychometric properties of the new items are verified in the field trials prior to the main study, and parameters $\boldsymbol{\Lambda} = [\boldsymbol{\alpha}, \boldsymbol{\beta}]$ are estimated on the scale defined by $\boldsymbol{\Lambda}^*$ in the main study.

3.2.1. E-step

In E-steps, the conditional distribution of the missing data, conditional on the provisional item parameters $\boldsymbol{\Lambda}_g^{(t)}$ and fixed-parameters $\boldsymbol{\Lambda}^*$, is calculated:

Equation 17

$$h_{gi}^{(t)}(\theta) = h_{gi}\left(\theta|\boldsymbol{u}_{gi}, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}, \boldsymbol{\varphi}_g^{(t)}\right) = \frac{f\left(\boldsymbol{u}_{gi}|\theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}\right) h\left(\theta|\boldsymbol{\varphi}_g^{(t)}\right)}{\int_\theta f\left(\boldsymbol{u}_{gi}|\theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}\right) h\left(\theta|\boldsymbol{\varphi}_g^{(t)}\right) \mathrm{d}\theta}$$

Note $\boldsymbol{\Lambda}_g^{(0)}$ are the initial values of the item parameters. The expected frequency for category k of item j are calculated using the conditional distribution of θ defined in Equation 17.

$$F_{gjk}^{(t)}(\theta) = \sum_{i=1}^{N_g} u_{gijk} h_{gi}^{(t)}(\theta)$$

The student responses that appear in Equation 16 are replaced with the expectation $F_{gjk}^{(t)}(\theta)$ calculated in the cycle of the EM-algorithm.; thus,

Equation 18

$$\begin{aligned} \mathrm{E} \ln L_g\left(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}, \boldsymbol{\varphi}_g^{(t)}|\boldsymbol{u}_{gi}\right) &= \int_\theta \sum_{j=1}^{J} \sum_{k=0}^{K_j-1} F_{gjk}^{(t)}(\theta) \ln p_{gjk}^{(t)}(\theta) \, \mathrm{d}\theta \\ &\approx \sum_{q=1}^{Q} \sum_{j=1}^{J} \sum_{k=0}^{K_j-1} F_{gjk}^{(t)}(\theta_q) \ln p_{gjk}^{(t)}(\theta_q) \end{aligned}$$

Note that Equation 18 holds $\mathrm{E} \ln L = \sum_{g=1}^{G} \mathrm{E} \ln L_g$. At this point, $\mathrm{E} \ln L_g$ is marginalised with respect to θ. In practice, numerical integration is employed during the EM cycles. Equation 18 is approximated with Q quadrature points on the θ scale, and Equation 17 is evaluated at each q $(= 1, \ldots, Q)$.

M-step for $\boldsymbol{\Lambda}$

M-step is composed of two parts; one for $\boldsymbol{\Lambda}$ and another for $\boldsymbol{\varphi}$; in each M-step, $\boldsymbol{\Lambda}^{(t)}$ is updated first, then $\boldsymbol{\varphi}^{(t)}$. In ILSAs, item parameters $\boldsymbol{\Lambda}_{gj}$ are imposed to be equal ($\boldsymbol{\Lambda}_{gj} = \boldsymbol{\Lambda}_j$) across countries in order to set scalar invariance of the factor. Here, the procedure for estimating $\boldsymbol{\Lambda}$ is introduced.

Let $v\ (=1,\dots,V)$ be the index of parameter sets[4] which composes parameters that need to be estimated together, $r\ (=1,\dots,R_v)$ be the r-th parameter in parameter set v. Usually, v is a cluster of countries that share the same item parameters; thus, $v=j$ if $\boldsymbol{\Lambda}_{gj}=\boldsymbol{\Lambda}_j$. Λ_{vr} denotes the r-th parameter (i.e., α_j or β_{jk} in GPCM) in parameter set v. For example, if metric invariance is assumed between countries g and h for item j, the parameters in v are $\boldsymbol{\Lambda}_v=[\alpha_{(g=h)j},\beta_{gj0},\dots,\beta_{gj(k_j-1)},\beta_{hj0},\dots,\beta_{hj(k_j-1)}]'$. In M-step for $\boldsymbol{\Lambda}$, the parameters $\boldsymbol{\Lambda}^{(t)}$ are updated for all the sets of parameters parallelly or sequentially.

In M-step for $\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}_v$ for all v are updated independently by the Newton-Raphson method,

Equation 19

$$\boldsymbol{\Lambda}_v^{((n+1):t)}=\boldsymbol{\Lambda}_v^{(n:t)}-\boldsymbol{H}_v^{-1}\boldsymbol{g}_v$$

where index $n{:}\,t$ denotes the n-th iteration of the Newton-Raphson method in the t-th cycle of the EM-algorithm. $\boldsymbol{g}_v$ is the size $R_v\times 1$ gradient vector of parameter set v, which contains the 1st derivative functions with respect to the parameters.

$$\boldsymbol{g}_v=\frac{\partial\,\mathrm{E}\ln L}{\partial\boldsymbol{\Lambda}_v}$$

$\boldsymbol{H}_v$ is the size $R_v\times R_v$ Hessian matrix.

$$\boldsymbol{H}_v=\frac{\partial^2\,\mathrm{E}\ln L}{\partial\boldsymbol{\Lambda}_v\partial\boldsymbol{\Lambda}_v'}$$

In order to calculate the derivative functions of the parameters under equal constraints among countries, the following transformation is applied.

Equation 20

$$\frac{\partial\,\mathrm{E}\ln L}{\partial\Lambda_{vr}}=\sum_{g=1}^{G}\sum_{j=1}^{J}t_{vr,gjs}\sum_{k=0}^{K_j-1}\sum_{q=1}^{Q}F_{gjk}^{(t)}(\theta_q)\frac{\partial\ln p_{gjk}^{(t)}(\theta_q)}{\partial\Lambda_{gjs}}$$

$$\frac{\partial^2\,\mathrm{E}\ln L}{\partial\Lambda_{vr}\partial\Lambda_{vr'}}=\sum_{g=1}^{G}\sum_{j=1}^{J}t_{vr,gjs}t_{vr',gjs'}\sum_{k=0}^{K_j-1}\sum_{q=1}^{Q}F_{gjk}^{(t)}(\theta_q)\frac{\partial^2\ln p_{gjk}^{(t)}(\theta_q)}{\partial\Lambda_{gjs}\partial\Lambda_{gjs'}}$$

where Λ_{gjs} is the s-th $(=1,\dots,S_j)$ parameter of item j in country g, and $t_{vr,gjs}$ is an index that takes 1 if the s-th parameter of item j in country g belongs to the r-th parameter of parameter set v; otherwise, $t_{vr,gjs}=0$. Here, $r'(=1,\dots,R_v)$ and $s'(=1,\dots,S_j)$ are the same subscripts as r and s, respectively. In most cases, equality parameter constrains are not set among items in ILSAs. A parameter that holds $\sum_{g=1}^{G}t_{vr,gjs}=1$ is called national parameter, whereas a parameter that holds $\sum_{g=1}^{G}t_{vr,gjs}\geq 2$ is called unique parameter.

The statistical modelling that treats $t_{vr,gjs}$ as a random variable is called latent class modelling (Lazarsfeld, 1968[40]), where the information that a country belongs to a country cluster is provided as the membership probability. This modelling approach can be employed when identifying groups of unique parameters or detecting a group of students who have particular factor structures that are different from the vast majority of students.

