Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies

Robitzsch, Alexander; Lüdtke, Oliver

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Robitzsch, A., & Lüdtke, O. (2022). Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies. *Measurement Instruments for the Social Sciences*, *4*, 1-20. https://doi.org/10.1186/s42409-022-00039-w

Robitzsch and Lüdtke
Measurement Instruments for the Social Sciences (2022) 4:9
https://doi.org/10.1186/s42409-022-00039-w

Measurement Instruments for the Social Sciences

ADVANCES IN METHODOLOGY

Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies

Alexander Robitzsch[1,2*] and Oliver Lüdtke[1,2]

Abstract

International large-scale assessments (LSAs), such as the Programme for International Student Assessment (PISA), provide essential information about the distribution of student proficiencies across a wide range of countries. The repeated assessments of the distributions of these cognitive domains offer policymakers important information for evaluating educational reforms and received considerable attention from the media. Furthermore, the analytical strategies employed in LSAs often define methodological standards for applied researchers in the field. Hence, it is vital to critically reflect on the conceptual foundations of analytical choices in LSA studies. This article discusses the methodological challenges in selecting and specifying the scaling model used to obtain proficiency estimates from the individual student responses in LSA studies. We distinguish design-based inference from model-based inference. It is argued that for the official reporting of LSA results, design-based inference should be preferred because it allows for a clear definition of the target of inference (e.g., country mean achievement) and is less sensitive to specific modeling assumptions. More specifically, we discuss five analytical choices in the specification of the scaling model: (1) specification of the functional form of item response functions, (2) the treatment of local dependencies and multidimensionality, (3) the consideration of test-taking behavior for estimating student ability, and the role of country differential items functioning (DIF) for (4) cross-country comparisons and (5) trend estimation. This article's primary goal is to stimulate discussion about recently implemented changes and suggested refinements of the scaling models in LSA studies.

Keywords: Large-scale assessment, Item response models, Scaling, Linking, Differential item functioning, Partial invariance, Item response function, Trend estimation, PISA, Survey statistics, Educational assessment

Introduction

In the last two decades, international large-scale assessments (LSAs) have provided important information about the distribution of student proficiencies across a wide range of countries and age groups. For example, every 3 years since 2000, the Programme for International Student Assessment (PISA) reported international comparisons of student performance in three content areas (reading, mathematics, and science; OECD, 2014). The repeated assessments of these content domains provide policymakers with important information for the evaluation of educational reforms and also received considerable attention from the media. Furthermore, LSAs provide unique research opportunities (Singer & Braun, 2018) that are increasingly used by researchers from different fields to investigate the relations between student proficiency and other cognitive and noncognitive variables. From the beginning, LSAs have been confronted with many methodological challenges (Rutkowski et al., 2013). In

*Correspondence: robitzsch@leibniz-ipn.de

[1] IPN — Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany
Full list of author information is available at the end of the article

addition, it seems that the analytical strategies employed in LSAs often define methodological standards for applied researchers in the field. Hence, it is vital to critically reflect on the conceptual foundations of analytical choices in LSA studies.

In the present article, we reflect on methodological challenges in selecting and specifying the scaling model used to obtain proficiency estimates from the individual student responses in LSA studies. Our discussion distinguishes between design-based inference (based on sampling designs for specific populations of persons and test items) and model-based inference (based on specific assumptions of statistical models). It is argued that for the official reporting of LSA results, design-based inference should be preferred because it allows for a clear definition of the target of inference (e.g., country mean achievement) and is less sensitive to specific modeling assumptions. More specifically, we discuss five specific analytical choices for the scaling model that received considerable attention in the methodological literature and that they can affect the reporting of LSA results: (1) specification of the functional form of item response functions, (2) the treatment of local dependencies and multidimensionality, (3) the consideration of test-taking behavior for estimating student ability, and the role of country differential items functioning (DIF) for (4) cross-country comparisons and (5) trend estimation. The main goal of this article is to stimulate discussion about the role of recent changes that have been implemented in the scaling models of LSA studies (with a particular emphasis on PISA) or that were suggested by methodologists as further refinements of the currently used scaling models.

Model-assisted design-based inference

Model-assisted design-based inference for persons

In the remainder of the article, we consider statistics (e.g., mean, standard deviation, quantiles) of the distribution of an ability variable (e.g., reading ability). Let θ_n denote a corresponding ability of person n. In the usual sampling design of LSA studies, not all students in a population (e.g., a country) are sampled. Frequently, stratified multistage sampling is employed in which schools are sampled in the first stage, and students within a school are sampled in the second stage (Meinck, 2020). Consequently, not all students within a country have the same probability of being sampled, and it is important to take into account the different selection probabilities when inferring from the sample to the population. Hence, student weights $w_{\mathcal{P},n}$ are used where $w_{\mathcal{P},n}$ is the inverse of the probability that person n is sampled (Meinck, 2020; Rust et al., 2017). The subscript $\mathcal{P}$ indicates that the weights refer to the population $\mathcal{P}$ of persons (e.g., students). The inference for a statistic of the ability distribution (e.g., mean achievement) from the sample to the population of students in a country is also referred to as a design-based inference (Lohr, 2010; Särndal et al., 2003).

We illustrate the typical approach for statistical inference in LSA studies for the estimation of two distribution parameters of an ability distribution (e.g., reading ability for a country in the PISA study): the mean μ and the variance σ^2. Suppose that there are N sampled students within a country and unobserved (and error-free) latent abilities θ_n for all $n=1, \ldots, N$. Then, in a design-based (db) approach, sample estimates for the mean μ and the variance σ^2 are given by:

$$\hat{\mu}_{\text{db}} = \frac{\sum_{n=1}^{N} w_{\mathcal{P},n}\theta_n}{\sum_{n=1}^{N} w_{\mathcal{P},n}} \text{ and } \hat{\sigma}^2_{\text{db}} = \frac{\sum_{n=1}^{N} w_{\mathcal{P},n}\left(\theta_n - \hat{\mu}_{\text{db}}\right)^2}{\sum_{n=1}^{N} w_{\mathcal{P},n}}, \quad (1)$$

where ability values θ_n are weighted by student weights $w_{\mathcal{P},n}$. However, there are two obstacles to applying the estimation formulas in Eq. (1) and adopting a pure design-based approach in LSA studies. First, abilities cannot be directly measured in LSA studies but have to be inferred from a multivariate vector $\mathbf{x}_n$ of discrete item responses of student n. In the following, we only consider dichotomous items for the sake of notational simplicity. A scoring rule f that maps item responses $\mathbf{x}_n$ to estimated abilities $\hat{\theta}_n$ (i.e., $\hat{\theta}_n = f(\mathbf{x}_n)$) is required. Typically, the ability is considered as a latent random variable θ, but estimated abilities $\hat{\theta}_n$ for student n are prone to measurement errors. The extent of measurement errors relies on a specified measurement model (i.e., an item response theory (IRT) model; Yen & Fitzpatrick, 2006). The probability for item responses $\mathbf{X}=(X_1,\ldots,X_I)$ conditional on a latent ability θ is modeled by posing a local independence assumption:

$$P(\mathbf{X}=\mathbf{x}|\theta) = \prod\nolimits_{i=1}^{I} P\left(X_i = x_i|\theta;\boldsymbol{\gamma}_i\right), \quad (2)$$

where I is the number of items, X_i is the item response on item i, and $\boldsymbol{\gamma}_i$ denotes a vector of item parameters for item i. Note that error-prone ability estimates result in biased estimates of parameters for the distribution of θ, particularly for the standard deviation and quantiles, and biased correlation of abilities with covariates (Lechner et al., 2021; Wu, 2005).

The second obstacle in LSA studies like PISA is that not all students receive items in all ability domains (OECD, 2014; see also Frey et al., 2009). Hence, imputation procedures must be used to borrow for each student information from administered ability domains to obtain estimates for non-administered ability domains (Little & Rubin, 2002). The issue of non-administered ability domains is addressed using a so-called latent background model (LBM; Mislevy, 1991). The motivation for using an

LBM from which plausible values are drawn is twofold. First, there is a measurement error in estimated abilities because only a finite number of items are administered to each student. Plausible values are realizations of the ability variable that allow secondary data analysts to provide answers to substantive research questions that are not affected by measurement errors in estimated abilities. Second, plausible values can also be drawn for an ability domain for a student who did not receive items in this domain by taking into account the relationships across all ability domains and student covariates.

For a $C \times 1$ vector of observed covariates $\mathbf{z}_n$ (e.g., variables such as gender or sociodemographic status), the LBM for a target unidimensional ability θ (e.g., reading) and a vector of additional $D-1$ abilities $\boldsymbol{\eta}$ (e.g., mathematics and science) is defined as:

$$\begin{pmatrix} \theta \\ \boldsymbol{\eta} \end{pmatrix} = \mathbf{B}\mathbf{z}_n + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \mathbf{T}), \text{Cov}(\mathbf{z}_n, \boldsymbol{\varepsilon}) = 0, \tag{3}$$

where MVN denotes the multivariate normal distribution, $\mathbf{B}$ is a $D \times C$ matrix of regression coefficients, and $\mathbf{T}$ is a $D \times D$ matrix of residual covariances of the vector of random variables $(\theta, \boldsymbol{\eta})$. Note that the specification of the LBM in (3) also needs the specification of a measurement model such as the one in (2). More formally, for an extended vector of item responses $\mathbf{y}_n$ that are indicators of the vector of latent variables $(\theta, \boldsymbol{\eta})$, the probability distribution in the latent background model is defined as:

$$P(\mathbf{Y} = \mathbf{y}_n | \mathbf{z}_n) = \int P(\mathbf{Y} = \mathbf{y}_n | \theta, \boldsymbol{\eta}; \boldsymbol{\gamma}) P(\theta, \boldsymbol{\eta} | \mathbf{z}_n; \mathbf{B}, \mathbf{T}) \, \mathrm{d}\theta \mathrm{d}\eta, \tag{4}$$

where the measurement part $P(\mathbf{Y} = \mathbf{y}_n | \theta, \boldsymbol{\eta}; \boldsymbol{\gamma})$ is defined by the IRT model in Eq. (2), and the structural model $P(\theta, \boldsymbol{\eta} | \mathbf{z}_n; \mathbf{B}, \mathbf{T})$ is defined by the LBM in Eq. (3). Also, note that (3) can be rewritten as a conditional unidimensional normal distribution:

$$\theta = \mathbf{B}^*(\mathbf{z}_n, \boldsymbol{\eta}) + \varepsilon^* \text{and } \varepsilon^* \sim N\left(0, \tau^2\right) \tag{5}$$

using an appropriate $1 \times (C + D - 1)$ matrix of regression coefficients $\mathbf{B}^*$. It can be seen in Eq. (5) that in the LBM, the ability θ is inferred from student covariates $\mathbf{z}_n$ and other ability domains $\boldsymbol{\eta}$. Note that τ^2 is the residual variance for the ability θ, and the variances in $(\theta, \boldsymbol{\eta})$ are allowed to differ across all ability dimensions. Suppose items are administered in the target ability domain θ. In that case, the IRT model in Eq. (2) typically provides the major amount of information for the target ability. In contrast, for non-administered ability domains, only the LBM delivers information for the ability θ. That is, administered ability domains $\boldsymbol{\eta}$ and covariates $\mathbf{z}_n$ are used for imputing the target ability. In the operational practice of LSA studies, the imputations are called plausible values (Mislevy, 1991; von Davier & Sinharay, 2014). Plausible values $\left(\widetilde{\theta}_n, \widetilde{\boldsymbol{\eta}}_n\right)$ for student n are drawn from subject-specific posterior distributions $P(\theta, \boldsymbol{\eta} | \mathbf{y}_n, \mathbf{z}_n)$ (also referred to as predictive distributions for $(\theta, \boldsymbol{\eta})$; von Davier & Sinharay, 2014) that can be derived from Eq. (4):

$$\text{simulate} \begin{pmatrix} \widetilde{\theta}_n \\ \widetilde{\boldsymbol{\eta}}_n \end{pmatrix} \text{from } P(\theta, \boldsymbol{\eta} | \mathbf{y}_n, \mathbf{z}_n) = \frac{P(\mathbf{Y} = \mathbf{y}_n | \theta, \boldsymbol{\eta}; \boldsymbol{\gamma}) P(\theta, \boldsymbol{\eta} | \mathbf{z}_n; \mathbf{B}, \mathbf{T})}{\int P(\mathbf{Y} = \mathbf{y}_n | \theta, \boldsymbol{\eta}; \boldsymbol{\gamma}) P(\theta, \boldsymbol{\eta} | \mathbf{z}_n; \mathbf{B}, \mathbf{T}) \, \mathrm{d}\theta \mathrm{d}\boldsymbol{\eta}} \tag{6}$$

In the case of a unidimensional ability θ and normally distributed measurement errors $\text{SE}\left(\hat{\theta}_n\right)$ of the point estimate $\hat{\theta}_n$, plausible values $\widetilde{\theta}_n$ can be written as:

$$\widetilde{\theta}_n = \rho_c \hat{\theta}_n + (1 - \rho_c) \mathbf{B}^*(\mathbf{z}_n, \boldsymbol{\eta}) + e_n, e_n \sim N\left(0, \kappa^2\right), \tag{7}$$

where the conditional reliability ρ_c and the posterior variance κ^2 are determined by:

$$\rho_c = \frac{\tau^2}{\tau^2 + \text{E}\left[\text{SE}\left(\hat{\theta}_n\right)\right]^2} \text{ and } \kappa^2 = \left(1 - \rho_c\right)\left(1 - R^2\right), \tag{8}$$

and where $R^2 = \frac{\tau^2}{\text{Var}(\theta)}$ is the proportion of explained variance in Eq. (5) (see Mislevy, 1991), and $E\left[\text{SE}\left(\hat{\theta}_n\right)\right]^2$ is the average of squares of individual standard errors of measurement.

