Robitzsch and Lüdtke
Large-scale Assessments in Education (2023) 11:26
https://doi.org/10.1186/s40536-023-00176-6

Large-scale Assessments in Education

METHODOLOGY

Open Access

Comparing different trend estimation approaches in country means and standard deviations in international large-scale assessment studies

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

*Correspondence: robitzsch@leibniz-ipn.de

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

Abstract

One major aim of international large-scale assessments (ILSA) like PISA is to monitor changes in student performance over time. To accomplish this task, a set of common items (i.e., link items) is repeatedly administered in each assessment. Linking methods based on item response theory (IRT) models are used to align the results from the different assessments on a common scale. This work employs the one-parameter logistic (1PL) and the two-parameter logistic (2PL) IRT models as scaling models for dichotomous item response data. The present article discusses different types of trend estimates in country means and standard deviations for countries in ILSA. These types differ in three aspects. First, the trend can be assessed by an indirect or direct linking approach for linking a country's performance at an international metric. Second, the linking for the trend estimation can rely on either all items or only the link items. Third, item parameters can be assumed to be invariant or noninvariant across countries. It is shown that the most often employed trend estimation methods of original trends and marginal trends can be conceived as particular cases of indirect and direct linking approaches, respectively. Through a simulation study and analytical derivations, it is demonstrated that trend estimates using a direct linking approach and those that rely on only link items outperformed alternatives for the 1PL model with uniform country differential item functioning (DIF) and the 2PL model with uniform and nonuniform country DIF. We also illustrated the performance of the different scaling models for assessing the PISA trend from PISA 2006 to PISA 2009 in the cognitive domains of reading, mathematics, and science. In this empirical application, linking errors based on jackknifing testlets were utilized that adequately quantify DIF effects in the uncertainty of trend estimates.

Keywords: Large-scale assessment, Trend estimation, Item response models, Linking, Differential item functioning, Item parameter drift, Original trend, Marginal trend

Introduction

One primary aim of international large-scale assessments (ILSAs; Rutkowski & Rutkowski, 2022; Rutkowski et al., 2014) is to monitor changes in the levels of educational outcomes (e.g., student performance). Every three years, for example, the Programme for International Student Assessment (PISA) delivers international comparisons of student performance in three cognitive domains (reading, mathematics, and science; OECD, 2014). The repeated assessment of these cognitive domains makes it possible to estimate national trends within each participating country, providing policymakers with important information for evaluating educational reforms; these trends often receive considerable attention from the media. To estimate the trends in student performance, the assessment outcomes need to be reported on a common scale that is comparable across time (Mazzeo & von Davier, 2014). To accomplish this task, a set of common items is repeatedly administered in each assessment, and linking methods are used to align the results from the different assessments on a common scale. It is vital to use reliable and interpretable trend estimates.

In the present article, we discuss different trend estimation approaches in ILSAs. In the literature, competitive trend estimation approaches for country means and standard deviations have been discussed (Carstensen et al., 2008; Gebhardt & Adams, 2007; Robitzsch & Lüdtke, 2019; Sachse & Haag, 2017). These different trend estimation approaches can differ in three aspects. First, the trend estimate can be assessed by an indirect or direct linking approach for linking a country's performance at an international reference metric. In an indirect linking approach, the trend estimate for assessing a country's change between two time points is computed by twice linking the country to the international metric. In contrast, the direct linking approach only links a particular country to the international metric at time point 1 and directly assesses the change in the country mean or the country standard deviation by applying a linking from the first to the second time point based on data at the second time point only from the respective country. Second, the linking for the trend estimation can rely on either all items or only the link items. Using only the subset of link items could result in less precise trend estimates because part of the data is ignored (i.e., the items that are uniquely administered at one of the two time points) in linking. On the other hand, restricting the set of items to link items can reduce the variability in trend estimates if item-specific factors such as differential item functioning (DIF) impact the linking approach. Third, item parameters can be assumed invariant or noninvariant across countries. In the literature, it is often argued that the invariance assumption would lead to more stable trend estimates (von Davier et al., 2019).

It is shown that the most often employed trend estimation methods of original trends and marginal trends can be conceived as particular cases of indirect and direct linking approaches, respectively. Through a simulation study and analytical derivations, it is shown that trend estimates using a direct linking approach and those that rely on only link items outperformed alternatives for the 1PL model with uniform country DIF and the 2PL model with uniform and nonuniform country DIF. We also illustrated the performance of the different scaling models for assessing the PISA trend from PISA 2006 to PISA 2009 in the cognitive domains of reading, mathematics,

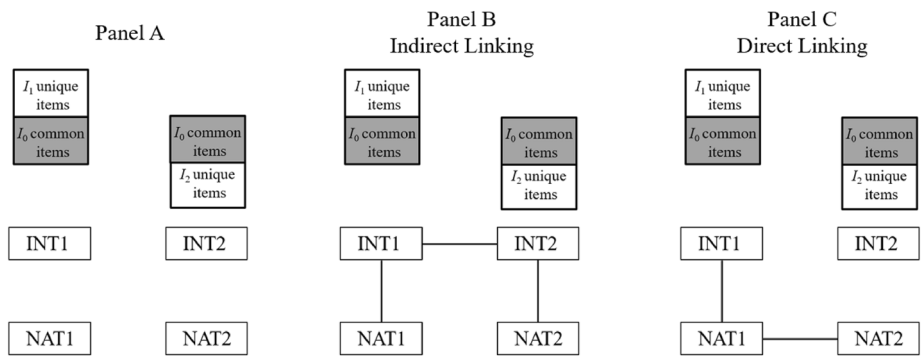

Fig. 1 Assessment designs for two time points

and science. In this empirical application, linking errors based on jackknifing testlets were computed that quantify DIF effects in the uncertainty of trend estimates.

Trend estimation in ILSA studies

In order to estimate country trends in student achievement, the results from different assessments need to be linked so that the achievement scores in a respective cognitive domain can be directly compared. The general idea of a linking approach is to use a set of common items that are administered in more than one assessment in order to establish a common metric that makes it possible to compare the test results across the different assessments (Dorans et al., 2007; Pohl et al., 2015; Sachse et al., 2019; von Davier and Sinharay, 2014). Panel A in Fig. 1 illustrates a typical linking design used in two assessments of an ILSA study. In both assessments, a set of I_0 link items (also referred to as common or anchor items) is administered to a cohort of students (e.g., 15-year-old students in a country; OECD, 2014). In addition, I_1 and I_2 unique items are presented in only one of the two assessments. One advantage of including unique items in an ILSA is that they can be made publically available for secondary analysis, and the item pool can be renewed in later assessments (Mazzeo & von Davier, 2014). In a linking design with two time points, four distributions can be distinguished in Panel A in Fig. 1. The international metric INT is based on all countries or a subset of reference countries (e.g., OECD countries). The corresponding ability distributions at the international metric of the two time points are denoted by INT1 and INT2. Furthermore, a particular country (i.e., nation or NAT) must be linked to the international metric to enable cross-sectional and longitudinal comparisons of the country with the international reference INT or other countries. The ability distributions of the country NAT at the two time points are denoted by NAT1 and NAT2, respectively.

In PISA until 2012, two assessments are linked at an international metric (OECD, 2014). Two calibration samples comprising all (selected) countries at time point 1 and time point 2 are utilized to obtain item parameter estimates from two separate item response models, such as the one-parameter logistic (1PL) model or the two-parameter logistic (2PL) model (Birnbaum, 1968). To place estimated item parameters onto a common international metric in order to assess differences in student abilities across time points, a linking approach (Kolen & Brennan, 2014) can be used. Until PISA 2012, the 1PL model was applied with subsequent mean-mean linking (OECD, 2014). For the 2PL

model, Haebara linking (Haebara, 1980) or Haberman linking (Haberman, 2009) can be used. Panel B in Fig. 1 illustrates the link between the two assessments at the international metric by the line between INT1 and INT2. In a cross-sectional assessment, the countries can be linked to the international metric by assuming common (i.e., invariant or international) item parameters. More specifically, a particular country NAT1 is linked to the international metric INT1 by using the international item parameters as fixed parameters in the country-specific scaling model. By doing so, the ability distribution of a country NAT (i.e., country mean and country standard deviation) can be compared to the international average at time point 1 (see Panel B in Fig. 1). Alternatively, one could employ a separate scaling model for each country at time point 1. In this approach, all item parameters are noninvariant, and a subsequent linking approach must be performed to enable comparisons of countries with each other and the international metric. Moreover, countries have to be linked to the international metric at time point 2. Again, a country (NAT2 in Panel B in Fig. 1) can be linked to the common international metric INT2 by assuming invariant item parameters across countries or relying on country-specific noninvariant item parameters with subsequent linking. The approach depicted in Panel B in Fig. 1 follows the rationale of the original trend estimation because two cross-sectional estimates can be subtracted to obtain a trend estimate (Gebhardt & Adams, 2007). Notably, countries are linked twice in this approach. At both time points, there is a linking of a nation NAT to an international metric, and the link across the two different time points is established by linking international item parameters obtained from calibration samples. The linking illustrated in Panel B in Fig. 1 can be referred to as indirect longitudinal linking because the trend estimate for country NAT is obtained by linking twice to the international metric.

Alternatively, the longitudinal linking of a country can be established with a direct linking approach. This situation is depicted in Panel C in Fig. 1. The linking of countries onto the international metric at time point 1 is the same as the linking at the international metric depicted in Panel B in Fig. 1. However, the crucial difference in the direct linking of a country is that there is no reference to the international metric obtained at time point 2. That is, ability changes for a country NAT at time point 2 are directly assessed as differences to the item parameters NAT1 obtained from the scaling at the first time point. Typically, the link between NAT1 and NAT2 is conducted by applying separate scaling models with subsequent linking (Gebhardt & Adams, 2007). Marginal trend estimates are an example of such a direct linking approach (Gebhardt & Adams, 2007). Importantly, Panel C in Fig. 1 indicates that only one link of a country to the international metric at time point 1 is required. At the same time, the linking approach at the international metric depicted in Panel B in Fig. 1 also requires a linking at the second time point.