[4] If no parameter constraints are set among items, v can be considered as a country-cluster that shares the same item parameters.

Hence, the latent class modelling (LCM) can be applied in the context of the residual analysis mentioned in section 2.1.

$\boldsymbol{\Lambda}_v^{(t)}$ is updated to $\boldsymbol{\Lambda}_v^{(t+1)}$ if $\boldsymbol{\Lambda}_v^{((n+1):t)} - \boldsymbol{\Lambda}_v^{(n:t)}$ satisfies the convergence criteria. In M-step for $\boldsymbol{\Lambda}$, V sets of the item parameters are updated parallelly[5] or sequentially in each M-step.

M-step for $\boldsymbol{\varphi}$

In an M-step within the t-th EM cycles, the country/economy parameter $\boldsymbol{\varphi}_g^{(t)}$ is updated based on the fixed item parameters $\boldsymbol{\Lambda}^*$ and the item parameters $\boldsymbol{\Lambda}^{(t+1)}$ that have just been updated within the same t-th EM cycle. In a unidimensional IRT model, $\boldsymbol{\varphi}_g$ composes μ_g and σ_g^2; thus, $\boldsymbol{\varphi}_g$ are updated as follows:

Equation 21

$$\mu_g^{(t+1)} = \frac{1}{N_g}\int_\theta \theta \sum_{i=1}^{N_g} h_{gi}\left(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t+1)}\right) \mathrm{d}\theta$$

$$\sigma_g^{2(t+1)} = \frac{1}{N_g}\int_\theta \left(\theta - \mu_g^{(t+1)}\right)^2 \sum_{i=1}^{N_g} h_{gi}\left(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t+1)}\right) \mathrm{d}\theta$$

Note that h_{gi} appearing in Equation 17 is replaced with the one based on $\boldsymbol{\Lambda}^{(t+1)}$ when calculating Equation 21. The above updates can be computed parallelly for all g. Thus, the updated $\boldsymbol{\varphi}_g$ for all g are obtained.

After updating $\boldsymbol{\varphi}_g$, $\mathrm{E}\ln L^{(t+1)} - \mathrm{E}\ln L^{(t)}$ based on $\boldsymbol{\Lambda}^{(t+1)}$ and $\boldsymbol{\varphi}_g^{(t+1)}$ is computed in order to evaluate the convergence of the optimisation. The iteration is completed if it reaches the convergence criteria; otherwise, another cycle of the E and M computations is iterated. It is important to run the EM cycles multiple times with different sets of initial values of $\boldsymbol{\Lambda}^{(0)}$, especially when a latent class structure is involved in the model.

3.3. Model evaluation

Various goodness-of-fit indices that are used in the framework of SEM, such as the root mean square error of approximation (RMSEA) and the comparative fit index (CFI), can be employed to evaluate the model-data fit[6] of the IRT modelling, although some require estimators other than the MLE. In addition to evaluating the goodness-of-fit indices, it is important to scrutinise the residual structure of the model. In particular, conducting the stratified analysis for the residual covariance/variance $\widehat{\boldsymbol{\Psi}}$ is critical in ILSAs. Moreover, monitoring $\mathrm{Cov}\left[\widehat{\boldsymbol{\theta}}, \hat{\boldsymbol{e}}\right]$ stratified by student backgrounds such as gender, index of economic, social and cultural status (ESCS), and school information (e.g., rural-urban) is also essential.

In ILSAs, once the construct validity of the assessment is verified in the pilots/trials, the main psychometric interest shifts towards the score comparability across the countries/groups and the testing cycles. In PISA, the root mean square deviation (RMSD) index is employed to assess the item equivalence (item functioning) of item j in country g against the international parameters. RMSD measures the discrepancy between the

[5] Parallel computing can be implemented in this step, which makes convergence speed faster.

[6] The goodness of fit index tends to be poor since the number of observed variables is large in multi-group factor analysis modelling.

expected category response probability based on the fixed item parameters and the observed pseudo frequencies appearing in EM cycles.

$$\mathrm{RMSD}_{gj} = \frac{1}{K_j}\sum_{k=0}^{K_j-1}\sqrt{\int_{\theta}\left(o_{gjk}(\theta) - p_{gjk}(\theta)\right)^2 f_g(\theta)\,\mathrm{d}\theta}$$

where $f_g(\theta)$ is the proficiency distribution of the population and replaced with $N(\hat{\mu}_g, \hat{\sigma}_g^2)$. Here, $o_{gjk}(\theta)$ is the pseudo observed frequency calculated based on the conditional distribution of θ defined in Equation 17; namely,

$$o_{gjk}(\theta) = \frac{\sum_{i=1}^{N_g} u_{gijk} h_{gi}(\theta|\boldsymbol{\varphi}_g)}{\sum_{k=0}^{K_j-1}\sum_{i=1}^{N_g} u_{gijk} h_{gi}(\theta|\boldsymbol{\varphi}_g)}$$

In PISA, the value of 0.12 is set as the cut-off criterion for RMSD. Items that exceed the thresholds will be investigated for the possibility of differential item functioning (DIF) by looking into the translated and adapted materials. In practice, the threshold is gradually reduced from 0.30, 0.18, and then 0.12, in which the parameter constraints for the DIF items are released in each of the steps (i.e., updating $t_{vr,gjs}$ in each step).

The DIF check appears in two different contexts; one is in examining the model-data fit against the international parameters (horizontal comparability), and the other is in checking the discrepancy of the trend items between cycles (vertical comparability). The model evaluation procedure in the item scaling phase is the crucial part of scaling ILSAs.

3.4. Parameter constraints and likelihood function

Scale comparability across testing cycles is ensured by the unchanged characteristics of the trend items and their parameters $\boldsymbol{\Lambda}^*$ that are invariant across testing cycles. In PISA 2015, sets of item parameters $\boldsymbol{\Lambda}^* = [\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*]$ of all countries/economies were re-calibrated (OECD, 2017[10]), where most of them are common among the groups (i.e., treated as international parameters) and some were estimated as country-specific (i.e., national parameters) or group-specific (i.e., unique parameters). Thus, fixed-parameters $\boldsymbol{\Lambda}^*$ secure the comparability and continuity of ILSAs. In this section, the impact of parameter constraints on the likelihood function is discussed.

3.4.1. Model without horizontal parameter constraints

Let all j hold $\sum_{g=1}^{G} t_{vr,gjs} = 1$; namely, $\boldsymbol{\Lambda}_{gj}$ is independent among countries for all item j. Furthermore, suppose the parameters of the trend items $\boldsymbol{\Lambda}^*$ are given, and that the model-data fit (e.g., RMSD index) for the trend items are good enough. In this case, the likelihood function with respect to $\boldsymbol{\varphi}_g$ and $\boldsymbol{\Lambda}_g$ can be described as follows.

Equation 22

$$L_g(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g, \boldsymbol{\varphi}_g | \boldsymbol{u}_g) = \prod_{i=1}^{N_g}\int_{\theta} f(\boldsymbol{u}_{gi}|\theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g) h(\theta|\boldsymbol{\varphi}_g)\,\mathrm{d}\theta$$

The likelihood function of country g, $L_g(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g, \boldsymbol{\varphi}_g | \boldsymbol{u}_g)$, is not confounding with the data or the parameters of any other countries g' ($g' \neq g$). Equation 22 is maximised through the EM-algorithm introduced in section 3.2. Since $\sum_{g=1}^{G} t_{vr,gjs} = 1$ for all j in this case, the gradient $\boldsymbol{g}_v$ and the Hessian matrix $\boldsymbol{H}_v$ shown in Equation 19 can be calculated based on E ln L_g; namely, $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\varphi}_g$ are estimated independently from parameters $\boldsymbol{\Lambda}_{g'}$ and $\boldsymbol{\varphi}_{g'}$.