If the IRT model in Eq. (2) is misspecified, the likelihood part $P(\mathbf{y}_n | \theta, \boldsymbol{\eta}; \boldsymbol{\gamma})$ in Eq. (6) will be misspecified. Consequently, the model-implied reliability will be incorrect and plausible values do not correctly reflect the uncertainty associated with the ability variable θ. In practice, item parameters $\boldsymbol{\gamma}$ are fixed in Eq. (6) when drawing plausible values and the likelihood part can be written as a function of θ and $\boldsymbol{\eta}$, that is, there is a multidimensional function $h_n(\theta, \boldsymbol{\eta}) = P(\mathbf{y}_n | \theta, \boldsymbol{\eta}; \boldsymbol{\gamma})$. The amount of error associated with $(\theta, \boldsymbol{\eta})$ is quantified by the peakedness of the function h_n. The measurement error assumption can be modified by adjusting the function h_n to be steeper (i.e., increase reliability) or more flat (i.e., decrease reliability; see Chandler & Bate, 2007; Mislevy, 1990). In more detail, the unidimensional person-specific likelihood function is approximated with an unnormalized normal density function; that is:

$$h_n(\theta) = P(\mathbf{x}_n|\theta; \boldsymbol{\gamma}) = c_{n,\theta}\phi\left(\theta; \mu_{n,\theta}, \sigma_{n,\theta}\right) \quad (9)$$

where ϕ is the normal density, and $c_{n,\theta}$ is a scaling factor. We set $\mu_{n,\theta} = \hat{\theta}_n$ and $\sigma_{n,\theta} = \mathrm{SE}\left(\hat{\theta}_n\right)$. Methods that resample items (see Design-based or model-based inference for items? section) can be used to estimate reliability (or the standard error $\mathrm{SE}\left(\hat{\theta}_n\right)$) in misspecified IRT models (Wainer & Wright, 1980). Hence, the person-specific standard deviation $\sigma_{n,\theta}$ in Eq. (9) can be modified by posing different assumptions about the reliability of the ability scores.

The statistical inference in LSA studies almost exclusively relies on plausible values (von Davier & Sinharay, 2014). It is evident that the effects of misspecifications in the LBM vanish with an increasing number of items because individual squared standard errors $\left[\mathrm{SE}\left(\hat{\theta}_n\right)\right]^2$ converge to zero (and ρ_c in Eq. (8) will be close to 1; Marsman et al., 2016). The current approach in LSA studies that relies on plausible values can be described as a model-assisted design-based inference (Binder & Roberts, 2003; Brewer, 2013; Little, 2004; Särndal et al., 2003; Ståhl et al., 2016). With the model-assisted approach, as it has been called, one tries to construct estimators with good design-based properties (Gregoire, 1998). However, the finite population is never considered generated according to model parameters (Särndal et al., 2003). In contrast, the model is only a statistical device to allow a design-based inference with desirable statistical properties. The model-assisted design-based approach in LSA studies is design-based because the inference to a concrete population of students in a country is warranted, but—at the same time—it is model-assisted because a model (IRT model and the LBM) is utilized for computing plausible values that substitute the non-observable ability θ_n. In practice, for reducing the simulation error and enabling the estimation of standard errors with imputed data, several plausible values (e.g., $M = 10$) are generated; that is, for each student n, there are M plausible values $\overset{\sim(m)}{\theta}_n$ $(m = 1, \ldots, M)$. The sample estimates based on all M plausible values for the mean μ and the variance σ^2 of the ability variable θ are given as (see Mislevy, 1991):

$$\hat{\mu}_{\mathrm{db,PV}} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} w_{\mathcal{P},n}\overset{\sim(m)}{\theta}_n}{M\sum_{n=1}^{N} w_{\mathcal{P},n}} \text{ and } \hat{\sigma}^2_{\mathrm{db,PV}} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} w_{\mathcal{P},n}\left(\overset{\sim(m)}{\theta}_n - \hat{\mu}_{\mathrm{db,PV}(m)}\right)^2}{M\sum_{n=1}^{N} w_{\mathcal{P},n}}, \quad (10)$$

where the mean of the mth plausible value is given as:

$$\hat{\mu}_{\mathrm{db,PV}(m)} = \frac{\sum_{n=1}^{N} w_{\mathcal{P},n}\overset{\sim(m)}{\theta}_n}{\sum_{n=1}^{N} w_{\mathcal{P},n}}. \quad (11)$$

Note that the subject-specific posterior distribution $\mathrm{P}(\theta, \boldsymbol{\eta}|\mathbf{y}_n, \mathbf{z}_n)$ that is used to generate plausible variables in (6) is a continuous function of θ and $\boldsymbol{\eta}$. Hence, the statistics in Eq.(10) relying on plausible values are shortcuts for evaluating person-specific integrals. In more detail, for an infinite number of plausible values, the estimates in (10) can be written as:

$$\hat{\mu}_{\mathrm{db,PV}} = \frac{\sum_{n=1}^{N} w_{\mathcal{P},n}\int \theta \mathrm{P}\left(\theta, \boldsymbol{\eta}|\mathbf{y}_n, \mathbf{z}_n\right) \mathrm{d}\theta \mathrm{d}\boldsymbol{\eta}}{\sum_{n=1}^{N} w_{\mathcal{P},n}} \quad (12)$$

$$\hat{\sigma}^2_{\mathrm{db,PV}} = \frac{\sum_{n=1}^{N} w_{\mathcal{P},n}\int \left(\theta - \hat{\mu}_{\mathrm{db,PV}}\right)^2 \mathrm{P}\left(\theta, \boldsymbol{\eta}|\mathbf{y}_n, \mathbf{z}_n\right) \mathrm{d}\theta \mathrm{d}\boldsymbol{\eta}}{\sum_{n=1}^{N} w_{\mathcal{P},n}} \quad (13)$$

Comparing these estimates with the design-based estimates $\hat{\mu}_{\mathrm{db}}$ and $\hat{\sigma}^2_{\mathrm{db}}$ (see Eq. (1)) highlights that $\hat{\mu}_{\mathrm{db,PV}}$ and $\hat{\sigma}^2_{\mathrm{db,PV}}$ depend on both the design (i.e., relying on weights $w_{\mathcal{P},n}$) and model assumptions (i.e., relying on individual posterior distributions $\mathrm{P}(\theta, \boldsymbol{\eta}|\mathbf{y}_n, \mathbf{z}_n)$). Hence, the choice of a particular IRT model (see Eq. (2)) and the specification of the LBM (see Eq. (3)) have the potential to change the meaning of θ, and, hence, can affect the meaning of μ and σ^2 and their corresponding estimates.

Equations (12) and (13) also clarify that statistical inference in LSA studies can be described as model-assisted design-based inference. The design-based inference is represented by including student weights $w_{\mathcal{P},n}$, but it is model-assisted because the ability variable θ is represented by the posterior distribution $\mathrm{P}(\theta, \boldsymbol{\eta}|\mathbf{y}_n, \mathbf{z}_n)$ that relies on the chosen IRT model and the LBM. In a further alternative hybrid approach of design-based inference and model-based inference (see Ståhl et al., 2016), subjects can additionally be weighted by including weights $\nu_{\mathcal{P},n}$ according to their fit to a statistical model. For example, model-based student-specific weights $\nu_{\mathcal{P},n}$ can be derived according to their fit to the scaling model (person fit; see Conijn et al., 2011; Hong & Cheng, 2019; Raiche et al., 2012; Schuster & Yuan, 2011). In such an approach, students whose item responses are atypical with respect to the IRT model (e.g., non-scalable students; see Haertel, 1989) would be downweighted compared to students whose item responses are consistent with the IRT model. Doing so might increase the information function when using student-specific weights. However, a critical issue might be that reweighting based on $\nu_{\mathcal{P},n}$ can change the

representativity of a sample regarding a target population of students. Corresponding sample estimates in such a hybrid design-model-based (dmb) are given by:

$$\hat{\mu}_{\text{dmb,PV}} = \frac{\sum_{n=1}^{N} w_{\mathcal{P},n} \nu_{\mathcal{P},n} \int \theta \mathrm{P}\left(\theta, \boldsymbol{\eta} | \mathbf{y}_n, \mathbf{z}_n\right) \mathrm{d}\theta \mathrm{d}\boldsymbol{\eta}}{\sum_{n=1}^{N} w_{\mathcal{P},n} \nu_{\mathcal{P},n}} \tag{14}$$

$$\hat{\sigma}^2_{\text{dmb,PV}} = \frac{\sum_{n=1}^{N} w_{\mathcal{P},n} \nu_{\mathcal{P},n} \int \left(\theta - \hat{\mu}_{\text{dmb,PV}}\right)^2 \mathrm{P}\left(\theta, \boldsymbol{\eta} | \mathbf{y}_n, \mathbf{z}_n\right) \mathrm{d}\theta \mathrm{d}\boldsymbol{\eta}}{\sum_{n=1}^{N} w_{\mathcal{P},n} \nu_{\mathcal{P},n}} \tag{15}$$

It might be tempting to identify subgroups of students that do not fit the IRT model as a threat to validity and, subsequently, to eliminate these students from the final analysis by effectively setting $\nu_{\mathcal{P},n}$ to zero. Clearly, the estimates $\hat{\mu}_{\text{db,PV}}$ and $\hat{\mu}_{\text{dmb,PV}}$ will turn out to be different in practice and likely target different estimands. There is a danger that estimates in Eqs. (14) and (15) generalize to a different population of students compared to the model-assisted design-based estimates in Eqs. (12) and (13). In a hybrid design model-based inference, the specification of a model allows the target estimand to differ from the estimand in a design-based approach because, in the former, observations are weighted by $w_{\mathcal{P},n}\nu_{\mathcal{P},n}$, while in the latter observations are weighted by $w_{\mathcal{P},n}$. This hybrid approach should also be clearly distinguished from model-assisted design-based inference in which the model is only considered as a tool that is used to implement a design-based inference approach.

Standard errors can be computed by resampling methods (e.g., jackknife or balanced repeated replication methods; Kolenikov, 2010; Rust et al., 2017) in which subgroups of students are resampled. The multi-stage clustered sampling with explicit and implicit stratification can easily be accommodated in these resampling methods (Meinck, 2020).

We argue that a fully design-based inference should be the first analysis option in LSA studies. Obviously, this could only be realized if an infinite (or very large) number of items would be administered in the ability domain of interest so that the variance of the measurement error is negligible. However, the number of administered items in most applications is not large enough such that measurement errors in abilities can be neglected. Hence, the statistical inference employed in LSA studies (i.e., model-assisted design-based inference) depends on measurement error assumptions in the IRT model and the specified LBM. However, we would argue that misspecifications in the IRT model can be accepted (see Functional form of item response functions section) because the choice of the IRT model should be driven by the meaning of the ability variable (e.g., equal weighting of items in the scoring rule) and not by the model fit. In contrast, the degree of misspecification in the LBM should be minimized, even though it can be challenging to adequately treat the high dimensionality of the predictor variables (Grund et al., 2021; von Davier, Khorramdel, et al., 2019). Overall, we believe that the hybrid design-model-based inference poses threats to validity because the fit of each subject in a model can redefine the contribution of subjects by additionally incorporating weights $\nu_{\mathcal{P},n}$ in the analysis. Thus, a statistical model (and, hence, psychometrics) is allowed to change the target of inference. We prefer a design-based approach that is less sensitive to specific modeling assumptions when reporting LSA results.

Design-based or model-based inference for items?