Previous literature on trend estimation focuses on particular indirect and direct linking approaches. If all $I_0+I_1+I_2$ items (i.e., "all items") are used in indirect linking (see Panel B in Fig. 1), trend estimates from the corresponding indirect linking method have been termed as original trend estimates (Gebhardt & Adams, 2007). Moreover, if all items are used in the direct linking approach (see Panel C in Fig. 1), trend estimates based on this method are labeled as marginal trend estimates (Gebhardt & Adams, 2007). It has been pointed out that marginal trend estimates (i.e., as an example of direct linking) can

result in more precise (i.e., more efficient) estimates than original trend estimates (i.e., as an example of indirect linking) (see Robitzsch & Lüdtke, 2019; Sachse et al., 2016). Consequently, the direct linking approach could be more efficient than the indirect approach if it cannot be assumed that item parameters are invariant across all countries and the two time points. The consequences of the violation of measurement invariance can be distinguished with regard to three different aspects (Robitzsch & Lüdtke, 2019). First, items can function differently across assessments, known as item parameter drift (IPD; Meade et al., 2005; Hanson & Beguin, 2002; Kang & Petersen, 2012). Second, an item can function differently (differential item functioning; DIF) across countries at one time point, indicating that an item is relatively easier or more difficult for a specific country than at the international level (Camilli, 2006; Holland & Wainer, 1993). These cross-national differences have been studied extensively and termed country DIF (Kreiner & Christensen, 2014; Oliveri & von Davier, 2014, 2017). Third, country DIF can vary across time points, meaning the relative difficulty changes across assessments (DIF × IPD; Carstensen, 2013; Wetzel & Carstensen, 2013).

Furthermore, the magnitude of the efficiency gains by using direct linking instead of indirect linking also depends on the linking design (see Weeks et al., 2014). A good showcase of the relevance of the linking design can be seen in the PISA study with its distinction between major and minor domains (e.g., reading was the major domain in PISA 2009 and, therefore, a large number of reading items were used in PISA 2009, whereas, in PISA 2006, reading was a minor domain and a much smaller number of reading items were used). In this linking design, additional uncertainty in the linking is caused by the substantially lower number of items when a cognitive domain changes from being a major to a minor domain (Weeks et al., 2014).

Cross-sectional estimation of country means and standard deviations

In the following, we discuss the variability of cross-sectional estimates of the country mean and country standard deviations in the 1PL model. The variability comprises standard errors due to the sampling of persons and linking errors due to a random functioning of items[1] (Hastedt & Desa, 2015; Monseur & Berezner, 2007; Monseur et al., 2008). It is assumed that the 1PL model with country-specific DIF effects holds. Notably, the precision of the two country distribution parameters is influenced by two aspects. First, the sample size per country affects the standard error. With larger sample sizes, less variability can be expected. Second, the extent of DIF effects and the number of items affect the precision of country-specific distribution parameters (Robitzsch & Lüdtke, 2019; Wu, 2010).

Assume that the estimated identified item difficulty $\widehat{\beta}_{ic}$ of item i in country c in a country-wise scaling model is given by

$$\widehat{\beta}_{ic} = \mu_c + b_i + \nu_{ic} + e_{ic}, \tag{1}$$

where μ_c is the true country mean, b_i is the common item difficulty of item i, ν_{ic} is the true DIF effect of item i in country c, and e_{ic} is the country-specific sampling error of the

[1] Alternatively, a sampling of items can be regarded as a source of uncertainty (Brennan, 2001; Wu, 2010).

estimated item difficulty. We assume that DIF effects ν_{ic} and sampling errors e_{ic} have an expected value of zero (i.e., $\mathrm{E}(\nu_{ic}) = 0$ and $\mathrm{E}(e_{ic}) = 0$).[2]

It is evident that there are two error terms in Eq. (1). The joint effect $\nu_{ic} + e_{ic}$ confounds sampling error and linking error due to country DIF in a concurrent estimation approach that ignores the presence of DIF effects.

Two estimation approaches will be distinguished: concurrent scaling and separate scaling with subsequent linking. These approaches are now being discussed.

First, concurrent scaling specifies a multiple group model by assuming invariant item parameters. In operational practice, the IRT model is estimated with marginal maximum likelihood (MML; von Davier & Sinharay, 2014). However, we approximate MML by diagonally weighted least squares (DWLS; Cai & Moustaki, 2018) estimation (see Robitzsch & Lüdtke, 2020) in the analytical treatment. DWLS is a limited information estimation approach that fits country-specific thresholds and tetrachoric (or polychoric) correlations of all item pairs. If DWLS approximates ML, the weights of all input statistics are determined as precision weights that are given as the inverse of the variance estimate of a corresponding statistic. For example, the weight of an input country-specific threshold statistic increases with increasing country sample size and if the threshold is close to zero (i.e., it lies in the center of the ability distribution of the country). We want to emphasize that obtaining common item parameters in a concurrent scaling model involving multiple groups with equal contributions from all countries is very similar to the approach that relies on a pooled calibration sample comprising students of all countries. The only difference between the two approaches consists of the specification of the distribution of the latent ability variable. Empirical evidence shows that the specification of this prior distribution is practically inconsequential regarding item parameter estimation. The population distribution of a single country can be obtained in an anchoring approach using fixed item international parameters from the calibration sample (OECD, 2012).

Second, the countries can be separately scaled for each country. The country-specific item parameters are brought onto a common international metric in a subsequent linking procedure (Kolen & Brennan, 2014). Note that the item parameters on the international metric are obtained by estimating a single-group IRT model that includes students from all countries. In contrast to concurrent scaling, the linking approach allows non-invariant item parameters. There is some belief that this property must always result in less stable estimates (von Davier et al., 2019). However, we refute this claim with analytical and simulation results.

In the following, we derive the variability of the country mean and standard deviation estimates for the concurrent scaling and the linking approach. It is shown that the variability depends on the relationship between sampling error variance and linking error due to the variance of DIF effects and the number of items.

[2] Sampling errors are denoted by Latin letters, while Greek letters are used for item effects.

Variance of estimated country means

Let $\widehat{\beta}_{ic} = \mu_c + b_i + \nu_{ic} + e_{ic}$ be the identified threshold of item i in country c (see Eq. (1)). Furthermore, let $\phi_{ic} = \mathrm{Var}(e_{ic})$ be the sampling variance of the estimated item difficulty $\widehat{\beta}_{ic}$. First, we derive the variability of the estimated country mean under concurrent scaling that relies upon invariant item parameters (labeled as "inv"). The MML estimation method is approximated by DWLS. Hence, the country mean is estimated by minimizing the following weighted least squares function for item difficulties (or item thresholds, respectively)

$$\widehat{\mu}_{c,\mathrm{inv}} = \underset{\mu}{\mathrm{argmin}} \sum\nolimits_{i=1}^{I} \omega_{ic} \left(\widehat{\beta}_{ic} - b_i - \mu \right)^2, \tag{2}$$

where the precision weights are given as $\omega_{ic} = 1/\phi_{ic}$. The ML estimate $\widehat{\mu}_{c,\mathrm{inv}}$ is given by

$$\widehat{\mu}_{c,\mathrm{inv}} = \frac{\sum_{i=1}^{I} \omega_{ic} \widehat{\beta}_{ic}}{\sum_{i=1}^{I} \omega_{ic}} = \mu_c + \frac{\sum_{i=1}^{I} \omega_{ic} (\nu_{ic} + e_{ic})}{\sum_{i=1}^{I} \omega_{ic}}. \tag{3}$$

Equation (3) highlights that the two types of errors (i.e., DIF effects and sampling error) affect the estimated country mean. Because ν_{ic} and e_{ic} have expected values of zero and are assumed to be uncorrelated with ω_{ic}, we obtain that $\widehat{\mu}_{c,\mathrm{inv}}$ is an unbiased estimate of μ_c. Moreover, the random variables ν_{ic} and e_{ic} are all uncorrelated.

Let $\sigma^2_{\mathrm{DIF},c}$ be the country-specific variance $\mathrm{Var}(\nu_{ic})$ of the DIF effects ν_{ic}. Hence, the variance of the estimated country mean can be determined as

$$\mathrm{Var}(\widehat{\nu}_{c,\mathrm{inv}}) = \frac{\sum_{i=1}^{I} (\sigma^2_{\mathrm{DIF},c} + \phi_{ic})}{I^2} = \frac{\sigma^2_{\mathrm{DIF},c}}{I} + \frac{m_{\phi,c}}{I} \tag{4}$$

where $m_{\omega,c}$ and $s_{\omega,c}$ denote the empirical mean and standard deviation of precision weights ω_{ic}, respectively. The derivation relies on the assumption that the sampling errors of estimated difficulties are uncorrelated, which is approximately fulfilled for a sufficiently large number of items (Yuan et al., 2014). Also, note that $m_{\omega,c}$ is the harmonic mean of variances ϕ_{ic}; that is, $1/m_{\omega} = 1/m_{1/\phi}$, where $m_{1/\phi}$ is the arithmetic mean of the quantities $1/\phi_{ic}$. It can be shown that the harmonic mean is always smaller than the arithmetic mean (Xia et al., 1999); that is, $\frac{1}{m_{1/\phi}} \leq m_{\phi}$.