Furthermore, Equation 22 shows the number of countries G and the other countries' student responses $\boldsymbol{u}_{g'}$ are no longer involved in $L_g(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g, \boldsymbol{\varphi}_g | \boldsymbol{u}_g)$, which indicates G and $\boldsymbol{u}_{g'}$ in the model do not affect $\widehat{\boldsymbol{\Lambda}}_g$ and $\widehat{\boldsymbol{\varphi}}_g$ in case $\boldsymbol{\Lambda}_g$ is independent across countries.

Since the model has the scale indeterminacy between $\boldsymbol{\varphi}$ and $\boldsymbol{\Lambda}$, the point estimates of $\boldsymbol{\varphi}_g$ are essentially determined by $\boldsymbol{\Lambda}^*$ and their corresponding responses $\boldsymbol{u}_g$. During the EM cycles, $\boldsymbol{\Lambda}_g^{(t)}$ and $\boldsymbol{\varphi}_g^{(t)}$ are updated consistently on the scale calibrated on $\boldsymbol{\Lambda}^*$; thus, $\boldsymbol{\Lambda}_g$ does not contribute to the point estimates of $\boldsymbol{\varphi}_g$ if $\widehat{\boldsymbol{\Lambda}}_g$ reach to the local maximum of the (log)likelihood function given by student responses $\boldsymbol{u}_g$ and the fixed item parameters $\boldsymbol{\Lambda}^*$. $\boldsymbol{\Lambda}_g$ and the corresponding responses provide additional information to the likelihood, increasing thereby the reliability and validity of $\widehat{\boldsymbol{\varphi}}_g$; furthermore, it contributes to the student-level information. Consequently, $\widehat{\boldsymbol{\Lambda}}_g$ and $\widehat{\boldsymbol{\varphi}}_g$ are not affected by $\widehat{\boldsymbol{\Lambda}}_{g'}$ and $\widehat{\boldsymbol{\varphi}}_{g'}$ under the situation that any equality parameter constraints are not set.

3.4.2. Model with horizontal parameter constraints

Suppose $\sum_{g=1}^{G} t_{gjrv} \geq 2$ in item j, which implies two or more countries/economies share the same item parameters. Unlike Equation 22, $\boldsymbol{\Lambda}_{gj}$ and some of $\boldsymbol{\Lambda}_{g'j}$ are imposed to be equal, and the gradient $\boldsymbol{g}_v$ and the Hessian matrix $\boldsymbol{H}_v$ are computed as shown in Equation 20. Therefore, $\widehat{\boldsymbol{\Lambda}}_v$ is influenced by $\partial \, \mathrm{E} \ln L_g / \partial \boldsymbol{\Lambda}_{gj}$ and $\partial^2 \, \mathrm{E} \ln L_g / \partial \boldsymbol{\Lambda}_{gj} \, \partial \boldsymbol{\Lambda}_{gj}$ of all g in v. The impacts of $\widehat{\boldsymbol{\Lambda}}_{gj}$ to $\widehat{\boldsymbol{\Lambda}}_v$ are approximately proportional to the number of students involved in the likelihood function; hence, it is important to take the number of students of countries/economies into account when estimating parameters in ILSAs. In PISA, senate weights are employed in the item scaling phase (OECD, n.d.[41]), which implies, on the other hand, that representativeness of the population is not a necessary condition when estimating item parameters.

Consider the situation that $\boldsymbol{u}_{gj}$ ($\boldsymbol{o}_{gj}(\theta)$) of all countries/economies that belong to v fit all the elements of $\widehat{\boldsymbol{\Lambda}}_v$ perfectly, which indicates $\widehat{\boldsymbol{\Lambda}}_v$ is equal to $\widehat{\boldsymbol{\Lambda}}_{gj}$ in the model without horizontal parameter constraints (i.e., independent model). In this case, no bias is introduced for $\widehat{\boldsymbol{\varphi}}_g$ of all countries/economies in v since $\widehat{\boldsymbol{\Lambda}}_v$ and the model-data fit are the same as that of the independent model. However, if student responses to item j in country g^*, $\boldsymbol{u}_{g^*j}$ ($\boldsymbol{o}_{g^*j}(\theta)$), do not follow $p_j(\theta | \widehat{\boldsymbol{\Lambda}}_v)$, then $\widehat{\boldsymbol{\Lambda}}_v$ is biased toward $\boldsymbol{\Lambda}_{g^*j}$ that is not estimated in the model actually. In this case, $\boldsymbol{\Lambda}_{g^*j}$ should be estimated independently from the parameter set v; otherwise, $\widehat{\boldsymbol{\varphi}}_g$ of all countries in parameter set v are deviated from the true parameters $\boldsymbol{\varphi}_g$ because of the biased $\widehat{\boldsymbol{\Lambda}}_v$. The size of bias is larger in $\widehat{\boldsymbol{\varphi}}_{g^*}$ and less in the other $\widehat{\boldsymbol{\varphi}}_g$ if majority of g follow $p_j(\theta | \widehat{\boldsymbol{\Lambda}}_v)$. The impact to $\widehat{\boldsymbol{\varphi}}_g$ depends on the values of $\boldsymbol{\Lambda}_{g^*j} - \boldsymbol{\Lambda}_v$.

Note that DIF itself is not an issue; however, imposing inappropriate parameter constraints of the model is the issue. In the above-mentioned case, there will be no bias for $\widehat{\boldsymbol{\varphi}}_g$ of all countries in v if country g^* is excluded from v and the parameters $\boldsymbol{\Lambda}_{g^*j}$ are estimated independently from the other countries in v or set as unique parameters for some countries that have the equivalent item functioning with country g^*. Therefore, setting the appropriate parameter constraints $t_{vr,gjs}$ is essential in ILSAs. The cut-offs of RMSD should be carefully considered in order to balance the reliability, validity, and comparability of the assessments. If the parameter constraints are set appropriately and the model assumptions of IRT model are fulfilled, it would not jeopardise the reliability and the validity of the scales.

4. Latent regression model and prior distribution

4.1. Auxiliary information and prior distribution

In the item scaling phase, item parameters are estimated (or validated) in conjunction with the country/economy parameters on the PISA scale using the unidimensional IRT models where $\boldsymbol{\Lambda}_m$ and $\boldsymbol{\varphi}_m$ are estimated for each domain separately. In ILSAs, the student-level proficiencies are provided together with the student background information in a dataset for further data analysis. In order to produce a valid, consistent, and unbiased dataset in terms of student proficiencies of the population, the prior distribution of $\boldsymbol{\theta}_i$ is applied for each student's likelihood function with respect to θ_{im} giving the posterior distribution from which random values are drawn to perform multiple imputations. The main role of the prior distribution in ILSAs is to provide an unbiased dataset with regard to the student proficiencies of students who did not take any items in a domain due to the booklet design of the assessment.