In the previous subsection, we discussed the kind of statistical inference for the population of persons. It is not apparent which kind of statistical inference is needed to represent the process of choosing test items in LSA studies. The test items should cover the ability domain defined by the test framework (test blueprint; see also Pellegrino & Chudowsky, 2003; Reckase, 2017). It might be legitimate to assume that there exists a larger population of test items (henceforth, labeled by $\mathcal{I}$) from which the items are chosen in a particular study, and true ability values would be defined as outcomes in a study in which all items from the population would have been chosen (Cronbach & Shavelson, 2004; see also Ellis, 2021, Kane, 1982; Brennan, 2001). Interestingly, it has been argued that classical test theory (CTT) or generalizability theory (GT; Cronbach et al., 1963) treats items in a study as random and, as a consequence, allows the inference to a larger set of items in a population of items (see also Nunnally & Bernstein, 1994; Markus & Borsboom, 2013). In contrast, IRT treats items as fixed (Brennan, 2010) and restricts the statistical inference to the items chosen in a test. This distinction is strongly related to the question of whether the representation of item responses in the ability θ_n follows a design-based (i.e., CTT or GT) or a model-based inference (i.e., IRT). In CTT or GT, items are treated as exchangeable by posing assumptions about the sampling process. Notably, if the selection (or sampling) of items from the domain of test items is appropriately conducted, the inference for the ability from the chosen items to the population of items would be valid. From a design-based perspective, substantive theory (e.g., by test domain experts, item developers) should define the contribution of each chosen item. In more detail, there are a priori defined item-specific weights $w_{\mathcal{I},i}$ that enter the scoring rule for the ability estimate $\hat{\theta}_n$:

$$\hat{\theta}_n = f\left(\sum_{i=1}^{I} w_{\mathcal{I},i} x_{ni}\right) \tag{16}$$

If the administered test mimics the population of items, all item weights will be set to be equal to each other; that is $w_{\mathcal{I},i} = 1$ for all $i = 1,...,I$, and $\hat{\theta}_n$ is given by monotone transformation of the sum score. If the item selection in a study is adequately made, a subsequent post hoc elimination of items based on a fit in the IRT model (e.g., item fit statistics for the IRT model in Eq. (2)) potentially changes the target of inference (Brennan, 1998; see also Uher, 2021). By choosing an IRT model, there are model-based derived item weights $v_{\mathcal{I},i}(\theta)$ (so-called locally optimal item weights) that define a local scoring rule for the ability (see Eq. (45) in Appendix 1)

$$\hat{\theta}_n = f\left(\sum_{i=1}^{I} v_{\mathcal{I},i}(\theta_n) x_{ni}; \theta_n\right), \tag{17}$$

where the item weights $v_{\mathcal{I},i}(\theta_n)$ are given by (Birnbaum, 1968; Chiu & Camilli, 2013; Yen & Fitzpatrick, 2006):

$$v_{\mathcal{I},i}(\theta) = \frac{P_i'(\theta)}{P_i(\theta)[1 - P_i(\theta)]}. \tag{18}$$

The main consequence of the local scoring rule in Eq. (17) is that the choice of the IRT model implicitly defines the contribution of items in the ability, and the model-based approach (see Eq. (17)) can deviate from the design-based approach (see Eq. (16)) in which weights $w_{\mathcal{I},i}$ are defined by sampling considerations (Camilli, 2018). By posing a particular IRT model, locally optimal item weights $v_{\mathcal{I},i}(\theta)$ are determined that provide the best-fitting model in terms of the potentially misspecified maximum likelihood function (White, 1982). Items that are most informative for θ in the IRT model receive the largest weights, which, in turn, can influence the interpretations of the ability score. The item weights $v_{\mathcal{I},i}(\theta)$ are locally defined for every ability value θ. To summarize the effects of item scoring at the country level, Camilli (2018) defined effective country-specific item weights $v_{\mathcal{I},ic}$ that integrate locally optimal item weights for the country-specific ability density f_c:

$$v_{\mathcal{I},ic} = \int v_{\mathcal{I},i}(\theta) f_c(\theta) \mathrm{d}\theta. \tag{19}$$

The quantity $v_{\mathcal{I},ic}$ allows the evaluation of whether the effective contribution of an item in the ability score θ varies across countries.

If an IRT model were used for scoring, the measurement error in estimated abilities $\hat{\theta}_n$ is mainly driven by the observed information function (Magis, 2015). Hence, the statistical model defines the extent of error associated with ability scores. In contrast, in a design-based approach of CTT or GT, sampling assumptions regarding selecting items from the population of items define the extent of measurement errors. In such a design-based perspective, no assessment of the model fit for the set of item responses $\mathbf{x}_n$ is required. For example, the use of Cronbach's alpha (Cronbach, 1951) as a reliability measure for the sum score does not require that a model with equal item loadings and uncorrelated residual errors have to fit the data of item responses (Cronbach, 1951; Cronbach & Shavelson, 2004; Ellis, 2021; Meyer, 2010; Nunnally & Bernstein, 1994; Tryon, 1957). In the same manner, as for persons, resampling methods for items can be used to determine standard errors in estimated abilities (Liou & Yu, 1991; Wainer & Thissen, 1987; Wainer & Wright, 1980) by resampling items or groups of items for which abilities are reestimated (see also Michaelides & Haertel, 2014). It is also possible to include additional dependence by item stratification (e.g., multiple test components; Cronbach et al., 1965; Meyer, 2010) or item clustering (e.g., due to the arrangement of items in testlets, that is, several items share a common item stimulus such as a common reading text; Bradlow et al., 1999)[1] in resampling methods for items.

We tend to favor the scoring rules from a design-based perspective in Eq. (16) over the model-based perspective in Eq. (17) because, in our view, substantive theory should define the contribution of items in the ability score for carefully constructed test items.

We also want to emphasize that item fit statistics are related to the local fit of single items in an IRT model that treats items as fixed. Notably, the assessment of item fit statistics does not follow the perspective that treats items as random, and removing items (due to poor model fit) from the computation of the ability has the potential to change the target of statistical inference. We elaborate on these issues in detail in Specific analytical choices in scaling models section.

A plea for a symmetric role of persons and items

In the last two subsections, we discussed statistical inference for the populations of persons and items in LSA studies. For both populations of persons and items, (model-assisted) design-based, model-based, or hybrid variants of statistical inference can be employed. In most LSA studies, statistical inference for the population of persons is primarily handled under a design-based perspective. At the same time, the model-based inference is also present for the population of items. We argue that persons and items should have

[1] As mentioned by an anonymous reviewer, local dependence of m dichotomous items within a testlet can be avoided by forming a single item with $m+1$ categories that is defined as the sum of the single items. This polytomous item can be used the IRT modeling without violating the local independence assumption.

symmetric roles in LSA studies based on previous arguments. We believe that design-based inference should rule out model-based inference for both facets. There seems to be a consensus among researchers that students who do not fit a particular IRT model should not be removed from the analysis in LSA studies. By doing so, the sample of students would no longer be representative of the population of students. We argue that the same perspective should be taken for items: one should not simply remove items from the scoring rule for ability or country comparisons because they do not follow a particular IRT model. In contrast, items should be considered random, and IRT models should be regarded as statistical devices to achieve the inferential goals of LSA studies. In this sense, these psychometric models merely define estimating equations, and the fit of the chosen model is not of central relevance. The employed likelihood functions in estimating abilities in LSA studies are likely to be misspecified. We argue that their sole role is the (implicit) definition of target estimands of interest (Boos & Stefanski, 2013). Statistical inference should preferably rely on resampling methods for persons and items because these do not rely on a correctly specified statistical model. Also, note that local fit statistics can be computed for each person and item. However, atypical persons or items (with respect to a model) do not invalidate statistical inference from a design-based perspective.

Specific analytical choices in scaling models

In the following, we discuss five topics that are of central relevance in the specification of the scaling modeling in LSA studies: (1) the choice of the functional of the item response function, (2) the role of local dependence and multidimensionality, (3) the treatment of additional information from the test-taking behavior (e.g., response times), (4) the role of country DIF in cross-country comparisons, and (5) trend estimation. In this discussion, we highlight the consequences of a design-based perspective for the specification of the scaling model.

Functional form of item response functions

As argued in Model-assisted design-based inference section, the choice of the IRT model can affect the meaning of the latent ability variable θ_n. Of particular importance is the specification of the item response function (IRF) that describes the relationship between item responses and ability. In the following, we discuss the most common IRFs and use locally optimal weights (see Design-based or model-based inference for items? section) to show how the choice of different IRFs affects item contributions in the scoring rule for the latent ability variable (see Eq. (21)).

Probably the most popular IRT model is the one-parameter logistic (1PL) IRT model (also known as the Rasch model; Rasch, 1960), which employs the IRF:

$$\mathrm{P}(X_{ni} = 1|\theta_n) = \Psi(\theta_n - b_i), \tag{20}$$

where $\Psi(x) = \exp(x)/[1 + \exp(x)]$ denotes the logistic distribution function and b_i is the item difficulty. For the 1PL model, the sum score is a sufficient statistic; that is, the scoring rule in Eq. (17) is given by:

$$\theta_n = f\left(\sum_{i=1}^{I} 1 \cdot x_{ni}\right) \tag{21}$$

Hence, all items are equally weighted in the ability variable θ_n and receive the local item score $v_{\mathcal{I},i}(\theta) = 1$. Note that this weight is independent of θ. If the set of selected items in the test adequately represents the population of items (i.e., $w_{\mathcal{I},i} = 1$), it can be argued that the 1PL should be the preferred measurement model because the uniform weighting in the sum score in Eq. (21) can be considered as a proxy of an equally weighted sum score for the population of items (see also Stenner et al., 2008, 2009). The 1PL model was used in PISA as a scaling model until PISA 2012 (OECD, 2014).

In the two-parameter logistic (2PL; Birnbaum, 1968) model, items are allowed to have different item discriminations a_i:

$$\mathrm{P}(X_{ni} = 1|\theta_n) = \Psi(a_i(\theta_n - b_i)) \tag{22}$$

The sufficient statistic is given by the weighted sum score in which locally optimal item weights $v_{\mathcal{I},i}(\theta)$ are given by a_i that are independent of θ:

$$\theta_n = f\left(\sum_{i=1}^{I} a_i x_{ni}\right) \tag{23}$$

In most applications, item discriminations a_i are estimated from data and are determined to maximize model fit in terms of the log-likelihood function. However, the empirically determined weights can differ from a priorily specified item weights $w_{\mathcal{I},i}$ in Eq. (16) in a design-based inference. In this case, model-based and design-based inference will not provide the same results. However, if a design-based inference and the scoring rule in Eq. (16) are desired, the 2PL model can be utilized as a measurement model with fixed item discriminations; that is, $a_i = w_{\mathcal{I},i}$ (see, e.g., Haberkorn et al., 2016). The 2PL model is used in PISA as a scaling model since PISA 2015 (OECD, 2017; see also Jerrim et al., 2018).

In the three-parameter logistic (3PL; Birnbaum, 1968) model, an additional guessing parameter g_i is included in the IRF:

$$\mathrm{P}(X_{ni}=1|\theta_n)=g_i+\left(1-g_i\right)\Psi(a_i(\theta_n-b_i)) \tag{24}$$

For the 3PL model, locally optimal item weights indeed depend on the ability θ (Chiu & Camilli, 2013):

$$v_{\mathcal{I},i}(\theta)=\frac{a_i}{1+g_i\exp\left[-a_i(\theta-b_i)\right]} \tag{25}$$

Note that the contribution of item i in the ability value θ increases as a function of θ. In this sense, the model-implied effective item scores for countries depend on the country-specific ability distributions (Camilli, 2018). Another objection against the 3PL model is that g_i is not the probability of guessing for multiple-choice items (Aitkin & Aitkin, 2006; von Davier, 2009). Alternative IRT models have been proposed that circumvent this issue (Aitkin & Aitkin, 2006). Occasionally, arguments against using the 3PL model are made for reasons of a lack (or weak empirical) identification of model parameters of the 3PL model (Maris & Bechger, 2009; San Martín et al., 2015). However, these concerns vanish with sufficiently large samples, distributional assumptions for the ability variable, or weakly informative prior distributions for item parameters. The 3PL model is in operational use in PIRLS (Foy & Yin, 2017) and TIMSS (Foy et al., 2020).

In the psychometric literature, there is recent interest in the four-parameter logistic (4PL; e.g., Culpepper, 2017; Loken & Rulison, 2010) that also allows a slipping parameter s_i in the IRF:

$$\mathrm{P}(X_{ni}=1|\theta_n)=g_i+\left(1-s_i-g_i\right)\Psi(a_i(\theta_n-b_i)). \tag{26}$$

Students can receive a very large ability θ_n, even though their item response probabilities can be substantially smaller than one due to the presence of slipping parameters. As a consequence, a failure on some items is not so strongly penalized in the 4PL model because a wrong item response can be attributed to a slipping behavior. Like in the 3PL model, the locally optimal item weights in the 4PL model also depend on the ability (see Magis, 2013). It is unlikely that these θ-dependent item weights in a model-based perspective will coincide with apriori specified item weights in a design-based perspective. To our knowledge, the 4PL model is not currently in operational practice in any international LSA study.

Alternatively, asymmetric IRFs (Bolt et al., 2014; Goldstein, 1980) can be used that allow item weights to depend on item difficulty (see also Dimitrov, 2016). The most flexible approach would be achieved by a semiparametric or a nonparametric specification of IRFs (Falk & Cai, 2016; Feuerstahler, 2019; Ramsay & Winsberg, 1991). These IRFs imply model-based item weights that might strongly differ from weights that are specified under a design-based perspective and may therefore distort the test composition that is defined in test blueprints (Camilli, 2018).

Brown et al. (2007) showed that using the 3PL model instead of the 1PL or the 2PL model might have non-negligible consequences for low-performing students. Hence, country comparisons involving low-performing countries in LSAs or low-performing subgroups of students can be affected by a particular choice of a scaling model. Overall, country standard deviations and percentiles (Brown et al., 2007; Robitzsch, 2022c) are much more affected by choosing a particular IRT model than country means (Jerrim et al., 2018).