The mean-mean linking method is used in the linking approach (Kolen & Brennan, 2014). The estimated country-specific item parameters $\widehat{\beta}_{ic}$ are linked to international item parameters b_i. To ease the analytical derivations, we assumed that international item parameters are obtained without sampling error because, in LSA studies with many countries, the sample size at the international level is sufficiently large compared to the sample size at the level of a particular country. Moreover, DIF effects are frequently more important than sampling errors at the country level. The estimated country mean under linking based on noninvariant item parameters (labeled as "noninv") is given by

$$\widehat{\mu}_{c,\text{noninv}} = \frac{1}{I}\sum_{i=1}^{I}\left(\widehat{\beta}_{ic} - b_i\right) = \mu_c + \frac{1}{I}\sum_{i=1}^{I}(\nu_{ic} + e_{ic}). \tag{5}$$

Note that $\widehat{\mu}_{c,\text{noninv}}$ coincides with the country mean $\widehat{\mu}_{c,\text{inv}}$ obtained from concurrent scaling if all precision weights ω_{ic} in Eq. (3) would be equal. Moreover, the estimated country mean based on noninvariance is also unbiased because the expected values of ν_{ic} and e_{ic} are zero. The variance of the estimated country mean $\widehat{\mu}_{c,\text{noninv}}$ obtained with mean-mean linking is given by

$$\text{Var}\left(\widehat{\mu}_{c,\text{noninv}}\right) = \frac{\sum_{i=1}^{I}\left(\sigma^2_{\text{DIF},c} + \phi_{ic}\right)}{I^2} = \frac{\sigma^2_{\text{DIF},c}}{I} + \frac{m_{\phi,c}}{I}. \tag{6}$$

Hence, one can analyze which of the two estimators $\widehat{\mu}_{\text{inv}}$ and $\widehat{\mu}_{\text{noninv}}$ is more efficient. We compute the difference in variance estimates and obtain

$$\text{Var}\left(\widehat{\mu}_{c,\text{inv}}\right) - \text{Var}\left(\widehat{\mu}_{c,\text{noninv}}\right) = \frac{\sigma^2_{DIF,c}}{I}\frac{s^2_{\omega,c}}{m^2_{\omega,c}} - \frac{1}{I}\left(m_{\phi,c} - \frac{1}{m_{\frac{1}{\phi},c}}\right). \tag{7}$$

The first term in (7) is always positive in the presence of DIF effects (i.e., $\sigma^2_{\text{DIF},c} > 0$). The second term is always negative. Hence, the estimates under invariance (i.e., $\widehat{\mu}_{c,\text{inv}}$) are more efficient than estimates under noninvariance of item parameters (i.e., $\widehat{\mu}_{c,\text{noninv}}$) if the DIF variance $\sigma^2_{\text{DIF},c}$ is small compared to the variability in the precision-weighted estimate. More formally, by using Eq. (7), this property is fulfilled if

$$\sigma^2_{\text{DIF},c} \leq \frac{m^2_{\omega,c}}{s^2_{\omega,c}}\left(m_{\phi,c} - \frac{1}{m_{\frac{1}{\phi},c}}\right) \tag{8}$$

We also want to point out that both estimators (i.e., country means estimated with concurrent scaling and linking) provide unbiased estimates of the country means if one assumes that the DIF effects ν_{ic} vanish on average (i.e., $\text{E}(\nu_{ic}) = 0$) and are uncorrelated with common item difficulties b_i. However, the variances of these estimates generally differ. In contrast to the statement in von Davier et al. (2019), country means based on linking can be more efficient than concurrent scaling if the influence of DIF variance exceeds the effects of sampling error (see Eq. (8)). If there would be no sampling error, an equal weighting of item difficulties as conducted in mean-mean linking for computing the country mean is preferable. In contrast, in the absence of DIF effects, the precision-weighted computation of country means provides the most efficient estimates. This property might motivate the claim that concurrent scaling will always result in more stable estimates. However, we think a complete absence of DIF is unrealistic in practical applications. We would also like to note that the mean-mean linking in Eq. (5) can be replaced by a robust mean that downweighs outlying DIF effects (Magis & De Boeck, 2011; Robitzsch, 2021b; von Davier & Bezirhan, 2023; Wang et al., 2022). By doing so, the variance in estimated country means might be even more reduced. Moreover, a downweighting or complete removal of some

items from linking might sometimes be also preferred for bias reasons if items with large DIF effects can be considered as construct-irrelevant (Camilli, 1993). Such approaches are similar to concurrent scaling approaches relying on partial invariance (Oliveri & von Davier, 2011). The different consequences of removing specific items (e.g., items with large DIF effects) from the linking will be further addressed in the Discussion section.

Variance of estimated country standard deviations

We now derive the variance of the estimated country standard deviations. We still rely on the 1PL model. In the linking approach of the 1PL model, the country-wise estimated standard deviations must not be linked and can be directly compared across countries because the 1PL model defines a common metric. Hence, the two estimation approaches of concurrent scaling with MML and separate scaling with MML (no subsequent linking is required for estimating standard deviations) can be approximated by DWLS. In this approach, estimated tetrachoric correlations are used as the input. If the 1PL model with a standard deviation σ holds, the population tetrachoric correlation of all item pairs (i,j) is given by $\rho_{ij} = \tau = \sigma^2/(1+\sigma^2)$. Because τ is a monotone transformation of the standard deviation σ, we present the analytical findings for the country-specific parameter τ_c of country c because this strongly simplifies resulting formulas.

Let the estimated tetrachoric correlation $\widehat{\rho}_{ijc}$ for item pair (i,j) in country c be a function of observed frequencies $\widehat{p}_{ijc}$ and assumed item difficulties β_i and β_j in a scaling model. More formally, we can write

$$\widehat{\rho}_{ijc} = \rho_{ijc} + \varepsilon_{ijc} = f\left(\beta_i, \beta_j, \tau_c\right) + \varepsilon_{ijc}, \tag{9}$$

where ε_{ijc} is the sampling error of the estimated tetrachoric correlation. The population tetrachoric correlation ρ_{ijc} is a function of assumed item difficulties and the transformed standard deviation τ_c (i.e., $\rho_{ijc} = f\left(\beta_i, \beta_j, \tau_c\right)$). A DWLS estimate of the transformed standard deviation is given by

$$\widehat{\tau}_c = \frac{\sum_{i \neq j} \omega_{ijc} \widehat{\rho}_{ijc}}{\sum_{i \neq j} \omega_{ijc}}, \tag{10}$$

where ω_{ijc} are precision weights of estimated tetrachoric correlations.[3] Under a concurrent scaling model that relies on invariant item parameters, the item difficulties are fixed to the international parameters $b_i = b_{ic} - \nu_{ic}$. Using Eq. (9) and a Taylor expansion, we obtain

$$\widehat{\rho}_{ijc,inv} = f\left(b_{ic} - \nu_{ic}, b_{jc} - \nu_{jc}, \tau_c\right) + \varepsilon_{ijc} \simeq \tau_c - f_{1ijc}\nu_{ic} - f_{2ijc}\nu_{jc} + \varepsilon_{ijc}, \tag{11}$$

where f_{1ijc} and f_{2ijc} are appropriate first-order derivatives.[4] Note that $\rho_{ijc} = f\left(b_{ic}, b_{jc}, \tau_c\right)$. Also, note that unmodelled DIF effects ν_{ic} enter estimated tetrachoric correlations in Eq. (11). The estimated transformed standard deviation $\widehat{\tau}_{c,\text{inv}}$ can be written as

[3] Note that $\omega_{ijc} = \omega_{jic}$.

[4] Note that it holds $f_{1ijc} = f_{2jic}$.

$$\widehat{\tau}_{c,\text{inv}} = \tau_c - \frac{\sum_{i=1}^{I} \omega_{ic} \nu_{ic}}{2\omega_c} + \frac{\sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \omega_{ijc} \varepsilon_{ijc}}{\omega_c}, \tag{12}$$

where $\omega_{ic} = \sum_{j \neq i} f_{1ij} \omega_{ijc}$ and $\omega_c = \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \omega_{ijc}$. As for the country means, the transformed standard deviation $\widehat{\tau}_{c,\text{inv}}$ is unbiased because DIF effects ν_{ic} and sampling errors ε_{ijc} have expected values of zero.

Under a separate scaling that relies on noninvariance, estimated country-specific item parameters $b_{ic} + e_{ic}$ are used, where e_{ic} denotes the sampling error for the item difficulty of item i. We can derive the estimated tetrachoric correlation under noninvariance as

$$\widehat{\rho}_{ijc,noninv} = f\left(b_{ic} - \nu_{ic}, b_{jc} - \nu_{jc}, \tau_c\right) + \varepsilon_{ijc} \simeq \tau_c + f_{1ijc} e_{ic} + f_{2ijc} e_{jc} + \varepsilon_{ijc} \tag{13}$$

Note that we again use a Taylor expansion for the analytical treatment. The transformed standard deviation $\widehat{\tau}_{c,\text{noninv}}$ can be expressed as

$$\widehat{\tau}_{c,\text{noninv}} = \tau_c + \frac{\sum_{i=1}^{I} \omega_{ic} e_{ic}}{2\omega_c} + \frac{\sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \omega_{ijc} \varepsilon_{ijc}}{\omega_c} \tag{14}$$

The only difference of the competitive estimates in Eqs. (12) and (14) lies in the second term. For comparing the variance of the two estimates, the different signs can be ignored. It is important that DIF effects influence the variance of the estimate under invariance (i.e., concurrent scaling) in Eq. (12) while sampling errors in item difficulties influence the estimate based on noninvariant item parameters in Eq. (14). Because the third term in Eq. (12) coincides with the third term in (14), the efficiency of the two estimates $\widehat{\tau}_{c,\text{inv}}$ and $\widehat{\tau}_{c,\text{noninv}}$ depends on which of the variances $\text{Var}(\nu_{ic})$ or $\text{Var}(e_{ic})$ is larger. Hence, in the presence of DIF effects in item difficulties, the standard deviation estimate based on separate scaling (i.e., a transformation of $\widehat{\tau}_{c,\text{noninv}}$) can have lower variance than the estimate based on concurrent scaling (i.e., a transformation of $\widehat{\tau}_{c,\text{inv}}$) that relies on invariant item parameters. Interestingly, DIF effects in item difficulties impact the variability of estimated standard deviations, which might be a surprising finding. Consequently, separate scaling can provide more efficient estimates for country standard deviations than concurrent scaling. This finding contradicts previous claims in the literature that concurrent scaling will result in more stable estimates (see von Davier et al., 2019).

Linking error of original and marginal trend estimates in the 1PL model Due to DIF, IPD, and DIF × IPD.

This section reviews the linking error in original and marginal trend estimates for the country means in the 1PL model. The detailed derivations can be found in Robitzsch and Lüdtke (2019).