The prior distribution of each student is inferred based on auxiliary information taken from the student questionnaires. In the prior distribution inference phase, the hyper-parameters of the prior distribution of each student are estimated using the latent regression model (LRM), in which student proficiencies are regressed on the students' covariates calculated from their responses to the student questionnaire. The approach of estimating parameters of prior distributions of latent variables based on the observed data is called Empirical Bayes. Note the LRM is performed for each country/economy separately in PISA; thus, country index g is excluded from the equations in section 4.

4.2. Latent regression model

In PISA, the student's responses to the student questionnaire and the indices that are calculated based on the responses are compressed into S covariates using the principal component analysis, and the compressed variables and some direct variables such as gender are used as covariates in order to estimate the hyper-parameters of each student's proficiencies. Covariates are denoted as $\boldsymbol{y}$, containing S centred random variables and the constant for intercept. Let $\boldsymbol{t}$ and $\boldsymbol{v}$ be the structural variable vector and the residual variable vector, respectively, and let $\boldsymbol{\Delta}$ be the coefficients matrix of the structural equation modelling. Here, $\boldsymbol{\theta}$ and $\boldsymbol{z}$ are the endogenous variables, while $\boldsymbol{d}$, $\boldsymbol{e}$, and $\boldsymbol{y}$ are the exogenous variables. The structural equation of the LRM is described as follows:

Equation 23

$$\boldsymbol{t} = \boldsymbol{\Delta t} + \boldsymbol{v}$$
$$\begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{z} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \boldsymbol{\Gamma} \\ \boldsymbol{\lambda}^* & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{z} \\ \boldsymbol{y} \end{bmatrix} + \begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{e} \\ \boldsymbol{y} \end{bmatrix}$$

where $\boldsymbol{\theta}$ is the $M \times 1$ random variables regarding student proficiencies and $\boldsymbol{\Gamma}$ is the $M \times (S+1)$ matrix that contains the regression parameters. Here, the second row represents the measurement model defined in Equation 1 where $\boldsymbol{\lambda}$ is fixed at $\boldsymbol{\lambda}^*$ estimated in the item scaling phase. Hence, the constructs measured with $\boldsymbol{\Gamma}$ and $\boldsymbol{y}$ are ensured to be the same as the $\boldsymbol{\theta}$ calibrated with $\boldsymbol{\Lambda}$ and $\boldsymbol{x}$. In PISA, $\boldsymbol{\lambda}^*$ is modelled as

Equation 24

$$\boldsymbol{\lambda}^* = \begin{bmatrix} \boldsymbol{\lambda}_1^* & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\lambda}_M^* \end{bmatrix}$$

In this model, the residual covariance $\boldsymbol{\Upsilon}$ is modelled as

Equation 25

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathrm{V}[\boldsymbol{y}] \end{bmatrix}$$

and the model covariance structure $\boldsymbol{\Xi}$ is

Equation 26

$$\boldsymbol{\Xi} = (\boldsymbol{I} - \boldsymbol{\Delta})^{-1}\boldsymbol{\Upsilon}((\boldsymbol{I} - \boldsymbol{\Delta})^{-1})'$$

which shows each element of $\boldsymbol{\Xi}$ can be calculated based on the residual matrix, the structural variables, and the coefficients matrix. The endogenous variables cannot specify their covariance structure directly in the model; hence, the covariance structure $\boldsymbol{\Xi}$ is modelled with $\boldsymbol{\Delta}$ and $\boldsymbol{\Upsilon}$.

Note the model assumes

Equation 27

$$\mathrm{E}[\boldsymbol{d}] = \mathbf{0}$$
$$\mathrm{E}[\boldsymbol{e}] = \mathbf{0}$$

Based on the model shown in Equation 23 and Equation 25, the residual variance of the measurement equation is described as

Equation 28

$$\boldsymbol{\theta} - \boldsymbol{\Gamma}\boldsymbol{y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

Here, the marginalised data distribution is described as follows:

Equation 29

$$p(\boldsymbol{x}_i|\boldsymbol{\Lambda}, \boldsymbol{y}_i, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) = \int_{\boldsymbol{\theta}_i} p(\boldsymbol{x}_i|\boldsymbol{\theta}_i, \boldsymbol{\Lambda})\, h(\boldsymbol{\theta}_i|\boldsymbol{y}_i, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})\, \mathrm{d}\boldsymbol{\theta}_i$$

The data distribution does not contain $\boldsymbol{\theta}$; therefore, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ can be estimated by maximising the likelihood function via the EM-algorithm.

4.3. Multivariate prior distribution

Equation 29 is formulated as a M-dimensional model; however, evaluating the M-dimensional expectation at quadrature points in each of the EM cycles requires enormous amounts of calculation[7]. In order to ease the computation burden under the model condition shown in Equation 24, the LRM parameters $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_m, \dots, \boldsymbol{\gamma}_M]$ of each domain are estimated separately, and $\boldsymbol{\Sigma}$ is calculated subsequently based on the estimated $\boldsymbol{\theta}_m$, $\boldsymbol{\gamma}_m$, and $\boldsymbol{y}_m$of all domains and sub-domains.

From Equation 26 and the assumption $\mathrm{Cov}[\boldsymbol{\Gamma}\boldsymbol{y}, \boldsymbol{d}] = \mathbf{0}$, $\boldsymbol{\Sigma}$ is formulated as follows:

$$\mathrm{V}[\boldsymbol{\Gamma}\boldsymbol{y} + \boldsymbol{d}] = \mathrm{V}[\boldsymbol{\theta}]$$
$$\mathrm{V}[\boldsymbol{\Gamma}\boldsymbol{y}] + \boldsymbol{\Sigma} = \mathrm{V}[\boldsymbol{\theta}]$$

[7] For example, 51 quadrature points for 5 dimensions requires $51^5 = 345{,}025{,}251$ calculation for each of EM cycles.

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N}(\boldsymbol{\theta}\boldsymbol{\theta}' - \boldsymbol{\Gamma}\boldsymbol{y}\boldsymbol{y}'\boldsymbol{\Gamma}')$$

With the estimated $\boldsymbol{\Sigma}$, the M-dimensional prior distribution with respect to $\boldsymbol{\theta}$ of student i is described as follows:

Equation 30

$$\boldsymbol{\theta}_i \sim N(\widehat{\boldsymbol{\Gamma}}\boldsymbol{y}_i, \widehat{\boldsymbol{\Sigma}})$$

The important feature of Equation 30 is that, unlike the prior mean, $\widehat{\boldsymbol{\Sigma}}$ is common to all i in a country/economy and does not depend on θ. Therefore, the variance of the prior distribution is constant regardless of the student proficiencies.

5. Population modelling and plausible values draw

5.1. Multiple imputations

The final phase of the PISA scaling procedure is the multiple imputations phase, in which the plausible values (PVs) are generated for each student. The advantages of using PVs instead of point estimates are discussed in Mislevy (1987[42]; 1991[43]) and Rubin (1987[44]). The PVs are drawn from the M-dimensional posterior distribution where M depends on the number of domains in which the country participated[8] (OECD, 2017[10]). Furthermore, the PVs of sub-domains in a domain are drawn together with the other remaining domains. For example, the PVs of MATH sub-domains are drawn together with READ and SCIE domains; thus, the PVs are drawn from a 5 $(= 3 + 2)$ dimensional distribution (OECD, n.d.[41]).