To summarize, choosing a particular IRF implies different item weights and scoring rules for the ability variable θ. It can be questioned whether IRFs should be chosen for the sole purpose of increasing reliability (and model fit) because different IRFs correspond to different estimation targets. In our view, the choice of an IRF should be mainly a question of validity and cannot be answered by model fit or item fit statistics. However, if the superior model fit is defined as the primary goal of model choice in LSA studies, more complex IRFs (3PL, 4PL, semiparametric IRFs) will almost always outperform simpler IRFs (1PL, 2PL) (see Robitzsch, 2022c). The switch from the 1PL to the 2PL model in recent PISA studies can, therefore, in our opinion, not be defended for reasons of better model fit because the 4PL model or alternative flexible IRT models outperform the 1PL, 2PL, and 3PL model in PISA in terms of model fit (Culpepper, 2017; Liao & Bolt, 2021; Robitzsch, 2022c). However, the crucial question is whether the derived ability from the 4PL model is constituted valid. Following Brennan (1998), we believe that a psychometric model should not prescribe the contribution of items in the ability score. If items in the test represent items in a (hypothetical) larger item domain, the 1PL model can be defended even though it will likely not fit empirical data. From a design-based inference, the employed likelihood function in the ability estimation is intentionally misspecified (see Model-assisted design-based inference for persons section). Overall, we tend to favor the operational use of the 1PL model in LSA studies because the ability score primarily reflects an equal contribution of items that appear in the test. Nevertheless, the likelihood part associated with the misspecified IRT model has to be modified adequately to reflect the reliability (see Model-assisted design-based inference for persons section). Two further misspecifications—in addition to the functional form of the item response function—will be discussed in the next section.

Local dependence and multidimensionality

In the previous section, we discussed the choice of the IRF in unidimensional IRFs. Typical abilities assessed in LSA studies will be multidimensional, which, in turn, causes a violation of the local independence assumption. In Model-assisted design-based inference for persons section, we argued that a misspecified unidimensional IRT model could be defended from a design-based inference point of view. However, the model-implied reliability obtained from fitting a unidimensional model can be incorrect. If an ability domain is multidimensional (e.g., subdimensions in reading ability), and the multidimensionality is considered construct-relevant (Shealy & Stout, 1993), the model-implied reliability from a fitted unidimensional model will underestimate the true reliability (Zinbarg et al., 2005). Moreover, items are frequently arranged in testlets such that different items share the same item stimulus. This deviation from local independence can introduce additional error (Monseur et al., 2011) if testlet effects are considered construct-irrelevant[2] (Sireci et al., 1991). In this case, the reliability of ability scores decreases (Zinbarg et al., 2005). There will be construct-relevant multidimensionality and construct-irrelevant testlet effects in empirical LSA data. However, the unidimensional IRT model is used as a scoring model because a unidimensional summary is required for country comparisons. Resampling methods for items (see Design-based or model-based inference for items? section) can be used for determining standard errors associated with estimated abilities $\hat{\theta}_n$. The estimated standard errors can be used to adjust the likelihood part of the measurement model to generate plausible values (see Model-assisted design-based inference for persons section; Bock et al., 2002; Mislevy, 1990).

Current operational practice in LSA studies ignores deviations from local independence in the scaling of item responses. While we do not think this introduces a large bias in country means or individual ability estimates, the estimated uncertainty associated with plausible values might be incorrect. However, the reliability estimate obtained from the misspecified IRT model could be defended on the rationale that the residual covariance of items is assumed to be zero in IRT modeling. In practice, positive and negative residual covariances cancel out on (a weighted) average. Continuing this argument, the recent practice of ignoring local dependence could be defended if the underestimation of reliability due to the multidimensionality is compensated by an overestimation due to testlet effects. However, there is no chance to test this property with empirical data. One always has to make assumptions about the true average residual correlation in latent variable models (Westfall et al., 2012). This view on latent variable models corresponds to a design-based perspective in which one defines the interchangeability of items by design assumptions that cannot be guaranteed by any statistical test.

For example, for fitting items in an LSA mathematics test, a multidimensional IRT (MIRT) model with exploratively defined dimensions will likely fit the data. However, a unidimensional summary mathematics score is vital to reporting, and the dimensions obtained from the MIRT model cannot be easily interpreted. The MIRT model can be interpreted as a domain sampling model (McDonald, 1978, 2003). In our view, reporting a summary scale score from a misspecified unidimensional IRT model if a MIRT model holds is justified because statistical models are not used to fit the data but to fulfill particular purposes defined by researchers and practitioners. A bifactor model with a general factor and specific dimensions might also be attractive for applied researchers (Reise, 2012).

We argue that the approximate unidimensionality assumption for ability in the scaling model can be defended in practice due to the following two empirical findings. First, there is frequently only a low amount of multidimensionality found in data (e.g., for subdimensions in reading or mathematics in PISA data; OECD, 2017). Second, when the number of items tends to infinity (with a bounded number of items in testlets), the local dependence of testlet effects asymptotically vanishes in the estimation of model parameters (Ellis & Junker, 1997; Stout, 1990). For example, with about 100 administered items per domain and at most 5 to 7 items per testlet, biases in scaling models due to local dependence might be small to moderate.

The role of test-taking behavior in the scaling model

It has frequently been argued that measured student performance in LSA studies is affected by test-taking strategies (Rios, 2021; Wise, 2020). For example, in a recent paper that was published in the highly-ranked *Science* journal, Pohl et al. (2021) argued that "current reporting practices, however, confound differences in test-taking behavior (such as working speed and item nonresponse) with differences in competencies (ability). Furthermore, they do so in a different way for different examinees, threatening the fairness of comparisons, such as country rankings." (Pohl et al., 2021, p. 338). Hence, the reported student performance (or, equivalently, student ability)

[2] For testlets administered for the ability domain reading, one might argue that testlets effects are (partly) construct-relevant because the common item stimulus (i.e., the whole reading text) have to be processed for answering the items.

would be confounded by a "true" ability and test-taking strategies. Importantly, the authors question the validity of country comparisons that are currently reported in LSA studies and argue for an approach that separates test-taking behavior (i.e., item response propensity and working speed) from a purified ability measure. In the following, we clarify that the additional consideration of test-taking behavior has the potential to change the meaning of the measured abilities substantially (within and also between countries). As the proposed approach focuses on the modeling of omitted responses (Pohl et al., 2021), we start with a brief summary of how missing responses are treated in LSA studies.

Missing item responses can be classified into omitted items within the test and not-reached items at the end of the test (see Rose et al., 2017). Until PISA 2012, not reached items were treated as incorrect in ability estimation, while they were not scored as incorrect since PISA 2015 (OECD, 2017). The proportion of not reached items is used as a covariate in the LBM since PISA 2015 while recoding all not-reached item responses as not administered in the scaling. We would argue that this treatment of not-reached items can decrease the validity of ability scores because countries can easily manipulate the scores on not-reached items by advising test takers to work slowly through the test and only produce missing item responses if there are many items they do not know. Thus, we would not concur with Pohl et al. (2021, p. 339), who conclude that "[…] scoring not-reached items as incorrect—as done in some LSAs—results in scores that differ in their meaning, depending on whether examinees do or do not show missing values. This jeopardizes the comparability of performance scores across examinees and, thus, fairness.". Unfortunately, the role of not-reached items becomes even more critical in scaling with the implementation of multi-stage testing because the proportion rates of not reached in some modules in recent PISA studies are considerable.

Pohl et al. (2021) propose the speed-accuracy and omission (SA+O) model (Ulitzsch et al., 2020b) that simultaneously models item responses, response indicators that indicate whether students omit items, and response times. Not-reached items are treated as non-administered. In the SA+O model, these observed variables are associated with four latent variables: an ability, a response propensity variable, and two speed variables (for observed and omitted items). In the following, we discuss the potential implications of using this model in LSA studies. Let X_{ni} be the item response of person n on item i, and R_{ni} be the response indicator that takes a value of 1 if the item is observed and a value of 0 if it is missing (i.e., omitted). Moreover, let T_{ni} be the logarithmized response time for person n on item i. In the SA+O model, the joint distribution of indicator variables (X_{ni}, R_{ni}, T_{ni}) is modeled as (Ulitzsch et al., 2020b)

$$\begin{aligned}&\mathrm{P}\big(X_{ni}=x_{ni},R_{ni}=r_{ni},T_{ni}=t_{ni}|\theta_n,\xi_n,\eta_{0n},\eta_{1n}\big)\\&\quad=\mathrm{P}\big(X_{ni}=x_{ni}|\theta_n\big)^{r_{ni}}\,\mathrm{P}\big(R_{ni}=r_{ni}|\xi_n\big)f_1\big(t_{ni}|\eta_{1n}\big)^{r_{ni}}f_0\big(t_{ni}|\eta_{0n}\big)^{1-r_{ni}}\end{aligned}\tag{27}$$

where ξ_n denotes the response propensity, η_{1n} is the speed variable associated with observed items, η_{0n} is the omission speed, and f_1 and f_0 are normal densities for response times of observed and omitted items, respectively.[3] To further illustrate the meaning of Eq. (27), we first consider the decomposition for observed item responses (i.e., $R_{ni}=1$):

$$\begin{aligned}&\mathrm{P}\big(X_{ni}=x_{ni},R_{ni}=1,T_{ni}=t_{ni}|\theta_n,\xi_n,\eta_{0n},\eta_{1n}\big)\\&\quad=\mathrm{P}\big(X_{ni}=x_{ni}|\theta_n\big)\,\mathrm{P}\big(R_{ni}=1|\xi_n\big)f_1\big(t_{ni}|\eta_{1n}\big)\end{aligned}\tag{28}$$

For missing item responses (i.e., $R_{ni}=0$ and $X_{ni}=\mathrm{NA}$), Eq. (27) simplifies to

$$\begin{aligned}&\mathrm{P}\big(X_{ni}=\mathrm{NA},R_{ni}=0,T_{ni}=t_{ni}|\theta_n,\xi_n,\eta_{0n,}\eta_{1n}\big)\\&\quad=\mathrm{P}\big(R_{ni}=0|\xi_n\big)f_0\big(t_{ni}|\eta_{0n}\big)\end{aligned}\tag{29}$$

In a model-based estimation approach of the SA+O model, effectively, missing item responses X_{ni} have to be imputed based on the latent variables $(\theta_n, \xi_n, \eta_{1n}, \eta_{0n})$ and response time T_{ni}. We now derive how item responses, response indicators, and response times are used for estimating the ability in the SA+O model. For the derivation, it is convenient to reparametrize the vector of latent variables $(\theta_n, \xi_n, \eta_{1n}, \eta_{0n})$ to $(\theta_n, \xi_n^*, \eta_{1n}^*, \eta_{0n}^*)$, where ξ_n^*, η_{1n}^* and η_{0n}^* are residualized latent variable in which the ability θ_n is partialled out:

$$\xi_n=\alpha_{\xi\theta}\theta_n+\xi_n^*\text{ and }\eta_{hn}=\alpha_{\eta_h\theta}\theta_n+\eta_{hn}^*\text{ for }h=0,1\tag{30}$$

Then, the IRT model in Eq. (27) can be equivalently written as:

$$\begin{aligned}&\mathrm{P}\big(X_{ni}=x_{ni},R_{ni}=r_{ni},T_{ni}=t_{ni}|\theta_n,\xi_n^*,\eta_{0n}^*,\eta_{1n}^*\big)\\&\quad=\mathrm{P}\big(X_{ni}=x_{ni}|\theta_n\big)^{r_{ni}}\,\mathrm{P}\big(R_{ni}=r_{ni}|\theta_n,\xi_n^*\big)f_1\big(t_{ni}|\theta_n,\xi_n^*,\eta_{1n}^*\big)^{r_{ni}}f_0\big(t_{ni}|\theta_n,\xi_n^*,\eta_{0n}^*\big)^{1-r_{ni}}\end{aligned}\tag{31}$$

For item responses X_{ni}, a 2PL model is assumed (see Eq. (22)). The probability of responding to an item is assumed to be

[3] It might be possible to included separate distributions for response times of correct and incorrect item responses (see Bolsinova et al., 2017).

$$\mathrm{P}(R_{ni} = r_{ni}|\xi_{\mathrm{n}}) = \Psi(\gamma_i(\xi_n - \beta_i)) \tag{32}$$

Logarithmized response times T_{ni} are modeled as conditional normal distributions:

$$T_{ni} \mid \eta_{1n}, \eta_{0n}, R_{ni} = h \sim \mathrm{N}(\eta_{hn} - \tau_{ih}, \lambda_{ih}^{-2} \text{ for } h = 0, 1.) \tag{33}$$

In Appendix 2, local item scores are derived that define local sufficient statistics of indicator variables for θ_n. For observed item responses, the item weight is given by the item discrimination from the 2PL model (i.e., a_i). The local item score for θ_n for a missing item response (see Eq. (53) in Appendix 2) is given by:

$$a_i P_i(\theta) - \alpha_{\xi\theta}\gamma_i - \left(\alpha_{\eta_1\theta}\lambda_{i1} - \alpha_{\eta_0\theta}\lambda_{i0}\right)t_{ni} \tag{34}$$

From Eq. (34), it can be seen that two (in general) positive terms will be subtracted from the item score $a_i P_i(\theta)$. Note that $a_i P_i(\theta)$ would be considered as an appropriate item score if the missingness mechanism is ignorable (i.e., treatment of omitted items as non-administered provides a valid strategy; Pohl & Carstensen, 2013; Pohl et al., 2014). In case of a positive correlation of θ and ξ, the imputed score is adjusted by the value $\alpha_{\xi\theta}\gamma_i$. Furthermore, because omitted items are typically associated with shorter response times, the adjustment term $\left(\alpha_{\eta_1\theta}\lambda_{i1} - \alpha_{\eta_0\theta}\lambda_{i0}\right)t_{ni}$ also plays a role in the scoring rule. Hence, the ability variable θ_n also enters the log-likelihood contributions of response indicators and response times. Consequently, response indicators and response times contribute to the imputation of omitted item responses and influence the model-based estimation of abilities defined in Eq. (27).