The item response function in the 1PL model for item i in study t=1,2 for country c is given by

$$\text{logit}\, \text{P}(X_{itc} = 1|\theta) = \theta - b_{itc}\,, \tag{15}$$

where item difficulties b_{itc} are centered for each country c in study t. In this case, the country mean μ_{tc} and the country standard deviation σ_{tc} can be estimated. We now assume a variance component model for DIF effects of item difficulties (see Robitzsch & Lüdtke, 2019)

$$b_{itc} = b_i + \nu_{ic} + \nu_{it} + \nu_{itc}, \tag{16}$$

where the item effects ν_{ic}, ν_{it}, and ν_{itc} have zero means and variances $\sigma^2_{\text{DIF}} = \text{Var}(\nu_{ic})$, $\sigma^2_{\text{IPD}} = \text{Var}(\nu_{it})$, and $\sigma^2_{\text{DIF}\times\text{IPD}} = \text{Var}(\nu_{itc})$.

The linking error for the original trend $\widehat{\Delta\mu}_{\text{orig}}$ in country means is given by

$$Var\left(\widehat{\Delta\mu}_{\text{orig}}\right) = \frac{2}{I_0}\sigma^2_{\text{IPD}} + \frac{I_1 + I_2}{(I_0 + I_1)(I_0 + I_2)}\sigma^2_{\text{DIF}} + \frac{2I_0 + I_1 + I_2}{(I_0 + I_1)(I_0 + I_2)}\sigma^2_{\text{DIF}\times\text{IPD}} \tag{17}$$

In contrast, the linking error for the marginal trend $\widehat{\Delta\mu}_{\text{marg}}$ is given by

$$\text{Var}\left(\widehat{\Delta\mu}_{marg}\right) = \frac{2}{I_0}\sigma^2_{\text{IPD}} + \frac{2}{I_0}\sigma^2_{\text{DIF}\times\text{IPD}} \tag{18}$$

Robitzsch and Lüdtke (2019) have shown that the marginal trend estimate often has a lower variance than the original trend estimate. This finding was obtained by Eqs. (17) and (18) for designs with unique items (i.e., $I_1 > 0$ or $I_2 > 0$) because the cross-sectional DIF variance σ^2_{DIF} frequently turns out to be much larger in empirical applications than the variances σ^2_{IPD} and $\sigma^2_{\text{DIF}\times\text{IPD}}$. In an empirical application using the PISA reading trend between 2006 and 2009, Robitzsch and Lüdtke (2019) found that the empirical variance in marginal trend estimates in country means was smaller than those for original trend estimates. This finding is in coherence with the derived formulas for the linking error.

The analytical results for the 1PL model in Robitzsch and Lüdtke (2019) have been generalized to the 2PL model resulting in slightly more complex formulas (Robitzsch, 2023). In the present article, it is investigated how large efficiency gains in alternative linking approaches to the original trend estimate can be observed for the 1PL and the 2PL models. Moreover, it seems that the discussion about the superiority of original or marginal trends is confounded with the type of linking (indirect vs. direct), the set of items chosen for linking (all items vs. link items), and whether item parameters are assumed to be invariant or noninvariant across countries. These three aspects are disentangled in the following simulation study by including an extended set of trend estimation methods.

Simulation study

In our simulation study, we investigated the performance of different trend estimates for countries for the 1PL and the 2PL model. Analytical results were only obtained for the 1PL model, and whether these findings generalize to the more complex 2PL model that typically requires larger sample sizes to estimate item discriminations could be questioned.

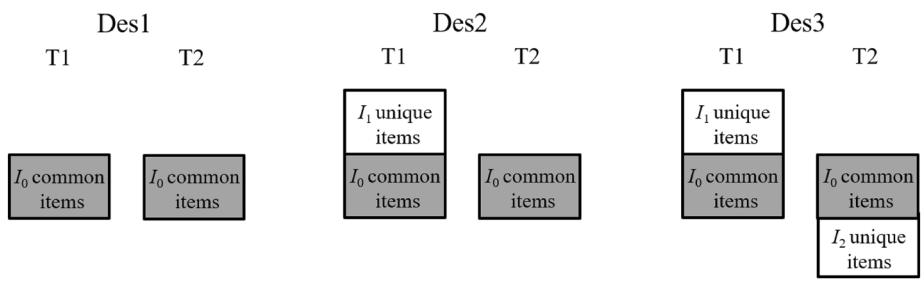

Fig. 2 Test designs Des1 (left panel), Des2 (middle panel), and Des3 (right panel) for time point 1 (T1) and time point 2 (T2) utilized in the simulation study

Method

The simulation study was designed to mimic test designs used in previous PISA (OECD, 2012, 2014) or PIRLS and TIMSS (Martin et al., 2017) assessments. The simulation design is a subdesign of the design used in the linking study of Robitzsch and Lüdtke (2019). Two international data sets referring to two measurement waves, each with the same 20 countries, were simulated. The primary parameters of interest were trend estimates for the mean and the standard deviation for each country (i.e., the difference between the country mean of the second and the first assessment). Moreover, cross-sectional estimates for country means and standard deviations were also reported.

To obtain realistic values for the data-generating model, we took the means and standard deviations for the reading achievement test of 20 selected OECD countries participating in PISA 2009 and 2012. We transformed them such that for the total population comprising all countries in this simulation study, the mean was set to 0, and the standard deviation was set to 1. The country means for the two assessments were similarly chosen (time 1: $M = 0.00$, $SD = 0.20$, $Min = -0.47$, $Max = 0.40$; time 2: $M = 0.05$; $SD = 0.20$; $Min = -0.55$; $Max = 0.28$). Moreover, the distribution of country standard deviations did not significantly differ between the two assessments (time 1: $M = 0.92$; $SD = 0.07$; $Min = 0.82$; $Max = 1.06$; time 2: $M = 0.93$; $SD = 0.08$; $Min = 0.78$; $Max = 1.09$). However, there were some country-specific variations in trend estimates for country means ($M = 0.05$; $SD = 0.11$; $Min = -0.17$; $Max = 0.27$) and country standard deviations ($M = 0.01$; $SD = 0.05$; $Min = -0.09$; $Max = 0.09$).

We implemented three different linking designs Des1, Des2, and Des3, in the simulation study, denoted as LA30, LB30a, and LC30 in Robitzsch and Lüdtke (2019), respectively. The three designs differ regarding the number of link items I_0 and the number of unique items I_1 and I_2 at time points 1 and 2. The design Des1 is a minor-minor design in PISA terminology that only contains link items ($I_0 = 30$, $I_1 = 0$, $I_2 = 0$). The design Des2 is a major-minor design that only has unique items at time 1 ($I_0 = 30$, $I_1 = 30$, $I_2 = 0$). The design Des3 contains unique items at both time points ($I_0 = 30$, $I_1 = 30$, $I_2 = 30$) and corresponds to a design that is implemented in PIRLS or TIMSS. The three test designs are displayed in Fig. 2. The number of administered items per student was fixed at 30 in all designs and at both time points. Because 60 items were administered in Des2 at time point 1 and Des3 at both time points, a

Table 1 Overview of different analysis models in the simulation study

Model	Link	Inv	Items
I1 (original)	Indirect	Yes	All
I2	Indirect	No	All
I3	Indirect	Yes	Link
I4	Indirect	No	Link
D2 (marginal)	Direct	No	All
D4	Direct	No	Link

Inv assumption of invariant item parameters across countries; *indirect* indirect linking; *direct* direct linking

balanced incomplete block design was used (Frey et al., 2009; see Robitzsch & Lüdtke, 2019, for the implementation) such that 30 out of 60 items appeared in each booklet.

The 1PL or the 2PL model was used for generating item responses. Item difficulties and item discriminations in the 2PL model were held constant across replications. However, DIF, IPD, and DIF × IDP were simulated, and the effects varied across replications within each condition. Furthermore, we assumed normal distributions for item effects that were homogeneous across countries (i.e., σ^2_{DIF}, σ^2_{IPD}, and $\sigma^2_{\text{DIF}\times\text{IPD}}$ did not differ across countries). The 2PL model was simulated with no DIF (NODIF), uniform DIF (UDIF), and nonuniform DIF (NUDIF). In the NODIF condition, no DIF, IPD, and DIF × IDP effects were simulated for item difficulties. For uniform DIF in the 1PL and the 2PL model, we used the variances $\sigma^2_{\text{DIF}} = 0.20$, $\sigma^2_{\text{IPD}} = 0.03$, and $\sigma^2_{\text{DIF}\times\text{IPD}} = 0.03$. Note that the variances were not made country-specific. Nonuniform DIF was simulated by assuming DIF effects for logarithmized item discriminations by choosing variances $\sigma^2_{\text{DIF}} = 0.05$, $\sigma^2_{\text{IPD}} = 0.00$, and $\sigma^2_{\text{DIF}\times\text{IPD}} = 0.00$ (i.e., there was no IPD and no DIF × IPD variance). The sizes of the DIF variances were chosen in alignment with previous studies (Monseur et al., 2008; Robitzsch & Lüdtke, 2019; Sachse et al., 2016), which reflect the extent of DIF that have been found in LSA studies. Note that additive DIF effects for logarithmized item discriminations correspond to multiplicative DIF effects for item discriminations.

The sample sizes N of each country and each time point were set to 500, 1000, or 2000. In contrast to typical large-scale assessments, we did not specify a clustered sampling design (i.e., students nested within schools). The main characteristic of clustered sampling designs is that they reduce the effective sample size. As we had already manipulated the sample size in our simulation study, it can be expected that the main findings would not change if clustered sampling designs were included.