The posterior distribution of student i is proportional to the product of the M independent likelihood functions $f(\boldsymbol{x}_{im}|\theta_{im}, \boldsymbol{\Lambda}_m)$ and the M -dimensional prior distribution $h(\boldsymbol{\theta}_i|\boldsymbol{y}_i, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$ shown in Equation 30.

Equation 31

$$p(\boldsymbol{\theta}_i|\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto f(\boldsymbol{x}_i|\boldsymbol{\theta}_i, \boldsymbol{\Lambda})\, h(\boldsymbol{\theta}_i|\boldsymbol{y}_i, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$$

Note that $\boldsymbol{\Lambda}$ contains the fixed item parameters $\boldsymbol{\Lambda}^*$ in this section. In PISA, a set of M PVs is drawn 10 times for each student. The process is performed independently by countries/economies. Equation 31 shows that using the point estimates with regard to the student proficiencies underestimates the variance of $\boldsymbol{\theta}$, and thus result in overestimating correlations between other variables.

5.2. Laplace approximation

In practice, it is difficult to evaluate the normalisation constant of Equation 31 and draw random values straightforwardly; therefore, Laplace approximation is applied when performing the multiple imputations. The Laplace approximation is the deterministic method for approximate inference in which a Gaussian approximation to the conditional distribution of a set of continuous variables is calculated.

Let $h(\boldsymbol{\theta})$ be the probability density function of parameters $\boldsymbol{\theta}$, and $h^*(\boldsymbol{\theta})$ be the function that is proportional to $h(\boldsymbol{\theta})$. Laplace approximation aims to find $\boldsymbol{\theta}_0$ and $\boldsymbol{H}$ such that

[8] In PISA 2018, the innovative domain was optional.

$$h(\boldsymbol{\theta}|\boldsymbol{u}) \approx N(\boldsymbol{\theta}_0, -\boldsymbol{H}^{-1})$$

For $\boldsymbol{\theta}_0$, find the mode (i.e., local maximum) of $h^*(\boldsymbol{\theta}|\boldsymbol{u})$ or $\ln h^*(\boldsymbol{\theta}|\boldsymbol{u})$; namely $\tilde{\boldsymbol{\theta}}$, which satisfies

$$\frac{\partial \ln h^*(\boldsymbol{\theta}|\boldsymbol{u})}{\partial \boldsymbol{\theta}} = 0$$

For $\boldsymbol{H}$, compute a truncated Taylor expansion of $\ln h^*(\boldsymbol{\theta})$ centre at the mode

$$\ln h^*(\boldsymbol{\theta}) \approx \ln h^*(\tilde{\boldsymbol{\theta}}) + g(\tilde{\boldsymbol{\theta}})'(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' H(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})$$

where $g(\tilde{\boldsymbol{\theta}})$ is a size $M \times 1$ vector that contains the 1st derivatives of $\ln h^*(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$

$$g(\tilde{\boldsymbol{\theta}}) = \left.\frac{\partial \ln h^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \left[\begin{matrix} \frac{\partial \ln h^*(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln h^*(\boldsymbol{\theta})}{\partial \theta_M} \end{matrix}\right]_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$$

and $H(\tilde{\boldsymbol{\theta}})$ is a size $M \times M$ Hessian matrix

$$H(\tilde{\boldsymbol{\theta}}) = \left.\frac{\partial^2 \ln h^*(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \left[\begin{matrix} \frac{\partial^2 \ln h^*(\boldsymbol{\theta})}{\partial \theta_1\, \partial \theta_1} & \cdots & \frac{\partial^2 \ln h^*(\boldsymbol{\theta})}{\partial \theta_1\, \partial \theta_M} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln h^*(\boldsymbol{\theta})}{\partial \theta_M\, \partial \theta_1} & \cdots & \frac{\partial^2 \ln h^*(\boldsymbol{\theta})}{\partial \theta_M\, \partial \theta_M} \end{matrix}\right]_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$$

Since $\tilde{\boldsymbol{\theta}}$ gives the local maximum of the function, it holds

$$g(\tilde{\boldsymbol{\theta}}) = 0$$

Thus,

$$\ln h^*(\boldsymbol{\theta}) \approx \ln h^*(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' H(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})$$

Since $h^*(\tilde{\boldsymbol{\theta}})$ does not contain $\boldsymbol{\theta}$,

Equation 32

$$\begin{aligned} h^*(\boldsymbol{\theta}) &\approx \exp\left(\ln h^*(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' H(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\right) \\ &= \exp\left(\ln h^*(\tilde{\boldsymbol{\theta}})\right) \times \exp\left(\frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' H(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\right) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \left(-H(\tilde{\boldsymbol{\theta}})\right)(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\right) \end{aligned}$$

Equation 32 shows that the distribution of $\boldsymbol{\theta}$ can be approximated as a multi-dimensional normal distribution with mean $\tilde{\boldsymbol{\theta}}$ and variance $-H(\tilde{\boldsymbol{\theta}})^{-1}$;

$$\boldsymbol{\theta} \sim N\left(\tilde{\boldsymbol{\theta}}, -H(\tilde{\boldsymbol{\theta}})^{-1}\right)$$

This is the general expression of the Laplace approximation.

5.3. Approximated multi-dimensional posterior distribution

In PISA, $h^*(\boldsymbol{\theta})$ is the product of likelihood function $f(\boldsymbol{u}_i|\boldsymbol{\theta}_i,\boldsymbol{\Lambda})$ and prior distribution $h(\boldsymbol{\theta}_i|\boldsymbol{y}_i,\boldsymbol{\Gamma},\boldsymbol{\Sigma})$, and is different from $h^*(\boldsymbol{\theta})$ defined in section 5.2. Therefore, $h^*(\boldsymbol{\theta})$ in the population modelling should be described as

Equation 33

$$h^*(\boldsymbol{\theta}) = f(\boldsymbol{u}_i|\boldsymbol{\theta}_i,\boldsymbol{\Lambda})\, h(\boldsymbol{\theta}_i|\boldsymbol{y}_i,\boldsymbol{\Gamma},\boldsymbol{\Sigma})$$
$$\ln h^*(\boldsymbol{\theta}) = \ln f(\boldsymbol{u}_i|\boldsymbol{\theta}_i,\boldsymbol{\Lambda}) + \ln h(\boldsymbol{\theta}_i|\boldsymbol{y}_i,\boldsymbol{\Gamma},\boldsymbol{\Sigma})$$

Since the likelihood function of each domain is independent of each other, the Hessian matrix of $\ln f(\boldsymbol{u}_i|\boldsymbol{\theta}_i,\boldsymbol{\Lambda})$ is the size $M \times M$ diagonal matrix

$$H_L(\tilde{\boldsymbol{\theta}}) = \begin{bmatrix} \frac{\partial^2 \ln f(\boldsymbol{u}_i|\boldsymbol{\theta}_i,\boldsymbol{\Lambda})}{\partial\theta_{i1}\,\partial\theta_{i1}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial^2 \ln f(\boldsymbol{u}_i|\boldsymbol{\theta}_i,\boldsymbol{\Lambda})}{\partial\theta_{iM}\,\partial\theta_{iM}} \end{bmatrix}_{\boldsymbol{\theta}_i=\tilde{\boldsymbol{\theta}}}$$

Moreover, the Hessian of the prior distribution $h(\boldsymbol{\theta}_i|\boldsymbol{y}_i,\boldsymbol{\Gamma},\boldsymbol{\Sigma})$ is