Like in our discussion for not-reached items, we would argue that the scoring rule implied by the SA+O model has substantial consequences for the interpretation of ability scores in LSA studies. Study results can be simply manipulated at the country level if students are advised to skip items they do not know or to produce very short response times in such cases. In our opinion, the possibility of influencing students' test-taking behavior severely threatens the validity and fairness of country comparisons. Furthermore, in our research with LSA data, we found that the conditional independence assumptions of item responses and response indicators in the SA+O model are strongly violated, resulting in a worse model fit of the SA+O model (see Robitzsch, 2021b). There is empirical evidence that students who do not know the answer to an item have a high probability of omitting this item even after controlling for latent variables. This seems to be particularly the case for constructed response items. Thus, we believe that the dependence of responding to an item from the true but unknown item response must be considered even after conditioning on latent variables (Robitzsch, 2021b). Given these concerns about the less plausible assumptions of the SA+O model and its consequences for the validity of country comparisons, we would argue that it cannot be recommended for operational use in large-scale assessment studies.

It is important to emphasize that the adjustments—and hence the scoring rules for ability—in the SA+O model will differ from country to country because the relationships between ability, response propensity, and speed differ across countries (Sachse et al., 2019; Pohl et al., 2021). In our view, a country comparison that does not employ the same scoring rule for each country cannot be considered valid (or fair).

Much psychometric work seems to imply that simulation studies demonstrate that missing item responses should never be scored as incorrect (Pohl & Carstensen, 2013; Pohl et al., 2014; Rose et al., 2017). We oppose such a perspective because simulation studies are not helpful in decisions about how to handle missing item responses (Rohwer, 2013; Robitzsch, 2021b). One can simulate data that introduce missing item responses only for incorrectly solved items (Rohwer, 2013). In this case, all IRT models that score missing item responses as incorrect provide biased model parameter estimates (Robitzsch, 2021b). Moreover, the scoring of items should always be conducted under validity considerations. We think that omitted (constructed response) items should always be scored as incorrect because alternative scoring rules decrease validity.

We would like to note that our discussion of always treating omitted responses as incorrect is mainly related to the reporting of country comparisons of ability variables. It might be valuable to investigate different missing data treatments to study the validity of the ability construct. In particular, it is interesting whether and how log data or response times are related to the omitted items. Moreover, not scoring omitted items as incorrect might be more valid for studying relationships of ability with covariates (e.g., student motivation). However, we nevertheless insist that one should not choose a scoring method (i.e., not scoring omitted items as incorrect) that can be simply manipulated at the country level to increase the country's scores in an LSA (see Robitzsch, 2021b).

In addition, continuing the arguments of Pohl et al. (2021), other test-taking behaviors could be used for purifying ability. For example, response effort such as rapid guessing (Deribo et al., 2021; Ulitzsch et al., 2020a) or performance decline (Debeer & Janssen, 2013; Jin & Wang, 2014) could be taken into account. Moreover, the ability variable θ could also be redefined in a scaling model in which item responses and response times load on θ, resulting in a purified latent variable for speed (Costa et al., 2021). Furthermore, measurement models could also involve an additional student latent variable α_n that characterizes person fit (Conijn et al., 2011; Ferrando, 2019; Raiche et al., 2012):

$$\mathrm{P}(X_{ni} = 1|\theta_n, \alpha_n) = \Psi(a_i\alpha_n(\theta_n - b_i)) \tag{35}$$

Such a model would weigh persons in the log-likelihood by a model-based weight α_n, and the model would certainly be justified by reasons of model fit against simpler alternatives. As a consequence, the local scoring rule for abilities also depends on the person fit variable α_n, which would further complicate the interpretation. We strongly believe that including latent variables that capture test-taking behavior in measurement models should be avoided in the official reporting of LSA results. In our view, the explicit modeling of test-taking behaviors leads to the opposite of fairer country comparisons. A design-based approach should be preferred for inferences regarding abilities in LSA. Test-taking behavior is always coupled with a realized test design in this approach. Researchers (as well as the public and policy) have to judge whether the assessed abilities—under the given design—are deemed valid.

Country DIF and cross-sectional country comparisons

For most international LSA studies, the comparability of test scores across countries is of crucial importance (Rutkowski & Rutkowski, 2019). We understand comparability as the possibility to conduct valid comparisons of statistical quantities across countries. Conceptual and statistical approaches for assessing comparability are distinguished in the literature. Statistical approaches include the assessment of differential item functioning (DIF; Holland & Wainer, 1993; Penfield & Camilli, 2007), focusing on the heterogeneity in item parameters across countries. In the 2PL model, there is empirical evidence that item difficulties vary from country to country (von Davier, Khorramdel, et al., 2019):

$$\mathrm{P}(X_{nci} = 1|\theta_{nc}) = \Psi(a_i(\theta_{nc} - b_{ic})), \tag{36}$$

where the index c denotes the country, and discrimination parameters a_i are assumed to be constant across countries (uniform DIF) in our treatment. Note that country-specific item difficulties b_{ic} (i.e., uniform DIF) are allowed. The presence of DIF with respect to countries is denoted as (cross-sectional) country DIF (see Monseur et al., 2008; Robitzsch & Lüdtke, 2019). If only a few item difficulties b_{ic} are allowed to deviate from a common item difficulty b_i, it is said that partial invariance holds (von Davier, Khorramdel, et al., 2019). Uniform DIF effects e_{ic} can be defined as:

$$e_{ic} = b_{ic} - b_i. \tag{37}$$

DIF in item difficulties is more apparent in practical applications than DIF in item discriminations. Therefore, we decide only to discuss findings for uniform DIF. In the case of nonuniform DIF, the arguments will not change, but some derivations do not result in closed formulas, as presented in the following.

Since PISA 2015, the assumption of partial invariance (Oliveri & von Davier, 2011; OECD, 2017; von Davier, Yamamoto, et al., 2019) has been incorporated into the scaling model. Non-invariant item parameters are determined utilizing item fit statistics such as the root mean square deviation (RMSD) statistic (Tijmstra et al., 2020). In the partial invariance approach, the majority of item parameters are assumed to be equal (i.e., invariant) across countries (e.g., more than 70% of the item parameters are invariant; see Magis & De Boeck, 2012), and there is a low proportion of country-specific item parameters. In PISA, the proportion of non-country-specific item parameters is defined as the comparability of a scale score (Joo et al., 2021). For example, Joo et al. (2021) noted that less than 10% of the items would be declared as misfitting items in PISA if default cutoffs of the RMSD statistic were used in PISA. In contrast, until PISA 2012, the 1PL scaling model with invariant item parameters was assumed, and country DIF was ignored unless it could be attributed to technical issues in item administration (e.g., translation errors; Adams, 2003).

The assumption of country-specific item parameters effectively eliminates some items from pairwise country comparisons (Robitzsch & Lüdtke, 2020a, 2022). Moreover, the set of effectively used items differs across comparisons (e.g., the comparison between country A and country B could be based on different items than the comparison between country A and country C; see also Zieger et al., 2019). It has been argued that this property poses a threat to validity, and researchers are comparing apples and oranges when pursuing the partial invariance approach (Robitzsch & Lüdtke, 2022). We believe that the decision of whether an item induces bias for country comparisons is not primarily of statistical nature. Camilli (1993) pointed out (see also Penfield & Camilli, 2007) that expert reviews of items showing DIF should accompany DIF detection procedures. Only those items should be excluded from country comparisons for which it is justifiable to argue that construct-irrelevant factors caused DIF (see also El Masri & Andrich, 2020; Zwitser et al., 2017). However, the purely statistical approach since PISA 2015 based on partial invariance disregards that DIF items could be construct-relevant. Also, note that PIRLS and TIMSS do not use country-specific item parameters and rely on a scaling model that assumes full invariance of item parameters across countries (Foy et al., 2020). From a validity perspective, we would prefer the approach that ignores country DIF if the DIF cannot be attributed to test the administration issues. This strategy more closely follows a design-based inference perspective for items because the test design and not a psychometric model

should guarantee whether the set of items in a test is presentative for a specific item population (Brennan, 1998).

In contrast to the partial invariance approach that assumes that most items do not show DIF (and only a few items possess large country DIF), the assumption of full noninvariance can be made, which assumes that all items show country DIF effects (Fox, 2010; Fox & Verhagen, 2010). In our experience and in line with other researchers, we find the partial invariance assumption unlikely to hold in empirical data. Instead, in our experience from empirical studies, DIF effects (see Eq. (37)) are frequently symmetrically distributed and closely follow a normal distribution (Robitzsch et al., 2020; Sachse et al., 2016). Moreover, a preference for partial invariance over the full nonivariance assumption with symmetrically distributed DIF effects is unjustified because there is always arbitrariness in defining identification constraints for DIF effects (Robitzsch, 2022a). Importantly, different identification constraints are employed by choosing different fitting functions (or linking functions) (Robitzsch, 2022a).

To acknowledge the dependence of country comparisons on the chosen set of items due to country DIF, linking errors (LE; Robitzsch, 2020, 2021c; Robitzsch & Lüdtke, 2019; Wu, 2010) have been proposed to quantify the heterogeneity in the country means due to the selection (or sampling) of items. The inclusion of the item facet for describing the uncertainty in group means has been studied in GT for a long time (Brennan, 2001; Kane & Brennan, 1977). Assume that the 2PL model with uniform country DIF effects (see Eq. (36)) holds and that the country-specific item parameters b_{ic} can be decomposed into a common item difficulty b_i and country-specific deviations e_{ic}:

$$b_{ic} = b_i + e_{ic}, \text{where } \mathrm{E}(e_{ic}) = 0 \text{ and } \mathrm{Var}(e_{ic}) = \tau^2_{\mathrm{DIF},c} \tag{38}$$

where $\tau^2_{\mathrm{DIF},c}$ is the country-specific DIF variance. For I items, the uncertainty due to the selection of items in the 2PL model is quantified in the following cross-sectional linking error:

$$\mathrm{LE}_{\mathrm{cs},c} = \sqrt{\frac{\sum_{i=1}^{I} a_i^2}{\left(\sum_{i=1}^{I} a_i\right)^2}}\, \tau_{\mathrm{DIF},c} \tag{39}$$

Moreover, due to $\mathrm{E}(e_{ic}) = 0$, estimated country means are unbiased for a large number of items I. For the 1PL model that assumes equal item discriminations a_i, Eq. (39) simplifies to (see Robitzsch & Lüdtke, 2019):

$$\mathrm{LE}_{\mathrm{cs},c} = \frac{1}{\sqrt{I}} \tau_{\mathrm{DIF},c} \tag{40}$$

Instead of establishing a scaling model assuming partial invariance, we prefer the additional error component associated with items for reporting in LSA studies. The total error $\mathrm{TE}_{\mathrm{cs},c}$ for a cross-sectional country mean contains the standard error $\mathrm{SE}_{\mathrm{cs},c}$ due to the sampling of persons as well the linking error $\mathrm{LE}_{\mathrm{cs},c}$ due to the selection of items (Wu, 2010)[4]:

$$\mathrm{TE}_{\mathrm{cs},c} = \sqrt{\mathrm{SE}^2_{\mathrm{cs},c} + \mathrm{LE}^2_{\mathrm{cs},c}} \tag{41}$$

The uncertainty of item selection also affects other statistical parameters (e.g., standard deviation, quantile, regression coefficient). Linking errors can be more flexibly obtained by resampling items (Brennan, 2001). It should be emphasized that linking errors also occur if invariant item parameters are assumed in the scaling model. There are still consequences of heterogeneity in item selection for the country means, even if no country-specific item parameters are explicitly modeled in the scaling model.

Previous studies have shown that the choice of how to handle DIF items can impact country means (Robitzsch, 2020, 2021c; Robitzsch & Lüdtke, 2020a, 2022). It is also possible that different DIF treatments can also impact country comparisons of relationships of abilities with covariates. Moreover, it could be argued that demonstrating partial invariance of item parameters across countries does not guarantee the invariance of relationships of abilities with covariates across countries. In such a case, invariance analysis must be performed for items by testing for potential interaction effects of countries and the covariate of interest (Davidov et al., 2014; Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Unfortunately, incorrect statements that metric invariance in a multiple-group model ensures the comparability of covariances of abilities and covariates across countries can be frequently found in the literature (e.g., He et al., 2017, 2019). In our view, we think that the assessment of measurement invariance is neither necessary nor sufficient for comparability (see also Robitzsch, 2022a). However, we would like to note that the reasoning is even inconsistent in the literature on measurement invariance (Davidov et al., 2014).