We now describe the general principle of the linking approach. In the first step, item response data at the international level, containing all 20 countries, were calibrated separately for both time points. This model specified a single group 1PL or 2PL with invariant item parameters across countries. Country means and country standard deviations assuming invariant item parameters were obtained by country-wise scaling models with item parameters fixed to those obtained at the international level (i.e., fixed item parameter calibration; see, e.g., König et al., 2021). Moreover, we performed separate

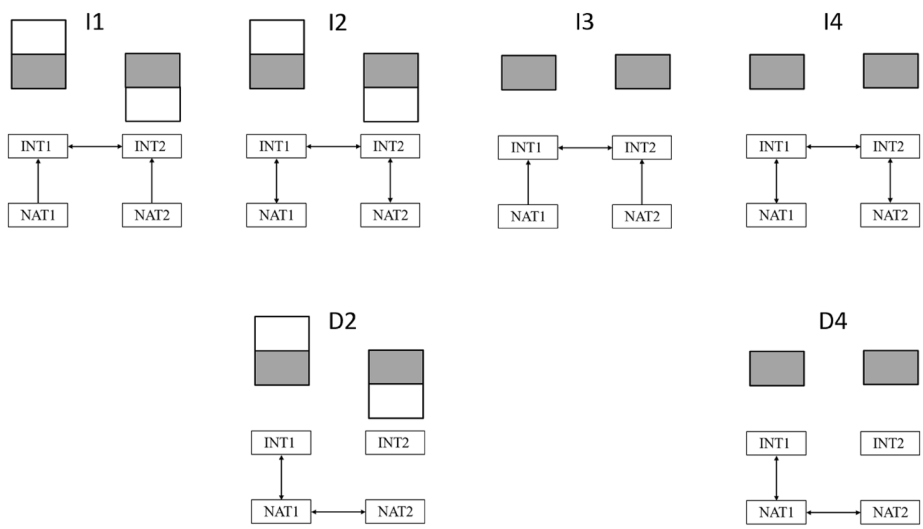

Fig. 3 Schematic representation of the six linking methods in the simulation study. *Double arrows indicate linking approaches, while single arrows correspond to a fixed item parameter scaling in which item parameters were fixed at the international metric.* The gray rectangle without a label corresponds to link items, while the white rectangle without a *label corresponds to unique items*

scaling models for each country at both time points that allowed for noninvariant item parameters. The country-specific distributions obtained from separate scaling were subsequently linked to the international level with Haberman linking (Haberman, 2009; Robitzsch, 2020), which simplifies to a log-mean-mean linking (Kolen & Brennan, 2014; Robitzsch, 2021b) for linking each of the countries at the international metric. Moreover, the link of both time points at the international level was also carried out using Haberman linking.[5]

To investigate the influence of item choice, we conducted linking based on all items or based only on the link items. For design Des1, the set of all items and link items coincided. However, different outcomes can be expected for designs Des2 and Des3 because there also exist unique items.

For analyzing the simulated item response datasets, we specified six different analysis models (see Table 1 and Fig. 3). These models differ concerning the method in which the longitudinal link was established. The first four methods I1, I2, I3, and I4, performed indirect linking methods to link the country performance to the international metric (i.e., using the pooled international dataset at the two time points for linking). In contrast, the last two methods D2 and D4 performed direct linking (i.e., using the country-specific datasets for longitudinal linking). Models I1 and I2 used all items, but I1 assumed invariant item parameters, and I2 relied on noninvariant item parameters. The analysis was repeated based on only using link items in models I3 and I4. The analysis method D2 performed direct linking and used all items, while D4 was restricted to using

[5] In Haberman linking, we excluded items at the level of countries and time points that had estimated item discriminations lower than 0.20. In earlier simulation studies, we found that items with small estimated item discriminations increased the variability in estimated linking constants in Haberman linking. This issue particularly occurs because Haberman linking utilizes logarithmized item discriminations. Obviously, negative item discriminations cannot be used at all.

only link items.[6] Note that we used the 1PL model as the analysis model when the data were generated by a 1PL and the 2PL model when the data were simulated with a 2PL.

The trend estimates in means and standard deviations of each country were obtained by calculating the difference between the country means and country standard deviations at time point 2 and time point 1. Moreover, the obtained abilities were linearly transformed to result in a mean of 0 and a standard deviation of 1 in the total population of all students at time 1.

More details about the linking designs, data-generating parameters for country means, country standard deviations, and used item parameters can be found in Additional file S1 at https://osf.io/n5zm6/?view_only=086ea651bbea49bb8b2aae44e3971db8. For all analyses, the R software (R Core Team, 2022) was employed. The R package TAM (Robitzsch et al., 2022) was used for estimating the IRT models, and the R package sirt (Robitzsch, 2022b) was used for Haberman linking. For all conditions in the simulation design, we conducted 1000 replications. In this simulation study, we consider the country mean and standard deviation as the parameter of interest. The bias was estimated by calculating the difference between the mean trend estimates from each design cell and the true parameter (i.e., the true trend estimate for a country). The overall accuracy of the parameter estimates was assessed with the root mean square error (RMSE), computed by calculating the square root of the mean square difference between the estimate and the true parameter. Note that the RMSE equals the standard deviation of an estimator for an unbiased parameter estimate (i.e., the efficiency of an estimator). We report a relative RMSE that is defined as the ratio of the RMSE of a method and the RMSE of the method I1 (i.e., the original trend), which serves as the reference method because it was used in operational practice in PISA studies 2000 to 2012. The results for the different criteria were aggregated across the 20 countries to reduce the information presented in the Results section. Replication material can be found in Additional file S1 at https://osf.io/n5zm6/?view_only=086ea651bbea49bb8b2aae44e3971db8.

Results

We now present the findings of our simulation study for trend estimates in country means and country standard deviations for the three different assessment designs. Across all estimation methods and simulation conditions, no biases existed on average. Hence, the RMSE only quantifies the variability (i.e., efficiency) of trend estimates.

In Table 2, the relative average root mean square error in the 1PL model for cross-sectional mean and standard deviation, as well as the trend in country means and standard deviations for each simulation condition, are displayed.

We first briefly discuss the cross-sectional estimates for the first time point. The method I1 relying on invariant item parameters was slightly more effective in terms of RMSE if there was no DIF (condition "NODIF"). However, cross-sectional country means, and standard deviations were more precisely estimated with method I2, which assumes noninvariant item parameters in the presence of DIF (condition "UDIF"). In

[6] One may wonder that no direct linking methods D1 and D3 were specified. These methods would assume invariant item parameters across countries. This would necessarily mean that a link utilizing the international metric were conducted. Hence, only the two methods D2 and D4 are considered reasonable direct linking approaches.

Table 2 Relative average root mean square error (RMSE) for cross-sectional estimates and trend estimates in country means and country standard deviations based on the 1PL model for assessment design Des1 ($I_0 = 30$, $I_1 = 0$, $I_2 = 0$), Design Des2 ($I_0 = 30$, $I_1 = 30$, $I_2 = 0$), and Design Des3 ($I_0 = 30$, $I_1 = 30$, $I_2 = 30$)

Design	Model	Link	Inv	Items	1PL NODIF			1PL UDIF		
					N			*N*		
					500	**1000**	**2500**	**500**	**1000**	**2500**
	Cross-sectional mean									
Des1	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	100.2	100.3	100.1	**98.2**	**98.4**	**97.3**
	Cross-sectional standard deviation									
Des1	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	102.6	102.7	102.1	**80.8**	**69.9**	**52.4**
	Trend in mean									
Des1	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	100.1	100.2	100.1	**99.2**	**98.7**	**98.2**
	D2	Direct	No	All	100.1	100.2	100.1	**99.2**	**98.7**	**98.2**
Des2	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	100.3	100.3	100.3	**99.2**	**98.9**	**98.6**
	I3	Indirect	Yes	Link	103.8	104.1	104.0	**89.8**	**84.9**	**81.4**
	I4	Indirect	No	Link	104.5	104.8	104.6	**89.6**	**84.2**	**80.1**
	D2	Direct	No	All	104.3	104.6	104.5	**89.6**	**84.4**	**80.2**
	D4	Direct	No	Link	104.5	104.8	104.6	**89.6**	**84.2**	**80.1**
Des3	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	100.5	100.3	100.3	**99.0**	**98.5**	**98.2**
	I3	Indirect	Yes	Link	107.4	107.0	107.7	**93.3**	**90.6**	**84.5**
	I4	Indirect	No	Link	108.7	107.8	108.6	**93.2**	**90.0**	**83.3**
	D2	Direct	No	All	108.3	107.5	108.4	**93.1**	**90.0**	**83.4**
	D4	Direct	No	Link	108.7	107.8	108.6	**93.2**	**90.0**	**83.3**
	Trend in standard deviation									
Des1	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	102.1	102.5	102.3	**97.8**	**93.3**	**83.9**
	D2	Direct	No	All	102.1	102.5	102.3	**97.8**	**93.3**	**83.9**
Des2	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	102.4	101.9	102.2	**92.5**	**84.8**	**72.2**
	I3	Indirect	Yes	Link	111.5	110.9	111.1	107.1	100.9	**96.8**
	I4	Indirect	No	Link	114.9	114.3	114.6	103.8	**93.6**	**80.8**
	D2	Direct	No	All	102.4	101.9	102.2	**92.5**	**84.8**	**72.2**
	D4	Direct	No	Link	114.9	114.3	114.6	103.8	**93.6**	**80.8**
Des3	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	102.5	102.3	102.3	**93.7**	**87.8**	**73.7**
	I3	Indirect	Yes	Link	119.3	119.8	121.1	117.1	113.1	103.1
	I4	Indirect	No	Link	124.5	124.3	126.1	115.4	107.6	**90.5**
	D2	Direct	No	All	102.5	102.3	102.3	**93.7**	**87.8**	**73.7**
	D4	Direct	No	Link	124.5	124.3	126.1	115.4	107.6	**90.5**

Inv assumption of invariant item parameters across countries; *indirect* indirect linking; *direct* direct linking; *NODIF* no differential item functioning; *UDIF* uniform differential item functioning; *NUDIF* nonuniform differential item functioning. Method I1 (original trend) was used as the reference for computing the relative RMSE. RMSE entries smaller than 100 are printed in bold

particular, there were efficiency gains of about 50% for cross-sectional standard deviations for the sample size $N=2500$.

Next, we discuss the findings for trend estimates for the country means in design Des1 that only contained link items (i.e., the sets "all items" and "link items" coincided). Because there are no unique items, only three of the six alternative approaches are displayed. It can be seen that there were minor efficiency losses in the absence of DIF but slightly more gains in the presence of DIF when using methods I2 and D2 that relied on noninvariant item parameters. Notably, methods I2 and D2 did not differ from each other.