Equation 34

$$H_h = \frac{\partial^2 \ln h(\boldsymbol{\theta}_i|\boldsymbol{y}_i,\boldsymbol{\Gamma},\boldsymbol{\Sigma})}{\partial\boldsymbol{\theta}_i\,\partial\boldsymbol{\theta}_i'} = -\boldsymbol{\Sigma}^{-1}$$

which is the inverse of the residual variance matrix with a negative sign defined in Equation 25. Therefore, the covariance of the posterior distribution originates only from the prior distribution since the Hessian matrix $H_L(\tilde{\boldsymbol{\theta}})$ is diagonal. Equation 33 indicates that the precision $\mathrm{V}[\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i]$ of the posterior distribution is composed of $H_L(\tilde{\boldsymbol{\theta}})$ and H_h in which the proportion between the two elements depends on $\tilde{\boldsymbol{\theta}}$. The Hessian matrix H_h forms the inverse matrix of the residual covariance matrix, which implies that the off-diagonal elements of $\boldsymbol{\Sigma}$ impact the corresponding diagonal elements of the posterior distribution; consequently, the covariances of the prior distribution improve the precision of $\boldsymbol{\theta}_i$.

One of the advantages of applying population modelling is that it contributes to the reliability of the proficiency distribution of students, especially for students with extreme proficiency levels. In practice, the more the students respond to the items, the more likelihood dominates the posterior distribution.

In order to take the uncertainty of the estimates of the LRM into account, $\boldsymbol{\Gamma}$ is drawn from the following normal distribution when calculating the prior mean of student i

$$\check{\boldsymbol{\Gamma}}_i \sim N(\hat{\boldsymbol{\Gamma}}, \mathrm{V}[\hat{\boldsymbol{\Gamma}}])$$

where $\hat{\boldsymbol{\Gamma}}$ and $\mathrm{V}[\hat{\boldsymbol{\Gamma}}]$ are the point estimate and the estimated variance of $\boldsymbol{\Gamma}$ obtained in the EM cycles (Rubin, 1987[44]), respectively. Further, $\check{\boldsymbol{\Gamma}}_i$ are the drawn regression parameters of a student. Thus, in practice, the prior distribution of student i is described as follows.

$$h(\boldsymbol{\theta}_i|\boldsymbol{y}_i,\check{\boldsymbol{\Gamma}}_i,\boldsymbol{\Sigma}) = N(\check{\boldsymbol{\Gamma}}_i\boldsymbol{y}_i,\boldsymbol{\Sigma})$$

The PVs are drawn from the approximated posterior distribution of a student where the prior distribution $h(\boldsymbol{\theta}_i|\boldsymbol{y}_i,\check{\boldsymbol{\Gamma}}_i,\boldsymbol{\Sigma})$ ensures the correlation structure within the same draws; thus, it is not appropriate to analyse the different sets of PVs in the same model, which results in underestimating the correlation structure. For example, PV1READ should be

analysed together with PV1MATH, not PV2MATH. In addition, using the average of PVs of a domain when performing a statistical analysis is not appropriate for the same reason.

6. Discussions

6.1. Scale robustness

In this section, possible factors that bias the PISA scale are discussed. In particular, the impacts of the growing number of countries/economies and their proficiency distributions in PISA are considered from a mathematical perspective. As a nature of ILSAs, continuous item development is essential to maintain the quality of item bank; therefore, a test that contains both trend items and newly developed items in booklets is discussed in this section. Moreover, it is supposed that the construct validity of the new items is verified already, and that the item functionings of the trend items are confirmed to be invariant from the previous testing cycles.

Under the situation that the parameters of the trend items $\boldsymbol{\Lambda}^*$ are given, the estimates of country parameters $\boldsymbol{\varphi}$ and the parameters of the new items $\boldsymbol{\Lambda}$ are not biased as far as the marginal likelihood estimator is employed and the model assumptions (Equation 3 and Equation 10) are satisfied. As explained in section 3.2, $\boldsymbol{\Lambda}$ and $\boldsymbol{\varphi}$ are estimated on the parameter space defined by $\boldsymbol{\Lambda}^*$; therefore, as far as $\widehat{\boldsymbol{\varphi}}$ and $\widehat{\boldsymbol{\Lambda}}$ are fully optimised on the scale set by $\boldsymbol{\Lambda}^*$ and show sufficient model-data fit, $\widehat{\boldsymbol{\varphi}}$ cannot be the factor to bias $\widehat{\boldsymbol{\Lambda}}$ (and vice versa). Note that $\widehat{\boldsymbol{\Lambda}}$ affects $\widehat{\boldsymbol{\theta}}_i$ and the precision of $\widehat{\boldsymbol{\varphi}}$ since $\widehat{\boldsymbol{\Lambda}}$ and the corresponding student responses $\boldsymbol{u}_j$ contribute to the likelihood. Same as the role of $\boldsymbol{\Lambda}$ in the item scaling phase, the role of national parameters is to keep the validity of the measures and to increase the precision of $\widehat{\boldsymbol{\varphi}}$; thus, the items with national parameters should be kept in further analyses as far as they have sufficient model-data fit. Furthermore, in the population modelling, it is important to keep items with national parameters since they improve the precision of the estimates of the LRM.

However, as shown in section 3.2, $\widehat{\boldsymbol{\Lambda}}$ and $\widehat{\boldsymbol{\varphi}}$ are biased in the item scaling phase if the parameter constraints across countries impair the model-data fit. A factor that causes the bias of $\widehat{\boldsymbol{\varphi}}_g$ is the inappropriate parameter constraint across the countries/economies. From this perspective, fixing slope parameters in the item response models across all items at a particular value (i.e., Rasch model and PCM) is not recommended in ILSAs. In order to avoid inappropriate parameter constraints, it is vital to detect country-DIF accurately. The cut-off criterion of DIF for cognitive items is set at $\mathrm{RMSD}_{gj} > 0.120$ in PISA; however, the validity of the criterion remains controversial. The alignment method (Asparouhov, 2014[22]), on the other hand, omits the DIF judgement of each item, estimates the item parameters based on the configural invariance assumption, and then performs a linear transformation to minimise the discrepancies of the non-invariant parameters of the countries/economies. The DIF issues in ILSAs are widely discussed in van de Vijver et al. (2019[33]), where flexible modelling, such as Bayesian DIF analysis (Soares, 2009[45]; Chen, 2022[46]) is introduced.

Misconfiguration of the residual structures $\boldsymbol{\Psi}$ can be another cause of the scale bias, as indicated in Equation 10. The violation of local independence introduces bias of the item parameters, in particular for the slope parameters. Therefore, examining local independence is important in ILSAs, especially in the assessments that repeatedly estimate the item parameters. When conducting a scale linking, $\widehat{\boldsymbol{\Psi}}$ corresponding to the anchor items should be checked carefully for the above reason. From a mathematical perspective, tests can be linked if one of the item parameter sets is fixed on the target scale; however, in

practice, it is essential that the anchor items cover a wide range of the item contents and the item formats so that $\hat{\boldsymbol{\Psi}}$ of various types of items can be examined. This is crucial when conducting a vertical linking. The residual invariance assumption can be checked across countries/economies or the testing cycles; however, it rarely holds in practice.