In this section, we argued against a partial invariance approach that removes items from particular country comparisons (see Robitzsch & Lüdtke, 2022). In empirical data, country DIF effects will almost always occur. There are two options for handling the presence of DIF effects in the scaling models. First, DIF effects can be ignored in a concurrent scaling approach in which the incorrect assumption of invariant item parameter is posed. Second, separate scaling can be performed at the country level, and linking methods

[4] The estimated linking error $\mathrm{LE}_{\mathrm{cs},c}$ in Eqs. (39) or (40) also contains a standard error associated with sampling of students. Hence, this variance contribution should be subtracted in Eq. (41). In practice, the positive bias in $\mathrm{TE}_{\mathrm{cs},c}$ might be regarded as negligible.

are used to compare countries (Robitzsch & Lüdtke, 2020a, 2022). Notably, concurrent scaling can only be more efficient than separate scaling for correctly specified IRT models, that is, in the absence of country DIF. In the presence of DIF, DIF effects are weighted by a likelihood discrepancy function in concurrent scaling. The estimates of country means can generally be less precise than separate scaling with subsequent linking. In these approaches, the weighing of DIF effects is determined by choosing a linking function. We think that a linking function should be chosen that does not automatically eliminate items with large DIF effects from comparisons (i.e., in robust linking; see He & Cui, 2020; Robitzsch & Lüdtke, 2022). Moreover, the concurrent scaling approach is probably based on a misspecified IRT model that can result in a biased estimation of the latent ability distribution parameters (i.e., standard deviation, quantiles). Interestingly, concurrent calibration or the anchored item parameter estimation approach that does not allow country-specific item parameters frequently results in less stable country mean or country standard deviation estimates than a linking approach (Robitzsch, 2021a). Finally, we believe that the sample sizes in typical LSA studies are large enough to apply a separate scaling approach with subsequent linking for the 1PL or the 2PL model.

Country DIF and trend estimation

One of the primary outcomes in LSA studies is trend estimation which enables monitoring of educational systems concerning students' abilities. The original trend estimate for two time points is computed by subtracting the cross-sectional country mean of the first time point from the second time point. As an alternative, a marginal trend estimate has been proposed that performs the linking across the two time points only on the link items administered in both studies (Gebhardt & Adams, 2007; Robitzsch & Lüdtke, 2019). Original trends have the advantage that officially reported cross-sectional country means can be utilized for computing the trend estimate (e.g., the difference). However, Robitzsch and Lüdtke (2019, 2021) showed analytically and with simulation studies that original trend estimates can be less precise than marginal trend estimates if there is a sufficiently large number of unique items; that is, items that are only administered at one of the two time points (see also Gebhardt & Adams, 2007). The primary reason for the increased precision of marginal trend estimates is that cross-sectional country DIF turns out to be relatively stable across time points. Consequently, unique items introduce additional variability in the country means due to DIF effects (Robitzsch & Lüdtke, 2019). In PISA, there is a switch from major to minor domains (or the other way around) for two of the three primary domains mathematics, reading, and science. If the number of unique items is large compared to the number of link items, original trend estimates in PISA tend to be much more variable than marginal trend estimates (Carstensen, 2013; Gebhardt & Adams, 2007; Robitzsch & Lüdtke, 2019; Robitzsch et al., 2020). On the other hand, by relying on the link items, country DIF effects are automatically controlled for in marginal trend estimates because the stable country DIF effects occur to the same extent at both time points and therefore cancel when calculating achievement trends.

We would like to note that marginal trend estimates were originally proposed at the country level, based on separate scaling with subsequent linking for each country (Gebhardt & Adams, 2007). In our experience based on simulation studies, it can be demonstrated that this requirement is not the essential reason that marginal trends can be more efficient than original trends (Robitzsch & Lüdtke, 2021). The linking could be conducted at the (international) level of all countries (i.e., in a joint scaling approach that involves all countries and assumes invariant item parameters across countries). However, the crucial point is that it should only involve the link items and not the unique items (see the analytical and simulation findings of Robitzsch & Lüdtke, 2019, 2021).

The variability in trend estimates due to the selection of items is quantified by linking errors (Gebhardt & Adams, 2007; OECD, 2017) in LSA studies. The linking error employed in PISA until PISA 2012 quantifies the uncertainty in trend estimates for the 1PL model based on the variance of item parameter drift (IPD; i.e., a difference of item difficulties across time; OECD, 2017). Notably, this error only assesses variability due to link items, ignoring the variability due to country DIF. Robitzsch and Lüdtke (2019) proposed a linking error for the 1PL model that also reflects the variability of original trend estimates due to item selection. Since PISA 2015, the computational approach for the linking error changed by utilizing a recalibration method (OECD, 2017; see also Martin et al., 2012). The motivation for the change in computation was that it should also apply to the recently implemented analytic changes (i.e., the 2PL model and the partial invariance approach). In the newly proposed method, data from the first time point are recalibrated using item parameters from the second time point. The linking error is defined as the variance of the average squared difference of original and recalibrated country means (OECD, 2017). If all item parameters were assumed invariant, the same item parameters as in the original calibration for the link items would be used. Hence, it can be shown that the newly proposed linking error will be small if most of the item parameters are assumed to be invariant. We would like to emphasize that there has been an essential conceptual change in how linking errors are defined in PISA since PISA 2015. In our opinion, the recalibration method might be helpful in defining an effect size of the extent of noninvariance in

item parameters in terms of variability in the recalibrated country means. Hence, in case of perfect comparability in the definition of PISA (Joo et al., 2021), no country-specific item parameters were used, and the newly proposed linking error would be zero. However, we do not believe that the new approach (implemented since PISA 2015) correctly reflects the variability in original trend estimates due to item selection because variability cannot vanish by assuming invariant item parameters. Consequently, the new linking error approach must differ from the previous approach. It has been shown analytically and in simulation studies that the newly proposed linking error substantially differs from the previously employed linking error in PISA (Robitzsch & Lüdtke, 2020b). While we admit that the computation of linking errors has to be modified in recent PISA cycles due to the use of the 2PL model, we would question that the recently proposed linking error provides a solid basis for statistical inference for trend estimates.

Finally, it can be discussed how several cycles of an LSA study should be optimally analyzed when trend estimates are of primary interest. While previous PISA cycles link subsequent PISA studies to each other in a chain linking (OECD, 2017), other researchers opted for a multiple group IRT concurrent scaling approach that assumes that most item parameters are invariant across time points and countries (von Davier, Yamamoto, et al., 2019). It has been argued that the concurrent scaling approach provides more stable trend estimates (von Davier, Yamamoto, et al., 2019; p. 485) by relying on the assumption of partial invariance (i.e., only a few item parameters are not invariant) because more stable item parameter estimates would be obtained. However, it should be noted that the claimed superiority of concurrent scaling was not confirmed by simulation studies. Moreover, the validity of such a statement would require that the partial invariance assumption holds. As for cross-sectional LSA data, we suppose there is no empirical evidence for this assumption. Hence, we would argue that there is lacking support for the higher efficiency of the concurrent scaling approach compared to separate scalings for each time point with subsequent linking (see also Robitzsch & Lüdtke, 2021).

Discussion

In this article, we reflected on several analytical choices in LSA studies. We illustrated that it could be crucial to distinguish between a design-based or model-based perspective on statistical inference. When it comes to official reporting in LSA studies, we argued that a design-based perspective should predominate a model-based perspective. In a part of the methodological LSA literature, there is a tendency to prefer more complex psychometric models with the promise that these complex models produce more stable and less biased estimates of student abilities. However, these claims are primarily made from a model-based perspective, and we clarified that model-based approaches often redefine the meaning of the abilities of interest. For example, using the 2PL model instead of the 1PL model implies a different weighing of items primarily defined through optimizing model fit. This contradicts a design-based perspective in which the contribution of items in a score is a priorily defined by a test framework. The reliance on a partial invariance model for scaling is another example of how a model-based perspective can change the meaning of country comparisons. In some sense, it can be argued that the partial invariance approach compares apples and oranges because the set of effectively used items differs across country comparisons.

From a design-based perspective, the likelihood function that involves the IRT model and the LBM is typically misspecified in LSA studies. It can always be acknowledged that models are only approximately true. However, we even do not believe that the concept of approximate fit makes sense when favoring the 1PL model over the 2PL model. The 1PL model is preferable because the main goal is to use an equally weighted sum score as a sufficient statistic for θ. The specified likelihood function can be interpreted as a pseudo-likelihood function that is only used to provide an estimating equation for the parameters of interest. As a consequence, we argue that model fit should not play a (primary) role in choosing psychometric models for LSA studies. Also, note that the likelihood-based inference (i.e., standard errors) obtained from a misspecified model will also be incorrect. We believe that resampling techniques for persons (see Model-assisted design-based inference for persons section) and items (see Design-based or model-based inference for items? section) allow valid statistical inference, even if the model is misspecified (Berk et al., 2014; White, 1982). In our view, scaling models for LSA studies should be defended from a design-based perspective. Hence, different researchers might opt for different psychometric models for modeling LSA data if the model fit is not considered the primary criterion. Note that there are typically very different approaches for assessing model fit. Depending on how model fit is defined, the complexity of chosen psychometric models will vary considerably. Hence, there will also be disagreement among psychometricians with respect to model choice in the case that model fit would serve as the main criterion.

The concept of model weighting (or model uncertainty) can quantify the extent of consequences in an uncertain space of models (Robitzsch, 2022b; Simonsohn et al., 2020; Young & Holsteen, 2017). For example, it might be beneficial to study the sensitivity of trend estimates for a country for different choices of the linking function. Researchers would be less confident in trend estimates that strongly depend on the chosen estimation method.

Appendix 1

Locally optimal item weights

In this appendix, we derive the locally optimal item weight that is based on the individual log-likelihood function $l_n(\theta) = \sum_{i=1}^{I} l_{ni}(\theta)$. The log-likelihood contribution $l_{ni}(\theta)$ of item i for person n is given by:

$$l_{ni}(\theta) = x_{ni} \log P_i(\theta) + (1 - x_{ni}) \log (1 - P_i(\theta)) \quad (42)$$

We now derive a Taylor approximation of $l_n(\theta)$ around an ability value θ_0 for deriving the contribution of items in a local sufficient statistic for θ. The derivative l_{ni} with respect to θ is given by:

$$\frac{\mathrm{d}l_{ni}}{\mathrm{d}\theta} = x_{ni} \frac{P_i'(\theta)}{P_i(\theta)} - (1 - x_{ni}) \frac{P_i'(\theta)}{1 - P_i(\theta)} = (x_{ni} - P_i(\theta)) \nu_i(\theta), \quad (43)$$

where item weights $\nu_i(\theta)$ are contributions of items in the weighted sum score and are denoted as locally optimal item weights (Birnbaum, 1968; Chiu & Camilli, 2013), where:

$$\nu_i(\theta) = \frac{P_i'(\theta)}{P_i(\theta)[1 - P_i(\theta)]} \quad (44)$$

Using (43) for a Taylor approximation of the log-likelihood function, we obtain:

$$l_n(\theta) \simeq l_n(\theta_0) + \left(\sum_{i=1}^{I} \nu_i(\theta) x_{ni} \right) (\theta - \theta_0) + \text{const}(\theta) \quad (45)$$

From Eq. (45), it can be seen that $\sum_{i=1}^{I} \nu_i(\theta) x_{ni}$ is a local sufficient statistic for θ. We now derive the locally optimal item weights for the 3PL model (see Eq. (24)). It holds that:

$$P_i'(\theta) = (1 - g_i) a_i \Psi(a_i(\theta - b_i))[1 - \Psi(a_i(\theta - b_i))] \quad (46)$$

We now use the short notation $\psi = \Psi(a_i(\theta - b_i))$. Then, we obtain:

$$\begin{aligned} \nu_i(\theta) &= \frac{P_i'(\theta)}{P_i(\theta)[1 - P_i(\theta)]} \\ &= \frac{(1 - g_i) a_i \psi (1 - \psi)}{[g_i + (1 - g_i)\psi](1 - g_i)(1 - \psi)} \\ &= \frac{a_i}{1 + g_i (1 - \psi) \psi^{-1}} \end{aligned} \quad (47)$$

A further simplification of (47) provides:

$$\nu_i(\theta) = \frac{a_i}{1 + g_i \exp(-a_i(\theta - b_i))} \quad (48)$$

For the 2PL model, we get $\nu_i(\theta) = a_i$ because it holds that $g_i = 0$. Furthermore, we get $\nu_i(\theta) = 1$ in the 1PL model.