The designs Des2 and Des3 also included unique items. It turned out that efficiency gains for the trend in the country mean for some methods compared to the reference method I1 can be at most 10% in the absence of DIF. However, efficiency gains in the presence of DIF turned out to be even more significant and were at most 20% (Des2, for method D4 and $N=2500$). Notably, methods that used only link items (I3, I4) or performed direct linking (D2, D4) were more efficient in the occurrence of DIF. Interestingly, method I4, which used only link items and noninvariant item parameters, performed similarly regarding the precision of trend in means to method I3, which assumes invariant item parameters.

Now, we turn to the assessment of trend estimates in country standard deviations for the 1PL model. In general, efficiency gains from alternative linking approaches to method I1 were more pronounced for trend estimates in standard deviations than in country means. As for the cross-sectional standard deviation, substantial efficiency gains from alternative linking approaches were observed (at most, 16.1% for Des1, 27.8% for Des 2, and 26.3% for Des3). As a general conclusion, it can be seen that the marginal trend estimate D2 can be recommended for obtaining precise trend estimates in the presence of DIF.

To summarize the performance of the six different analysis methods across conditions, we ranked the methods according to their RMSE performance. We assigned an average rank in the presence of ties. Two relative RMSE values in Table 2 were defined as equal if the values were equal after rounding to the first decimal place. In our summary, we included country mean and standard deviation trend estimates for designs Des2 and Des3. The results for country means and standard deviation were equally weighted in the rank statistic.

In the NODIF condition, the average ranks were: I1: 1.00, I2: 2.25, I3: 3.50, I4: 5.50, D2: 3.25, and D4: 5.50. Thus, in the absence of DIF linking at the international metric was most efficient if it was based on all items (I1 and I2). Note that the assumptions of method I1 were in complete alignment with the data-generating model in the case of no DIF. However, in the presence of DIF, the average ranks were I1: 5.38, I2: 3.25, I3: 4.88, I4: 2.69, D2: 2.13, and D4: 2.69. In this case, method D2 was the frontrunner, followed by methods I4 and D4. Note that method I4 relies on noninvariant item parameters and uses only link items.

In Table 3, the RMSE for the 2PL model under no DIF ("NODIF"), uniform DIF ("UDIF"), and nonuniform DIF ("NUDIF") is presented. Notably, there were larger efficiency losses for cross-sectional means in the 2PL model for models that relies on noninvariant parameters if DIF was absent. Efficiency gains for large samples ($N=1000$

Table 3 Relative average root mean square error (RMSE) for cross-sectional estimates and trend estimates in country means and country standard deviations based on the 2PL Model for assessment design Des1 ($I_0 = 30$, $I_1 = 0$, $I_2 = 0$), Design Des2 ($I_0 = 30$, $I_1 = 30$, $I_2 = 0$), and Design Des3 ($I_0 = 30$, $I_1 = 30$, $I_2 = 30$)

Design	Model	Link	Inv	Items	2PL NODIF			2PL UDIF			2PL NUDIF		
						N		*N*			*N*		
					500	1000	2500	500	1000	2500	500	1000	2500
	Cross-sectional mean												
Des1	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	122.8	119.4	117.9	115.7	**95.6**	**91.9**	114.3	**95.6**	**89.7**
	Cross-sectional standard deviation												
Des1	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	108.8	107.0	106.1	120.5	**62.6**	**45.0**	110.0	**89.7**	**85.5**
	Trend in mean												
Des1	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	121.6	118.0	116.6	128.6	106.3	100.5	132.3	109.8	**99.5**
	D2	Direct	No	All	121.3	118.0	116.6	115.4	106.3	100.5	115.9	106.2	**99.5**
Des2	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	182.4	123.1	118.8	197.5	112.2	**98.5**	213.5	115.9	**96.9**
	I3	Indirect	Yes	Link	104.2	103.5	103.7	**88.9**	**84.7**	**81.1**	**88.5**	**84.8**	**79.4**
	I4	Indirect	No	Link	234.8	137.6	132.4	222.2	109.5	**85.1**	244.1	127.0	**83.4**
	D2	Direct	No	All	139.3	132.1	128.7	115.1	**95.7**	**84.0**	112.5	**95.4**	**80.8**
	D4	Direct	No	Link	145.9	136.8	132.4	118.6	**98.2**	**85.1**	115.7	**98.0**	**81.7**
Des3	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	191.7	125.1	119.7	215.9	114.6	**98.3**	215.0	125.0	**96.0**
	I3	Indirect	Yes	Link	107.2	107.1	107.6	**93.1**	**88.1**	**83.5**	**91.7**	**87.7**	**82.7**
	I4	Indirect	No	Link	230.2	152.0	145.7	234.5	114.5	**90.8**	259.9	132.2	**87.9**
	D2	Direct	No	All	154.1	143.0	139.1	129.0	105.2	**88.8**	126.5	105.5	**85.8**
	D4	Direct	No	Link	165.4	151.1	145.7	136.8	110.1	**90.8**	132.8	110.1	**87.8**

Table 3 (continued)

Design	Model	Link	Inv	Items	2PL NODIF			2PL UDIF			2PL NUDIF		
						N		*N*			*N*		
					500	1000	2500	500	1000	2500	500	1000	2500
	Trend in standard deviation												
Des1	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	107.7	106.4	105.9	100.2	**92.4**	**80.1**	**99.8**	**92.6**	**76.7**
	D2	Direct	No	All	107.6	106.4	105.9	100.2	**92.4**	**80.1**	**99.4**	**92.4**	**76.7**
Des2	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	109.1	106.8	106.0	**93.8**	**83.6**	**65.3**	**98.5**	**92.4**	**83.2**
	I3	Indirect	Yes	Link	109.9	109.7	110.0	103.8	**98.7**	**91.8**	100.3	**94.1**	**85.7**
	I4	Indirect	No	Link	124.1	120.6	119.1	107.0	**94.4**	**72.8**	102.3	**88.7**	**67.0**
	D2	Direct	No	All	120.5	118.6	117.3	104.7	**92.6**	**71.9**	100.1	**86.7**	**65.9**
	D4	Direct	No	Link	122.4	120.3	119.1	106.2	**94.0**	**72.8**	100.9	**87.9**	**66.9**
Des3	I1	Indirect	Yes	All	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	I2	Indirect	No	All	109.5	107.6	106.2	**94.8**	**85.0**	**68.4**	100.0	**95.1**	**86.9**
	I3	Indirect	Yes	Link	118.7	119.9	119.3	111.9	106.0	**98.6**	108.6	102.2	**92.4**
	I4	Indirect	No	Link	136.1	133.5	131.4	119.7	105.2	**83.8**	114.8	101.0	**76.7**
	D2	Direct	No	All	131.3	129.8	127.7	116.5	102.1	**81.7**	111.3	**97.5**	**74.4**
	D4	Direct	No	Link	134.5	133.2	131.4	119.6	104.9	**83.8**	114.1	100.1	**76.6**

Inv assumption of invariant item parameters across countries; *indirect* indirect linking; *direct* direct linking; *NODIF* no differential item functioning; *UDIF* uniform differential item functioning; *NUDIF* nonuniform differential item functioning. Method I1 (original trend) was used as the reference for computing the relative RMSE. RMSE entries smaller than 100 are printed in bold

and $N = 2500$) in UDIF (at most 8.1%) and NUDIF (at most 10.3%) were smaller. In contrast, they were much more pronounced for standard deviations (UDIF: 55.0%, NUDIF: 14.0%). The efficiency gains for the cross-sectional standard deviation were generally smaller in the nonuniform DIF than in the uniform DIF condition.

Regarding the trend in means, efficiency losses in the absence of DIF were smallest when the international metric was used for linking, and only link items were utilized (method I3). However, efficiency gains for method I3 were apparent for designs Des2 and Des3 in the presence of DIF (Des2 UDIF: − 18.9%, Des2 NUDIF − 20.6%, Des3 UDIF − 16.5%, Des3 NUDIF − 17.3%, for a sample of $N = 2500$). Notably, methods I3, I4, D2, and D4 that relied on only using link items or utilized direct linking can also substantially reduce the variability in mean trend estimates for large sample sizes $N = 2500$. However, they can be ineffective for smaller sample sizes. Notably, the performance of the 1PL and the 2PL model differs for lower sample sizes $N = 500$ and $N = 1000$. Method D2 was slightly more efficient than method I2 for design Des1 for uniform or nonuniform DIF. The direct linking methods D2 and D4 were generally beneficial in reducing the variability for trend standard deviations. As an alternative, the indirect linking method I4 metric might be used that only relied on link items and assumed noninvariant item parameters.

As a general picture, we can state that the potential efficiency gains of alternative linking methods strongly depended on the sample size. As for the 1PL model, we also ranked the performance of the competitive methods for trend means and standard deviations for designs Des2 and Des3 in the conditions NODIF, UDIF, and NUDIF. In the absence of DIF (NODIF), we obtained the average ranks: I1: 1.00, I2: 2.83, I3: 2.50, I4: 5.83, D2: 3.83, and D4: 5.00. Again, it can be seen that in the condition with no DIF in the data-generating model the method I1 performed best, followed by methods I3 and I2. In the uniform DIF (UDIF) condition, the average ranks were I1: 3.83, I2: 3.17, I3: 2.75, I4: 4.75, D2: 2.67, and D4: 3.83. Overall, the marginal trend estimate D2 performed best, closely followed by I3 (only using link items in indirect linking but assuming invariant item parameters). In the nonuniform DIF (NUDIF) condition, we obtained the average ranks: I1: 3.88, I2: 3.79, I3: 2.83, I4: 4.83, D2: 2.25, and D4: 3.42. As for uniform DIF, D2 performed best, followed by I3.

To sum up, depending on the test design, the amount of DIF, and sample sizes, we found that there can be substantial efficiency gains in terms of RMSE when restricting the set of items used for linking to link items (I3, I4) or relying on direct linking (D2, D4) compared to the original trend (I1). The findings slightly differed between the 1PL and the 2PL models. Generally, smaller sample sizes were required in the 1PL model to realize efficiency gains with alternative trend estimation methods.