Consequently, IRT modelling with MMLE-EM provides us with a robust and unbiased scale against the growing number of countries with different proficiency levels as long as the measurement equation is appropriately set in each of the testing cycles. Namely, the item parameter constraints and the residual structure of the IRT model should fit the student response data in the item scaling phase to obtain valid system-level statistics. Since the LRM is modelled for each country separately under the assumption of Equation 25, Equation 27, and Equation 28, the LRM does not bias the point estimates of $\boldsymbol{\mu}$.

6.2. Analysing impact of nuisance factors

Nuisance factors that can confound the proficiency estimates of students are, for example, student engagement, and level of testing language proficiency. Moreover, nuisance factors such as an item-position effect and a testlet effect can confound the item parameter estimates; also, DIF by sub-populations (i.e., country/economy, language, gender, test delivery mode, etc.) is a nuisance factor in the scaling phase. Some of these nuisance factors can be cancelled out by the test design, but some cannot. Parametrising the nuisance factors in either $\boldsymbol{\lambda\Phi\lambda}'$ or $\boldsymbol{\Psi}$ is a reasonable option for minimising the impact on the student proficiency estimates. Another option is to involve the nuisance factors in the LRM modelling. Here, statistical modelling that can be applied within the methodological framework of PISA is discussed.

In PISA 2012, booklet effects were involved in the measurement model[9], and they were estimated together with the item parameters in order to prevent confounding of the parameters (OECD, 2014[47]); thus, the nuisance factors can be modelled within the framework of SEM. While the booklet effects were modelled in the structural variables in the same way as the items, Bradlow et al. (1999[48]) considered the model[10] in which additional student proficiencies $\boldsymbol{\theta}_t$ influence the items in testlet t. Namely, $p_j(\theta)$ is calculated based on both $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_t$ if item j belongs to testlet t, whereas $p_j(\theta)$ is calculated only on $\boldsymbol{\theta}_m$ if not. In the former measurement model, the nuisance factor is modelled as observed variables that affect the student responses constantly, whereas the latter model constitutes factors that affect student responses depending on $\boldsymbol{\theta}_t$. In addition to the above models, a model with structured parameters such as hierarchical linear model (Raudenbush, 2002[49]) can also be considered. Note that a better model-data fit can be expected by incorporating the factors into the model.

Another option is to structure the residual covariance in the model. The testlet effect can be incorporated in $\boldsymbol{\Psi}$, where the corresponding elements $\mathrm{Cov}[\boldsymbol{e}_t, \boldsymbol{e}_t]$ are assumed to be correlated in the model. The model requires fewer parameters than the model mentioned above; therefore, more stable estimates can be expected. However, unlike the model that incorporates the nuisance factors into the structural variables, the conditional expectations for each student cannot be estimated in the model since the factors are structured in $\boldsymbol{\Psi}$. Similarly, Fox, Wenzel, and Klotzke (2020[50]) showed a way to estimate testlet effects

[9] In this model, the factor loadings of the cognitive items and the booklets were imposed to be 1.0.

[10] In Bradlow et al. (1999[48]), A Bayesian solution is proposed. Some parameter constrains are required when estimating parameters in the framework of SEM.

under a Bayesian covariance structure model (BCSM) by extracting them from the residuals.

Some types of nuisance factors that have linear relation with a proficiency can be modelled in the LRM by including the variables into the regressors of the model. This enables us to mitigate the influences of the nuisance factors by fixing the estimates to 0 or a certain value, although the impacts are diminished due to the nature of the population modelling. However, the approach is valid only if the relation between the factor and the student proficiency is linear. Unlike incorporating the factors into the measurement model where the estimated parameters of the nuisance factors contribute to the likelihood function, this model influences the prior means of the students directly. The model is simple to implement; however, the impacts are not constant for all students and are different one country/economy to another.

In addition to the above options, the Bayesian solution can be a tool in the entire scaling process. Bayesian approaches allow flexible modelling, and a number of models have been proposed (Fox, 2010[51]) with powerful estimation methods such as Markov Chain Monte Carlo (Albert, 1992[37]; Patz, 1999[52]). For the local dependent items, Bayesian Testlet Model (BMT) is proposed in Wainer et al. (2007[53]) and Ip (2010[54]), where the residual correlations are parametrised. Note that ad-hoc analyses are not recommended since they make evaluations of the models difficult.

6.3. Population modelling and item assignment

In population modelling, the precision of the student proficiency estimates $\mathrm{V}\left[\widehat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i\right]$ is calculated based on $H_L\left(\widetilde{\boldsymbol{\theta}}\right)$ and H_h, in which $H_L\left(\widetilde{\boldsymbol{\theta}}\right)$ is dependent on $\boldsymbol{\theta}$, while $\boldsymbol{\Sigma}$ is not. Therefore, the influence of the prior distribution on the posterior distribution is larger in the proficiency range of low Fisher's information than high Fisher's information range. The effect of incorporating MSAT in ILSAs that employs the population modelling is therefore limited in terms of improving reliability, especially when $\boldsymbol{\Sigma}$ is small.

Regarding the dimensionality of LRM, the regression parameters can be estimated for each dimension separately without deteriorating the model-data fit; hence, it is possible to improve the precision of the student's posterior distribution by incorporating many domains into the population modelling. Since the elements of $\mathrm{offdiag}(\boldsymbol{\Sigma})$ contribute to the variance components of the posterior distribution as shown in Equation 34, the more domains (M) are incorporated in the population modelling, the smaller $\mathrm{V}\left[\widehat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i\right]$ are obtained. However, overfitting issues of the model should be carefully checked when M is large.

Evaluating $\widehat{\boldsymbol{\Psi}}$ and $\mathrm{Cov}\left[\widehat{\boldsymbol{\theta}}, \hat{\boldsymbol{e}}\right]$ of each sub-population is important, especially in an assessment whose item assignment is not balanced in terms of item properties such as response formats and sub-domains that may have interactions with the sub-populations. For example, in PISA, country-DIF is considered in the item scaling phase, while the other DIF, such as gender-DIF, is not. Thus, the unbalanced item exposure to students in a country may cause bias if a sub-population has DIF to a particular type of the item properties. Practically, the distributions of $\widetilde{\boldsymbol{\theta}}_m - \boldsymbol{\gamma}_m \boldsymbol{y}_m$ of each sub-population should be checked in order to keep comparability of the gaps across the testing cycles. As shown in Equation 27, $\mathrm{E}[\boldsymbol{d}] = \mathbf{0}$ is assumed in the LRM; however, if it does not hold for sub-populations, the score gaps will be biased because of the prior distributions.

The ideal booklet design in ILSAs is not only optimising the test information for each student but also balancing the item properties that may cause DIF for each sub-population. Suppose constructed-response items have gender-DIF and the proportions of constructed-

response items by item difficulty levels are imbalanced. If MSAT is employed in this situation, the gender gap will be dependent on the country parameters $\boldsymbol{\varphi}_g$ unless the proportions of the constructed-response items for each student are controlled appropriately in MSAT or different parameter sets are provided to each of the sub-populations. Consequently, enhancing the item bank is of importance in ILSAs, especially for those which employ a dynamic booklet design. It enables us to mitigate unexpected measurement errors caused by the item properties that are not taken into account in the item scaling phase.