Appendix 2

Local item scores in the SA+O model

In this appendix, we derive local item contributions for the ability score θ for the SA+O model studied in Pohl et al. (2021). The log-likelihood contribution for student n and item i in the reparametrized SA+O model (see Eqs. (30) and (31)) is given by:

$$\begin{aligned} l_{ni}(\theta) = \text{const}(\theta) &+ x_{ni} r_{ni} a_i \theta + r_{ni} \alpha_{\xi\theta} \gamma_i \theta \\ &+ (1 - r_{ni}) \log (1 + \exp (a_i (\theta - b_i))) \\ &+ r_{ni} t_{ni} \lambda_{i1} \alpha_{\eta_1 \theta} \theta + (1 - r_{ni}) t_{ni} \lambda_{i0} \alpha_{0\theta} \theta \end{aligned} \quad (49)$$

Using a Taylor approximation around $\theta = \theta_0$, we obtain:

$$\log (1 + \exp (a_i (\theta - b_i))) \simeq \text{const}(\theta_0) + a_i \Psi(a_i (\theta - b_i)) (\theta - \theta_0) \quad (50)$$

Where $\text{const}(\theta_0)$ is a function of θ_0. Using the approximation in Eq. (50), it can be seen that the multiplication factors of θ in Eq. (49) are given by:

$$\begin{aligned} &x_{ni} r_{ni} a_i + r_{ni} \alpha_{\xi\theta} \gamma_i + (1 - r_{ni}) a_i P_i(\theta) \\ &\quad + r_{ni} t_{ni} \lambda_{i1} \alpha_{\eta_1 \theta} + (1 - r_{ni}) t_{ni} \lambda_{i0} \alpha_{\eta_0 \theta}, \end{aligned} \quad (51)$$

where $P_i(\theta) = \Psi(a_i(\theta - b_i))$. We can extract the local item scores for θ from Eq. (51):

$$\begin{aligned} &\text{Observed item responses } (r_{ni} = 1): \; a_i + \alpha_{\xi\theta} \gamma_i + \alpha_{\eta_1 \theta} \lambda_{i1} t_{ni} \\ &\text{Omitted item responses } (r_{ni} = 0): \; a_i P_i(\theta) + 0 + \alpha_{\eta_0 \theta} \lambda_{i0} t_{ni}. \end{aligned} \quad (52)$$

These statistics are defined on the logit metric and are unique up to the addition of a constant c; that is:

$$\begin{aligned} \mathrm{P}(X_{ni} = 1) &= \frac{\exp(p_1)}{\exp(p_0) + \exp(p_1)} \\ &= \frac{\exp(p_1 + c)}{\exp(p_0 + c) + \exp(p_1 + c)} \end{aligned} \quad (53)$$

By defining $c = -(\alpha_{\xi\theta} \gamma_i + \alpha_{\eta_1 \theta} \lambda_{i1} t_{ni})$, the local item scores for observed and omitted item responses in Eq. (52) can be equivalently rewritten:

$$\begin{aligned} &\text{Observed item responses } (r_{ni} = 1): \; a_i \\ &\text{Omitted item responses } (r_{ni} = 0): \; a_i P_i(\theta) - \alpha_{\xi\theta} \gamma_i - (\alpha_{\eta_1 \theta} \lambda_{i1} - \alpha_{\eta_0 \theta} \lambda_{i0}) t_{ni} \end{aligned} \quad (54)$$

Abbreviations
1PL: One-parameter logistic; 2PL: Two-parameter logistic; 3PL: Three-parameter logistic; 4PL: Four-parameter logistic; CTT: Classical test theory; DIF: Differential items functioning; GT: Generalizability theory; IRF: Item response function; IRT: Item response theory; LBM: Latent background model; LE: Linking error; LSA: Large-scale assessments; MIRT: Multidimensional item response theory; PIRLS: Progress In International Reading Literacy Study; PISA: Programme for International Student Assessment; RMSD: Root mean square deviation; SA+O: Speed-accuracy and omission; SE: Standard error; TE: Total error; TIMSS: Trends in International Mathematics and Science Study.

Acknowledgements
We would like to thank Jules Ellis, Wolfgang Wagner, Sebastian Weirich, and Margaret Wu for the valuable comments on a previous version of this paper. The authors are responsible for any errors or incorrect statements in this article.

Authors' contributions
Both authors have made a substantial, direct, and intellectual contribution to the work and approved the final version for publication.

Funding
Open Access funding enabled and organized by Projekt DEAL. There is no external funding.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]IPN — Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany. [2]Centre for International Student Assessment (ZIB), Kiel, Germany.

Received: 6 September 2021 Accepted: 25 July 2022
Published online: 03 September 2022

References

Adams, R. J. (2003). Response to 'Cautions on OECD's recent educational survey (PISA)'. *Oxford Review of Education*, *29*(3), 379–389. https://doi.org/10.1080/03054980307445.

Aitkin, M. & Aitkin, I. (2006). Investigation of the identifiability of the 3PL model in the NAEP 1986 math survey. Technical report. https://bit.ly/35b79X0

Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., & Zhao, L. (2014). Misspecified mean function regression: Making good use of regression models that are wrong. *Sociological Methods & Research*, *43*(3), 422–451. https://doi.org/10.1177/0049124114526375.

Binder, D. A., & Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. In R. L. Chambers, & C. J. Skinner (Eds.), *Analysis of survey data*, (pp. 29–48). Wiley. https://doi.org/10.1002/0470867205.ch3.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental test scores*, (pp. 397–479). MIT Press.

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, *26*(4), 364–375. https://doi.org/10.1177/014662102237794.

Bolsinova, M., Tijmstra, J., Molenaar, D., & De Boeck, P. (2017). Conditional dependence between response time and accuracy: An overview of its possible sources and directions for distinguishing between them. *Frontiers in Psychology*, *8*, 202. https://doi.org/10.3389/fpsyg.2017.00202.

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*, *51*(2), 141–162. https://doi.org/10.1111/jedm.12039.

Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference*. Springer. https://doi.org/10.1007/978-1-4614-4818-1.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153–168. https://doi.org/10.1007/BF02294533.

Brennan, R. L. (1998). Misconceptions at the intersection of measurement theory and practice. *Educational Measurement: Issues and Practice*, *17*, 5–9. https://doi.org/10.1111/j.1745-3992.1998.tb00615.x.

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417.

Brewer, K. (2013). Three controversies in the history of survey sampling. *Survey Methodology*, *39*(2), 249–262 https://bit.ly/3mhYPxx.

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). International surveys of educational achievement: How robust are the findings? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(3), 623–646. https://doi.org/10.1111/j.1467-985X.2006.00439.x.

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning: Theory and practice*, (pp. 397–417). Erlbaum. https://doi.org/10.4324/9780203357811.

Camilli, G. (2018). IRT scoring and test blueprint fidelity. *Applied Psychological Measurement*, *42*(5), 393–400. https://doi.org/10.1177/0146621618754897.

Carstensen, C. H. (2013). Linking PISA competencies over three cycles – Results from Germany. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research on PISA*, (pp. 199–213). Springer. https://doi.org/10.1007/978-94-007-4458-5_12.

Chandler, R. E., & Bate, S. (2007). Inference for clustered data using the independence loglikelihood. *Biometrika*, *94*(1), 167–183. https://doi.org/10.1093/biomet/asm015.

Chiu, T. W., & Camilli, G. (2013). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement*, *37*(1), 76–86. https://doi.org/10.1177/0146621612459369.

Conijn, J. M., Emons, W. H., van Assen, M. A., & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research*, *46*(2), 365–388. https://doi.org/10.1080/00273171.2010.546733.

Costa, D. R., Bolsinova, M., Tijmstra, J., & Andersson, B. (2021). Improving the precision of ability estimates using time-on-task variables: Insights from the PISA 2012 computer-based assessment of mathematics. *Frontiers in Psychology*, *12*, 579128. https://doi.org/10.3389/fpsyg.2021.579128.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology*, *16*, 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x.

Cronbach, L. J., Schoenemann, P., & McKie, D. (1965). Alpha coefficient for stratified-parallel tests. *Educational and Psychological Measurement*, *25*, 291–312. https://doi.org/10.1177/001316446502500201.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. https://doi.org/10.1177/0013164404266386.

Culpepper, S. A. (2017). The prevalence and implications of slipping on low-stakes, large-scale assessments. *Journal of Educational and Behavioral Statistics*, *42*(6), 706–725. https://doi.org/10.3102/1076998617705653.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*(1), 55–75. https://doi.org/10.1146/annurev-soc-071913-043137.

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. https://doi.org/10.1111/jedm.12009.

Deribo, T., Kroehne, U., & Goldhammer, F. (2021). Model-based treatment of rapid guessing. *Journal of Educational Measurement*, *58*(2), 281–303. https://doi.org/10.1111/jedm.12290.

Dimitrov, D. M. (2016). An approach to scoring and equating tests with binary items: Piloting with large-scale assessments. *Educational Psychological Measurement*, *76*(6), 954–975. https://doi.org/10.1177/0013164416631100.

El Masri, Y. H., & Andrich, D. (2020). The trade-off between model fit, invariance, and validity: The case of PISA science assessments. *Applied Measurement in Education*, *33*(2), 174–188. https://doi.org/10.1080/08957347.2020.1732384.

Ellis, J. L. (2021). A test can have multiple reliabilities. *Psychometrika*, *86*(4), 869–876. https://doi.org/10.1007/s11336-021-09800-2.

Ellis, J. L., & Junker, B. W. (1997). Tail-measurability in monotone latent variable models. *Psychometrika*, *62*(4), 495–523. https://doi.org/10.1007/BF02294640.

Falk, C. F., & Cai, L. (2016). Maximum marginal likelihood estimation of a monotonic polynomial generalized partial credit model with applications to multiple group analysis. *Psychometrika*, *81*(2), 434–460. https://doi.org/10.1007/s11336-014-9428-7.

Ferrando, P. J. (2019). A comprehensive IRT approach for modeling binary, graded, and continuous responses with error in persons and items. *Applied Psychological Measurement*, *43*(5), 339–359. https://doi.org/10.1177/0146621618817779.

Feuerstahler, L. M. (2019). Metric transformations and the filtered monotonic polynomial item response model. *Psychometrika*, *84*(1), 105–123. https://doi.org/10.1007/s11336-018-9642-9.

Fox, J.-P. (2010). *Bayesian item response modeling*. Springer. https://doi.org/10.1007/978-1-4419-0742-4.

Fox, J.-P., & Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis: Methods and applications*, (pp. 461–482). Routledge Academic.

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. Mullis (Eds.), *TIMSS 2019 technical report*. Boston College: IEA.

Foy, P., & Yin, L. (2017). Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016*. Boston College: IEA.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53. https://doi.org/10.1111/j.1745-3992.2009.00154.x.

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement*, *8*(3), 305–322 https://bit.ly/2UDjWib.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, *33*(2), 234–246. https://doi.org/10.1111/j.2044-8317.1980.tb00610.x.

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Canadian Journal of Forest Research*, *28*(10), 1429–1447. https://doi.org/10.1139/x98-166.

Grund, S., Lüdtke, O., & Robitzsch, A. (2021). On the treatment of missing data in background questionnaires in educational large-scale assessments: An evaluation of different procedures. *Journal of Educational and Behavioral Statistics*, *46*(4), 430–465. https://doi.org/10.3102/1076998620959058.

Haberkorn, K., Pohl, S., & Carstensen, C. (2016). Scoring of complex multiple choice items in NEPS competence tests. In H. P. Blossfeld, J. von Maurice, M. Bayer, & J. Skopek (Eds.), *Methodological issues of longitudinal surveys*. Springer VS. https://doi.org/10.1007/978-3-658-11994-2_29.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 301–321. https://doi.org/10.1111/j.1745-3984.1989.tb00336.x.

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 369–385. https://doi.org/10.1080/0969594X.2018.1469467.

He, J., Van de Vijver, F. J. R., Fetvadjiev, V. H., de Carmen Dominguez Espinosa, A., Adams, B., Alonso-Arbiol, I., … Hapunda, G. (2017). On enhancing the cross–cultural comparability of Likert–scale personality and value measures: A comparison of common procedures. *European Journal of Personality*, *31*(6), 642–657. https://doi.org/10.1002/per.2132.

He, Y., & Cui, Z. (2020). Evaluating robust scale transformation methods with multiple outlying common items under IRT true score equating. *Applied Psychological Measurement*, *44*(4), 296–310. https://doi.org/10.1177/0146621619886050.

Holland, P. W., & Wainer, H. (Eds.) (1993). *Differential item functioning: Theory and practice*. Erlbaum. https://doi.org/10.4324/9780203357811.

Hong, M. R., & Cheng, Y. (2019). Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods*, *51*(2), 573–588. https://doi.org/10.3758/s13428-018-1150-4.

Jerrim, J., Parker, P., Choi, A., Chmielewski, A. K., Sälzer, C., & Shure, N. (2018). How robust are cross-country comparisons of PISA scores to the scaling model used? *Educational Measurement: Issues and Practice*, *37*(4), 28–39. https://doi.org/10.1111/emip.12211.

Jin, K. Y., & Wang, W. C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement*, *51*(2), 178–200. https://doi.org/10.1111/jedm.12041.

Joo, S. H., Khorramdel, L., Yamamoto, K., Shin, H. J., & Robin, F. (2021). Evaluating item fit statistic thresholds in PISA: Analysis of cross-country comparability of cognitive items. *Educational Measurement: Issues and Practice*, *40*(2), 37–48. https://doi.org/10.1111/emip.12404.

Kane, M. (1982). A sampling model for validity. *Applied Psychological Measurement*, *6*(2), 125–160. https://doi.org/10.1177/014662168200600201.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, *47*(2), 267–292. https://doi.org/10.3102/00346543047002267.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *Stata Journal*, *10*(2), 165–199. https://doi.org/10.1177/1536867X1001000201.

Lechner, C. M., Bhaktha, N., Groskurth, K., & Bluemke, M. (2021). Why ability point estimates can be pointless: A primer on using skill measures from large-scale assessments in secondary analyses. *Measurement Instruments for the Social Sciences*, *3*, 2. https://doi.org/10.1186/s42409-020-00020-5.

Liao, X., & Bolt, D. M. (2021). Item characteristic curve asymmetry: A better way to accommodate slips and guesses than a four-parameter model? *Journal of Educational and Behavioral Statistics*, *46*(6), 753–775. https://doi.org/10.3102/10769986211003283.

Liou, M., & Yu, L. C. (1991). Assessing statistical accuracy in ability estimation: A bootstrap approach. *Psychometrika*, *56*(1), 55–67. https://doi.org/10.1007/BF02294585.

Little, R. J. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, *99*(466), 546–556. https://doi.org/10.1198/016214504000000467.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley. https://doi.org/10.1002/9781119013563.

Lohr, S. L. (2010). *Sampling: Design and analysis*. Brooks/Cole Cengage Learning.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–525. https://doi.org/10.1348/000711009X474502.