Empirical example: trend estimates for trend from PISA 2006 to PISA 2009

In this empirical example, we investigated the PISA trend in reading, mathematics, and science from PISA 2006 to PISA 2009. The different trend estimates for trend estimates in country means and standard deviations utilized in the simulation study are compared. Moreover, the assessment of standard errors and linking errors is discussed.

Method

We chose 30 OECD countries that participated in PISA 2006 (OECD, 2009) and PISA 2009 (OECD, 2009) in our trend analysis (see Oliveri & von Davier, 2017 for a similar reanalysis). In PISA 2006, science was the major domain, while reading and mathematics were minor domains. In PISA 2009, reading was the major domain, while mathematics and science were minor domains. Hence, a nonnegligible number of unique items appeared in reading and science, while the mathematics test mostly possessed link items in our trend analysis.

Although the same countries were involved in our linking study, different schools within a country participated in the two time points. Consequently, the two different student samples can be considered independent. However, due to the stratified clustered sampling of students, students cannot be considered as independent realizations in the sample (see, e.g., Meinck, 2020). At each time point, test booklets contained two or three cognitive domains. This means that there was no item response data in some of the three domains for the majority of students. Missing item responses of non-administered items in one of the domains can be compensated by relying on plausible value imputation from a three-dimensional IRT model (von Davier & Sinharay, 2014). However, in our analysis, we relied on the unidimensional 2PL models and restricted the analysis to the students who had been administered items in the respective domain. To ensure that representativity is not impacted, we readjusted student weights for missing students using a non-response adjustment factor within a school. Hence, the sum of student weights within each school in our sample equaled the sum from the original (i.e., full) sample. For reading, the median sample size at the country level was 2683.5 in PISA 2006 ($Min = 2010$, $Max = 16215$). For PISA 2009, the median sample size was 5700.5 ($Min = 3628$, $Max = 38206$). The median of the average sample size per item at the country level was 1494.5 in PISA 2006 ($Min = 1142.3$, $Max = 8895.0$) and 1735.8 in PISA 2009 ($Min = 1114.4$, $Max = 11743.6$). For mathematics, the median sample sizes at the country level were 3797 ($Min = 2888$, $Max = 23788$) for PISA 2006 and 3972.5 ($Min = 2510$, $Max = 26406$) for PISA 2009. For science, the median sample size was 4933.5 ($Min = 3778$, $Max = 30872$) in PISA 2006 and 3803 ($Min = 2501$, $Max = 26374$) in PISA 2009.

In the official PISA datasets, some item responses were polytomous. For simplicity, all polytomous were dichotomously rescored, where only the highest category was rescored as correct. In the reading domain, 28 items and 101 items were administered in PISA 2006 and PISA 2009, respectively, from which 26 items were link items administered in both assessments. In the mathematics domain, 48 items and 35 items were administered in PISA 2006 and PISA 2009, respectively, from which 32 items were link items. Finally, 103 items and 53 items were administered in the science domain, respectively, from which 49 items served as link items.

Omitted and not-reached item responses were scored as incorrect in all scaling models and in the calibration samples to obtain item parameters at the international metric (see Robitzsch, 2021a, for a discussion).

We followed the same analysis strategy as in the simulation study. We applied the 1PL model with the six different linking methods I1, I2, I3, I4, D2, and D4. The obtained country means and standard deviations were subsequently linearly transformed such

that the total population comprising all students at the first time point[7] had a population mean of 500 and a standard deviation of 100. For assessing standard errors, plausible values were drawn, and repeated replicate weights were utilized to compute standard errors at the country level for both time points (Kolenikov, 2010; OECD, 2012). The standard error SE for a trend estimate in country means or country standard deviations was computed as

$$\mathrm{SE} = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}, \tag{19}$$

where SE_1 and SE_2 denote the standard errors in PISA 2006 and PISA 2009, respectively.

The linking to the international metric was conducted by using calibration samples for PISA 2006 and PISA 2009. For the three cognitive domains and the two time points (i.e., resulting in six calibration samples), 1000 students per country were randomly sampled. Hence, all calibration samples contained 30000 students.

The assessment of linking errors was carried out using a jackknife of testlets (Monseur & Berezner, 2007; Robitzsch & Lüdtke, 2019). Jackknife was applied at the testlet instead of the item level because previous studies showed that items within a testlet had similar DIF effects to some extent (Monseur et al., 2008). Note that jackknife was defined for testlets that referred to link items and unique items. For a trend estimate of interest γ, we computed trend estimates $\gamma_{(-t)}$ for all testlets $t = 1, \ldots, T$ in which testlet t was omitted from linking. The linking error LE was determined by

$$\mathrm{LE} = \sqrt{\sum_{t=1}^{T} \left(\gamma_{(-t)} - \gamma\right)^2}, \tag{20}$$

Note that this linking error quantifies DIF, IPD, and DIF $\times$ IPD effects.

However, we also computed the linking error $\mathrm{LE}_{\mathrm{IPD}}$ for the linking at the international metric in order to align the implemented approaches with the official PISA methodology (OECD, 2012). This has the potential disadvantage that this linking error quantifies only IPD effects, and it has been shown that using $\mathrm{LE}_{\mathrm{IPD}}$ could lead to severe underestimations of linking errors (compare Eqs. (17) and (18); Robitzsch & Lüdtke, 2019). Nevertheless, our primary goal for this empirical example was to assess differences between the linking methods and not the adequate assessment of linking errors. The total error for trend estimates was computed as

$$\mathrm{TE}_{\mathrm{IPD}} = \sqrt{\mathrm{SE}^2 + \mathrm{LE}_{\mathrm{IPD}}^2}\,. \tag{21}$$

To assess the statistical significance between two different linking methods, one can also compute the standard error by balanced repeated sampling (Robitzsch & Lüdtke, 2022a). Note that one cannot simply use the standard errors from the two methods because applying two different methods pertains to the same sample and results in highly dependent parameter estimates.

[7] In the computation of the total population, all countries equally contributed to the population mean and standard deviation. This corresponds to an analysis using house weights for all countries.

Table 4 Trend estimates in country means and their linking errors for the PISA Trend from 2006 to 2009 in reading

cnt	Estimate						Linking Error LE					
	I1	I2	I3	I4	D2	D4	I1	I2	I3	I4	D2	D4
AUS	4.8	3.3	3.4	2.0	**8.6**	2.0	8.1	5.6	4.6	4.6	3.7	4.6
AUT	−15.9*	−15.5*	**−9.2**	−10.9	**−6.0**	**−8.5**	7.7	9.6	6.9	8.9	6.2	7.8
BEL	7.0	**11.7**	**1.4**	4.0	6.0	4.0	7.7	6.4	6.4	6.2	5.7	6.2
CAN	2.8	4.4	**−2.8**	**−3.8**	−0.6	**−3.8**	11.5	5.8	8.1	6.7	3.0	6.7
CHE	6.5	11.4	5.0	2.8	12.2	2.8	11.1	11.4	4.5	6.2	5.7	6.2
CHL	15.5*	**6.9**	**−7.7**	**−13.7**	**5.9**	**−6.3**	18.9	20.2	10.9	22.7	11.9	19.1
CZE	3.3	4.2	3.7	6.8	**10.0**	6.8	12.5	13.2	5.5	5.3	4.0	5.3
DEU	4.6	**15.9***	**11.4**	10.5	11.5	10.5	6.3	17.3	8.5	5.7	7.9	5.7
DNK	0.5	−1.5	1.4	−5.4	0.1	−5.7	7.8	14.5	10.4	9.3	6.7	10.0
ESP	24.7*	23.7*	**6.5**	**2.4**	**9.4**	**1.7**	10.8	9.0	9.0	10.5	11.0	10.2
EST	10.2	11.6	**4.1**	**1.2**	7.0	**1.2**	17.7	14.5	12.2	11.2	9.5	11.2
FIN	−9.3*	**−3.3**	−6.2	−4.6	−6.3	−10.6	10.4	16.8	14.8	26.8	5.2	10.5
FRA	7.7	9.5	**−0.5**	**−3.9**	3.4	**−3.9**	9.5	8.1	14.1	4.5	5.9	4.5
GBR	5.1	6.9	**2.2**	6.0	**12.8**	6.0	10.9	7.7	5.7	4.1	8.0	4.1
GRC	33.8*	**22.0**	**11.1**	**12.3**	**19.0**	**12.3**	7.2	8.6	10.8	9.5	8.8	9.5
HUN	17.7*	16.1	**9.6**	**−0.5**	9.9	**4.9**	9.5	10.4	3.5	12.8	8.2	8.0
IRL	−12.6	−16.0	**−18.2**	**−21.7**	−15.2	**−21.7**	14.2	14.0	8.3	11.9	6.0	11.9
ISL	16.0*	17.5*	**4.1**	**2.3**	12.4	**2.3**	9.7	11.5	4.2	7.3	7.3	7.3
ISR	38.0*	**33.8***	**20.2***	**16.4**	**29.5***	**16.4***	8.6	8.4	10.8	7.1	7.5	7.1
ITA	24.1*	21.4*	**12.9***	**12.6***	21.0*	**12.6***	9.5	9.5	9.4	11.9	6.0	11.9
JPN	22.0*	**14.1**	**15.1**	**10.2**	**14.5**	**8.8**	15.5	11.7	8.4	9.0	4.5	9.4
KOR	−12.6*	−15.3*	−10.8	−11.5	**−1.4**	**−4.5**	13.5	8.1	11.0	9.8	6.3	5.7
LUX	2.1	7.4	**−7.6**	−7.2	0.6	−5.3	10.6	7.6	9.2	14.9	12.1	15.1
MEX	20.9*	24.2*	**3.7**	**1.3**	15.4*	**1.3**	12.5	10.7	10.6	8.2	13.2	8.2
NLD	3.1	7.0	0.7	2.0	2.6	2.0	22.3	20.4	4.0	5.9	7.5	5.9
NOR	21.2*	23.0*	20.8*	16.8	27.6*	24.6*	10.8	15.2	9.0	14.1	8.3	9.0
POL	3.6	**−1.5**	3.8	−2.8	1.8	−2.8	13.4	13.0	11.9	11.9	8.5	11.9
PRT	21.7*	19.9	**12.0**	**6.6**	18.4	**6.6**	10.7	25.9	9.3	12.6	6.5	12.6
SWE	−5.1	**5.3**	−7.0	1.7	**9.9**	1.7	10.2	15.3	7.6	6.6	11.1	6.6
TUR	33.3*	26.0*	**16.4***	**20.4**	27.8*	**20.4**	11.7	24.7	8.7	17.0	18.7	17.0

cnt country

* Statistically significant trend estimates ($p < .05$) with used standard error TE_{IPD} Estimates in I2, I3, I4, D2 and D4 that significantly differ from I1 are printed in bold

The replication code for the empirical application can be found in Additional file S2 at https://osf.io/n5zm6. As the original data can only be downloaded by individual researchers from the OECD websites and we are not allowed to make our processed PISA data publicly available, we created synthetic datasets (see Grund et al., 2022) that are very similar to the original datasets. The generation of the synthetic datasets is described in Appendix A. The average differences in median item discriminations and median item difficulties, as well as median absolute deviations in estimated item parameters between the original and synthetic datasets, were about 0.01 for item difficulties and 0.02 for item discriminations. Hence, perfect replicability of the original results cannot be expected, but findings for synthetic datasets would be sufficiently close.