References

Albert, J. (1992), "Bayesian estimation of normal ogive item response functions using Gibbs sampling", *Journal of Educational Statistics*, Vol. 17, pp. 251–269. [37]

Asparouhov, T. (2014), "Multiple-group factor analysis alignment", *Structural Equation Modeling*, Vol. 21, pp. 495–508. [22]

Baird, J. (2016), "On the supranational spell of PISA in policy", *Educational Research*, Vol. 58, pp. 121–138. [6]

Birnbaum, A. (1968), *Some latent trait models and their use in inferring an examinee's ability*, Addison-Wesley. [36]

Bock, R. (1981), "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm", *Psychometrika*, Vol. 46, pp. 443–459. [35]

Bock, R. (1972), "Estimating item parameters and latent ability when responses are scored in two or more nominal categories", *Psychometrika*, Vol. 37, pp. 29–51. [26]

Bradlow, E. (1999), "A Bayesian random effects model for testlets", *Psychometrika*, Vol. 64, pp. 153-168. [48]

Chen, C. (2007), "Effects of ignoring item interaction on item parameter estimation and detection of interacting items", *Applied Psychological Measurement*, Vol. 31, pp. 388-411. [17]

Chen, F. (2007), "Sensitivity of goodness of fit indexes to lack of measurement invariance", *Structural Equation Modeling*, Vol. 14, pp. 464-504. [29]

Chen, S. (2022), "Advantages of spike and slab priors for detecting differential item functioning relative to other Bayesian regularizing priors and frequentist lasso", *Structural Equation Modeling*, Vol. 29, pp. 122-139. [46]

Dempster, A. (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society, Series B.*, Vol. 39, pp. 1-38. [34]

Ferrara, S. (1999), "Contextual explanations of local dependencies in item clusters in a large scale hands-on science performance assessment", *Journal of Educational Measurement*, Vol. 36, pp. 119-140. [19]

Foster, N. (2022), "Assessing Creative Skills", *Creative Education*, Vol. 13, pp. 1-29. [8]

Fox, J. (2020), "The Bayesian Covariance Structure Model for Testlets", *Journal of Educational and Behavioral Statistics*, Vol. 46, pp. 219-243. [50]

Fox, J. (2010), *Bayesian Item Response Modeling: Theory and Applications*, Springer-Verlag. [51]

Hambleton, R. (1985), *Item response theory: principles and applications*, Kluwer-Nijhoff. [38]

Harwell, M. (1991), "The use of prior distribution in marginalized Bayesian item parameter estimation: A didactic", *Applied Psychological Measurement*, Vol. 15, pp. 375-389. [39]

Hopfenbeck, T. (2018), "Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment", *Scandinavian Journal of Educational Research*, Vol. 62, pp. 333-353. [4]

Hoskens, M. (1997), "A parametric model for local dependence among test items", *Psychological Methods*, Vol. 2, pp. 261-277. [18]

Ip, E. (2010), "Interpretation of the three parameter testlet response model and information function", *Applied Psychological Measurement*, Vol. 33, pp. 467-482. [54]

Lazarsfeld, P. (1968), *Latent Structure Analysis*, Houghton Mifflin. [40]

Lindblad, S. (2017), *International comparisons of school results: a systematic review of research on large scale assessments in education*, The Swedish Research Council. [5]

Lord, F. (1980), *Applications of item response theory to practical testing problems*, Erlbaum. [16]

Lord, F. (1968), *Statistical Theories of Mental Test Scores*, Addison-Wesley. [7]

Masters, G. (1982), "A Rasch model for partial credit scoring", *Psychometrika*, Vol. 47, pp. 149–174. [24]

Meredith, W. (1993), "Measurement invariance, factor analysis and factorial invariance", *Psychometrika*, Vol. 58, pp. 525–543. [31]

Mislevy, R. (1991), "Randomization-based inference about latent variables from complex samples", *Psychometrika*, Vol. 56, pp. 177–196. [43]

Mislevy, R. (1987), "Exploiting Auxiliary Information About Examinees in the Estimation of Item Parameters", *Applied Psychological Measurement*, Vol. 11, pp. 81–91. [42]

Mislevy, R. (1984), "Estimating latent distributions", *Psychometrika*, Vol. 49, pp. 359-381. [11]

Muraki, E. (1995), "Full-information factor analysis for polytomous item responses", *Applied Psychological Measurement*, Vol. 19, pp. 73-90. [13]

Muraki, E. (1992), "A generalized partial credit model: Application of an EM algorithm", *Applied Psychological Measurement*, Vol. 16, pp. 159-176. [23]

OECD (2019), *PISA 2018 Assessment and Analytical Framework*, OECD Publishing. [1]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing. [10]

OECD (2014), *PISA 2012 Technical Report*, OECD publishing. [47]

OECD (2000), *Measuring Student Knowledge and Skills: The PISA 2000 Assessment of Reading, Mathematical and Scientific Literacy, PISA*, OECD publishing. [2]

OECD (n.d.), *OECD PISA Database*, https://www.oecd.org/pisa/data/ (accessed on 1 October 2022). [3]

OECD (n.d.), *PISA 2018 Technical Report*, OECD publishing, https://www.oecd.org/pisa/data/pisa2018technicalreport/ (accessed on 1 October 2022). [41]

Patz, R. (1999), "A straightforward approach to Markov Chain Monte Carlo methods for item response models", *Journal of Educational and Behavioral Statistics*, Vol. 24, pp. 146–178. [52]

Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests*, Nielsen and Lydiche. [25]

Raudenbush, S. (2002), *Hierarchical Linear Models Applications and Data Analysis Methods*, Sage publishing. [49]

Reise, S. (1993), "Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance", *Psychological Bulletin*, Vol. 114, pp. 552-566. [32]

Rubin, D. (1987), *Multiple imputation for nonresponse in surveys*, John Wiley and Sons. [44]

Samejima, F. (1969), "Estimation of latent ability using a response pattern of graded scores", *Psychometric Monograph*, Vol. 17, pp. 1-100. [28]

Sijtsma, K. (1996), "A survey of theory and methods of invariant item ordering", *British Journal of Mathematical and Statistical Psychology*, Vol. 49, pp. 79-105. [21]

Soares, T. (2009), "An integrated Bayesian model for DIF analysis", *Journal of Educational and Behavioral Statistics*, Vol. 34, pp. 348-377. [45]

Stout, W. (1990), "A new item response theory modeling apploach with applications to unidimentionality assessment and ability estimation", *Psychometrika*, Vol. 55, pp. 293-325. [15]

Takane, Y. (1987), "On the relationship between item response theory and factor analysis of discretized variables", *Psychometrika*, Vol. 52, pp. 393–408. [14]

Thissen, D. (1986), "A taxonomy of item response models", *Psychometrika*, Vol. 51, pp. 567–577. [27]

van de Vijver, F. (2019), "Invariance analyses in large-scale studies", *OECD Education Working Papers*, Vol. 201, pp. 1-110. [33]

Wainer, H. (2007), *Testlet response theory and its applications*, Cambridge university Press. [53]

Warm, T. (1989), "Weighted likelihood estimation of ability in item response theory", *Psychometrika*, Vol. 54, pp. 427-450. [12]

Widaman, K. (1997), *Exploring the measurement invariance of psychological instruments: Applications in the substance use domain*, American Psychological Association. [30]

Yamamoto, K. (2019), "Introduction of multistage adaptive testing design in PISA 2018", *OECD Education Working Papers*, Vol. 209, pp. 1-29. [9]

Yen, W. (1993), "Scaling performance assessments: Strategies for managing local item dependence", *Journal of Educational Measurement*, Vol. 30, pp. 187-213. [20]