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. Applied Psychological Measurement, 37(4), 304–315. doi: https://doi.org/10.1177/0146621613475471

Magis, D. (2015). A note on the equivalence between observed and expected information functions with polytomous IRT models. *Journal of Educational and Behavioral Statistics*, *40*(1), 96–105. https://doi.org/10.3102/1076998614558122.

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent type I error inflation in differential item functioning. *Educational and Psychological Measurement*, *72*(2), 291–311. https://doi.org/10.1177/0013164411416975.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three parameter logistic model. *Measurement: Interdisciplinary Research and Perspectives*, *7*(2), 75–88. https://doi.org/10.1080/15366360903070385.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge. https://doi.org/10.4324/9780203501207.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *Psychometrika*, *81*(2), 274–289. https://doi.org/10.1007/s11336-016-9497-x.

Martin, M. O., Mullis, I. V., Foy, P., Brossman, B., & Stanco, G. M. (2012). Estimating linking error in PIRLS. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *5*, 35–47 https://bit.ly/3yraNrd.

McDonald, R. P. (1978). Generalizability in factorable domains: "Domain validity and generalizability". *Educational and Psychological Measurement*, *38*(1), 75–79. https://doi.org/10.1177/001316447803800111.

McDonald, R. P. (2003). Behavior domains in theory and in practice. *Alberta Journal of Educational Research*, *49*(3), 212–230 https://bit.ly/3O4s2I5.

Meinck, S. (2020). Sampling, weighting, and variance estimation. In H. Wagenmaker (Ed.), *Reliability and validity of international large-scale assessment*, (pp. 113–129). Springer. https://doi.org/10.1007/978-3-030-53081-5_7.

Meyer, P. (2010). *Understanding measurement: Reliability*. Oxford University Press.

Michaelides, M. P., & Haertel, E. H. (2014). Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education*, *27*, 46–57. https://doi.org/10.1080/08957347.2013.853069.

Mislevy, R. (1990). Scaling procedures. In E. Johnson, & R. Zwick (Eds.), *Focusing the new design: The NAEP 1988 technical report (ETS RR 19-20)*. Educational Testing Service https://bit.ly/3zuC5OQ.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196. https://doi.org/10.1007/BF02294457.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monographs Series. Issues and Methodologies in Large-Scale Assessments.*, *4*, 131–158 https://bit.ly/3k6wlyU.

Monseur, C., Sibberns, H., & Hastedt, D. (2008). Linking errors in trend estimation for international surveys in education. *IERI Monographs Series. Issues and Methodologies in Large-Scale Assessment*, *1*, 113–122 https://bit.ly/38aTVeZ.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.

OECD (2014). *PISA 2012 technical report*. OECD Publishing.

OECD (2017). *PISA 2015 technical report*. OECD Publishing.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333 https://bit.ly/3mkaRGO.

Pellegrino, J. W., & Chudowsky, N. (2003). The foundations of assessment. *Measurement: Interdisciplinary Research and Perspectives*, *1*(2), 103–148. https://doi.org/10.1207/S15366359MEA0102_01.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics*, (pp. 125–167). Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the national educational panel study - Many questions, some answers, and further challenges. *Journal of Educational Research Online*, *5*(2), 189–216.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests: Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*(3), 423–452. https://doi.org/10.1177/0013164413504926.

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science*, *372*(6540), 338–340. https://doi.org/10.1126/science.abd3300.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004.

Raiche, G., Magis, D., Blais, J. G., & Brochu, P. (2012). Taking atypical response patterns into account: A multidimensional measurement model from item response theory. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education*, (pp. 238–259). Routledge. https://doi.org/10.4324/9780203154519.

Ramsay, J. O., & Winsberg, S. (1991). Maximum marginal likelihood estimation for semiparametric item analysis. *Psychometrika*, *56*(3), 365–379. https://doi.org/10.1007/BF02294480.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Reckase, M. D. (2017). *A tale of two models: Sources of confusion in achievement testing. ETS Research Report, ETS RR-17-44*. https://doi.org/10.1002/ets2.12171.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555.

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing: A meta-analysis of interventions. *Applied Measurement in Education*, *34*(2), 85–106. https://doi.org/10.1080/08957347.2021.1890741.

Robitzsch, A. (2020). L_p loss functions in invariance alignment and Haberman linking with few or many groups. *Stats*, *3*(3), 246–283. https://doi.org/10.3390/stats3030019.

Robitzsch, A. (2021a). A comparison of linking methods for two groups for the two-parameter logistic item response model in the presence and absence of random differential item functioning. *Foundations*, *1*(1), 116–144. https://doi.org/10.3390/foundations1010009.

Robitzsch, A. (2021b). On the treatment of missing item responses in educational large-scale assessment data: An illustrative simulation study and a case study using PISA 2018 mathematics data. *European Journal of Investigation in Health, Psychology and Education*, *11*(4), 1653–1687. https://doi.org/10.3390/ejihpe11040117.

Robitzsch, A. (2021c). Robust and nonrobust linking of two groups for the Rasch model with balanced and unbalanced random DIF: A comparative simulation study and the simultaneous assessment of standard errors and linking errors with resampling techniques. *Symmetry*, *13*(11), 2198. https://doi.org/10.3390/sym13112198.

Robitzsch, A. (2022a). Estimation methods of the multiple-group one-dimensional factor model: Implied identification constraints in the violation of measurement invariance. *Axioms*, *11*(3), 119. https://doi.org/10.3390/axioms11030119.

Robitzsch, A. (2022b). Exploring the multiverse of analytical decisions in scaling educational large-scale assessment data: A specification curve analysis for PISA 2018 mathematics data. *European Journal of Investigation in Health, Psychology and Education*, *12*(7), 731–753. https://doi.org/10.3390/ejihpe12070054.

Robitzsch, A. (2022c). On the choice of the item response model for scaling PISA data: Model selection based on information criteria and quantifying model uncertainty. *Entropy*, *24*(6), 760. https://doi.org/10.3390/e24060760.

Robitzsch, A., & Lüdtke, O. (2019). Linking errors in international large-scale assessments: calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 444–465. https://doi.org/10.1080/0969594X.2018.1433633.

Robitzsch, A., & Lüdtke, O. (2020a). A review of different scaling approaches under full invariance, partial invariance, and noninvariance for cross-sectional country comparisons in large-scale assessments. *Psychological Test and Assessment Modeling*, *62*(2), 233–279 https://bit.ly/3kFiXaH.

Robitzsch, A., & Lüdtke, O. (2020b). Ein Linking verschiedener Linkingfehler-Methoden in PISA [Linking different linking errors] [Conference presentation]. In *Virtual ZIB Colloqium. Munich, Zoom, November 2020*.

Robitzsch, A., & Lüdtke, O. (2021). Comparing different trend estimation approaches in international large-scale assessment studies [Conference presentation]. In *6th International NEPS Conference (Virtual), Bamberg, Zoom, June 2021*.

Robitzsch, A., & Lüdtke, O. (2022). Mean comparisons of many groups in the presence of DIF: An evaluation of linking and concurrent scaling approaches. *Journal of Educational and Behavioral Statistics*, *47*(1), 36–68. https://doi.org/10.3102/10769986211017479.

Robitzsch, A., Lüdtke, O., Goldhammer, F., Kroehne, U., & Köller, O. (2020). Reanalysis of the German PISA data: A comparison of different approaches for trend estimation with a particular emphasis on mode effects. *Frontiers in Psychology*, *11*, 884. https://doi.org/10.3389/fpsyg.2020.00884.

Rohwer, G. (2013). *Making sense of missing answers in competence tests. NEPS working paper no. 30*. Otto-Friedrich-Universität, Nationales Bildungspanel https://bit.ly/3kzmEPc.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*(3), 795–819. https://doi.org/10.1007/s11336-016-9544-7.

Rust, K. F., Krawchuk, S., & Monseur, C. (2017). Sample design, weighting, and calculation of sampling variance. In P. Lietz, J. C. Creswell, K. F.

Rust, & R. J. Adams (Eds.), *Implementation of large-scale education assessments*, (pp. 137–167). Wiley. https://doi.org/10.1002/9781118762462.ch5.

Rutkowski, L., & Rutkowski, D. (2019). Methodological challenges to measuring heterogeneous populations internationally. In L. E. Suter, E. Smith, & B. D. Denman (Eds.), *The SAGE handbook of comparative studies in education*, (pp. 126–140). Sage. https://doi.org/10.4135/9781526470379.

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.) (2013). *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Chapman Hall/CRC Press. https://doi.org/10.1201/b16061.

Sachse, K. A., Mahler, N., & Pohl, S. (2019). When nonresponse mechanisms change: Effects on trends and group comparisons in international large-scale assessments. *Educational and Psychological Measurement*, *79*(4), 699–726. https://doi.org/10.1177/0013164419829196.

Sachse, K. A., Roppelt, A., & Haag, N. (2016). A comparison of linking methods for estimating national trends in international comparative large-scale assessments in the presence of cross-national DIF. *Journal of Educational Measurement*, *53*(2), 152–171. https://doi.org/10.1111/jedm.12106.

San Martín, E., González, J., & Tuerlinckx, F. (2015). On the unidentifiability of the fixed-effects 3PL model. *Psychometrika*, *80*(2), 450–467. https://doi.org/10.1007/s11336-014-9404-2.

Särndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer. https://doi.org/10.1007/978-1-4612-4378-6.

Schuster, C., & Yuan, K. H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics*, *36*(6), 720–735. https://doi.org/10.3102/1076998610396890.

Shealy, R., & Stout, W. A. (1993). Model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194. https://doi.org/10.1007/BF02294572.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z.

Singer, J. D., & Braun, H. I. (2018). Testing international education assessments. *Science*, *360*(6384), 38–40. https://doi.org/10.1126/science.aar4952.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., … Gregoire, T. G. (2016). Use of models in large-area forest surveys: Comparing model-assisted, model-based and hybrid estimation. *Forest Ecosystems*, *3*, 5. https://doi.org/10.1186/s40663-016-0064-9.

Stenner, A. J., Burdick, D. S., & Stone, M. H. (2008). Formative and reflective models: Can a Rasch analysis tell the difference? *Rasch Measurement Transactions*, *22*(1), 1152–1153 https://www.rasch.org/rmt/rmt221d.htm.

Stenner, A. J., Stone, M. H., & Burdick, D. S. (2009). Indexing vs. measuring. *Rasch Measurement Transactions*, *22*(4), 1176–1177 https://www.rasch.org/rmt/rmt224b.htm.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*(2), 293–325. https://doi.org/10.1007/BF02295289.

Tijmstra, J., Liaw, Y., Bolsinova, M., Rutkowski, L., & Rutkowski, D. (2020). Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *Journal of Educational Measurement*, *57*(4), 566–583. https://doi.org/10.1111/jedm.12263.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin*, *54*(3), 229–249. https://doi.org/10.1037/h0047980.

Uher, J. (2021). Psychometrics is not measurement: Unraveling a fundamental misconception in quantitative psychology and the complex network of its underlying fallacies. *Journal of Theoretical and Philosophical Psychology*, *41*(1), 58–84. https://doi.org/10.1037/teo0000176.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020a). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, *73*(S1), 83–112. https://doi.org/10.1111/bmsp.12188.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020b). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*, *55*(3), 425–453. https://doi.org/10.1080/00273171.2019.1643699.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002.

von Davier, M. (2009). Is there need for the 3PL model? Guess what? *Measurement: Interdisciplinary Research and Perspectives*, *7*(2), 110–114. https://doi.org/10.1080/15366360903117079.

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–705. https://doi.org/10.3102/1076998619881789.

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment*, (pp. 155–174). CRC Press. https://doi.org/10.1201/b16061.

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., … Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 466–488. https://doi.org/10.1080/0969594X.2019.1586642.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*(4), 339–368. https://doi.org/10.3102/10769986012004339.

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, *45*(3), 373–391. https://doi.org/10.1007/BF02293910.

Westfall, P. H., Henning, K. S., & Howell, R. D. (2012). The effect of error correlation on interfactor correlation in psychometric measurement. *Structural Equation Modeling*, *19*(1), 99–117. https://doi.org/10.1080/10705511.2012.634726.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. https://doi.org/10.2307/1912526.

Wise, S. L. (2020). Six insights regarding test-taking disengagement. *Educational Research and Evaluation*, *26*(5-6), 328–338. https://doi.org/10.1080/13803611.2021.1963942.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2-3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005.

Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*, 15–27. https://doi.org/10.1111/j.1745-3992.2010.00190.x.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement*, (pp. 111–154). Praeger Publishers.

Young, C., & Holsteen, K. (2017). Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research*, *46*(1), 3–40. https://doi.org/10.1177/0049124115610347.

Zieger, L., Sims, S., & Jerrim, J. (2019). Comparing teachers' job satisfaction across countries: A multiple-pairwise measurement invariance approach. *Educational Measurement: Issues and Practice*, *38*(3), 75–85. https://doi.org/10.1111/emip.12254.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. https://doi.org/10.1007/s11336-003-0974-7.

Zwitser, R. J., Glaser, S. S. F., & Maris, G. (2017). Monitoring countries in a changing world: A new look at DIF in international surveys. *Psychometrika*, *82*(1), 210–232. https://doi.org/10.1007/s11336-016-9543-8.

Publisher's Note