Results

Trend in country means for reading.

Table 4 contains trend estimates in the country means for reading from PISA 2006 to PISA 2009. Reading was a minor domain in PISA 2006 and a major domain in PISA 2009. At the country level, the median range between the six different linking methods was 11.4 ($Min = 6.3$, $Max = 29.2$), which can be considered a surprisingly large variability. For example, the trend estimates in means for Austria (AUT) ranged between − 15.9 and − 6.0. For this country, the trend estimates for I1 and I2 were negative and statistically significant according to the total error defined in Eq. (21). In contrast, the trend estimates for the other four methods did not reach the significance level of 0.05. Also note that I3, D2, and D4 significantly differed from I1. Furthermore, a very large range was obtained for Chile (CHL), with trend estimates between − 13.7 and 15.5. An interesting pattern was observed for Spain (ESP), which has a large positive and statistically significant trend of 24.7 and 23.7 for methods I1 and I2. However, the trend estimates of the other four methods I3, I4, D2, and D4 were substantially smaller (ranging between 1.7 and 9.4). The latter methods only involved link items (I3 and I4) or performed direct linking (D2 and D4).

The median change in trend estimates at the country level was I1: 6.8, I2: 10.5, I3: 3.7, I4: 2.0, D2: 9.7, and D4: 2.0. Overall, there was a positive average trend for reading between PISA 2006 and PISA 2009. However, substantially smaller average trend estimates were obtained for methods I3, I4, and D4 (3.7, 2.0, and 2.0) that only used link items for trend estimation. We also computed the median in absolute values of trend estimated to get further insights into the variability across countries. Again, the median was larger for the methods that relied on all items (I1: 11.4, I2: 12.9, D2: 9.9). At the same time, it was smaller for the methods that were only based on the link items (I3: 6.8, I4: 5.7, D4: 5.5). Interestingly and in line with previous findings in the literature (Gebhardt & Adams, 2007; Robitzsch & Lüdtke, 2019), the marginal trend estimate D2 showed slightly lower variability compared to the original trend estimate I1.

Furthermore, we now compare the estimated linking errors across the six different methods. The median linking errors across countries were I1: 10.8, I2: 11.5, I3: 8.9, I4: 9.2, D2: 7.4, and D4: 8.1. In line with our findings for the empirical variability of the country mean trend estimates, we found that the trend estimates using indirect linking that was based on all items (I1 and I2) resulted in larger linking errors than indirect linking trend estimates that were based on link items (I3 and I4) and the direct linking trend estimates (D2 and D4).

The estimated linking error LE_{IPD} for the international link between PISA 2006 and PISA 2009 was determined as 3.1. Notably, this linking error turned out to be substantially smaller than the linking errors that involved all items and were assessed by recomputing the linking based on jackknife. We also assessed the proportion of statistically significant trend estimates based on the total error TE_{IPD} for the trend estimates in the country means for all 30 countries, which were I1: 50%, I2: 33%, I3: 13%, I4: 3%, D2: 17%, and D4: 10%. It can be seen that fewer trend estimates became statistically significant at the country level if they were only based on link items or if direct linking was employed.

Table 5 Trend estimates in country standard deviations and their linking errors for the PISA Trend from 2006 to 2009 in reading

cnt	Estimate						Linking error					
	I1	I2	I3	I4	D2	D4	I1	I2	I3	I4	D2	D4
AUS	−2.1	**−4.7**	**11.8***	**9.5***	−5.0	**9.5***	3.8	3.7	6.6	4.1	4.8	4.1
AUT	−9.2	**−11.7**	**11.2**	**7.1**	**−4.3**	**6.3**	7.7	6.0	7.5	7.4	5.9	7.4
BEL	−9.1*	−8.6*	**−1.1**	**−0.4**	−12.3*	**−0.4**	6.3	6.0	4.2	4.0	3.4	4.0
CAN	−7.1	−7.6	**2.4**	**2.0**	**−15.5***	**2.0**	9.0	6.8	2.7	1.7	9.3	1.7
CHE	−6.9	**−9.9***	**8.5**	**6.8**	−5.5	**6.8**	5.9	6.4	10.4	9.1	6.5	9.1
CHL	−14.6*	**−24.1***	**−2.3**	**2.2**	−16.6*	**1.9**	9.6	6.3	4.9	5.5	19.2	6.0
CZE	−26.5*	−25.8*	**−6.7**	**−7.8**	**−21.2***	**−7.8**	10.7	9.6	5.9	2.9	10.5	2.9
DEU	−20.1*	−17.4*	**−7.4**	**−6.8**	−19.0*	**−6.8**	7.1	5.9	6.7	4.8	5.8	4.8
DNK	−11.3*	−9.7*	**7.5**	**5.6**	−8.4	**5.0**	6.1	5.9	3.5	3.5	7.2	3.6
ESP	0.5	**−11.6***	**12.5***	**5.6**	**−14.3***	**5.8**	9.2	5.2	10.0	5.4	8.3	5.0
EST	−7.0*	**−9.8***	**6.3**	**3.9**	−4.8	**3.9**	8.7	8.7	2.7	2.8	5.8	2.8
FIN	−3.9	**−7.4***	**10.8***	**7.6**	−4.9	**6.7**	12.6	6.6	7.5	4.0	5.4	3.2
FRA	1.3	**−4.2**	**9.8**	**11.8**	**−6.6**	**11.8**	9.4	5.6	9.3	12.3	7.8	12.3
GBR	−10.3*	−11.6*	**1.6**	**1.2**	−13.0*	**1.2**	4.6	5.6	6.4	7.3	6.1	7.3
GRC	−5.9	**−11.5***	**13.0***	**8.7**	**−24.4***	**8.7**	4.5	4.7	14.5	14.0	11.6	14.0
HUN	−7.4	−5.5	**7.6**	**5.7**	−5.3	**5.1**	6.6	5.9	3.4	4.7	6.3	3.0
IRL	−6.2	−6.9	**9.7***	**7.9**	−9.0*	**7.9**	6.8	8.1	6.4	4.4	7.1	4.4
ISL	−0.9	**−4.6**	**11.4***	**10.1**	−6.2	**10.1**	9.5	10.5	5.1	10.3	15.7	10.3
ISR	−5.4	−8.4	**11.3***	**14.6**	**−16.9***	**14.6***	12.3	10.1	3.8	4.0	10.3	4.0
ITA	−7.7	**−9.6***	**6.1**	**6.4**	**−16.9***	**6.4**	5.3	8.2	6.1	6.5	8.1	6.5
JPN	−7.8	−6.2	**11.4***	**8.1**	−5.0	**7.6**	3.6	7.5	10.9	5.5	4.1	5.8
KOR	−19.8*	**−11.7***	**−5.8**	**−2.3**	**−9.1**	**−1.1**	11.8	13.8	9.1	7.3	11.4	7.1
LUX	−4.1	**−7.1***	**12.5***	**9.6**	−7.9	**9.6**	7.1	8.2	4.3	4.6	9.5	4.7
MEX	−7.0*	**−10.0***	**10.1***	**8.9***	**−12.5***	**8.9***	10.6	6.9	5.4	7.6	12.3	7.6
NLD	−13.7*	**−10.6**	**−4.6**	**−3.3**	−9.8*	**−3.3**	9.0	8.6	4.5	5.1	9.9	5.1
NOR	−16.2*	−17.9*	**4.0**	**−0.8**	−18.3*	**0.3**	4.6	5.4	5.4	5.7	10.3	5.2
POL	−17.1*	−16.3*	**0.9**	**2.2**	−14.5*	**2.2**	9.3	8.4	5.3	6.1	7.9	6.1
PRT	−12.9*	**−20.6***	**3.9**	**−0.6**	−16.3*	**−0.6**	7.5	5.0	12.9	10.9	6.8	10.9
SWE	−7.9*	−5.1	**14.8***	**14.7***	−6.5	**14.7***	7.1	8.0	10.2	12.9	12.4	12.9
TUR	−6.4	−7.9	**1.6**	**0.4**	**−15.1**	**0.4**	10.8	7.0	8.7	6.1	8.1	6.1

cnt country

* Statistically significant trend estimates ($p < .05$) with used standard error $mathrmTE_{\mathrm{IPD}}$ Estimates in I2, I3, I4, D2 and D4 that significantly differ from I1 are printed in bold

Trend in country standard deviations for reading

Table 5 contains the trend for reading in country standard deviations. We observed substantial ranges in trends in standard deviations across models ($Med = 18.8$, $Min = 10.4$, $Max = 37.4$). The median of the trend estimates in standard deviations for the different models were I1: −7.6, I2: −9.8, I3: 7.6, I4: 5.7, D2: −11.1, D4: 5.5. Interestingly, a negative trend was observed for methods that relied on all items (I1, I2, D2). At the same time, there was a positive trend for models based on link items (I3, I4, D4). The variability of trend estimates in standard deviations was again assessed by computing the median of absolute values in trend estimates. The variability tended to be smaller for the methods based on link items (I3: 7.6, I4: 6.6, D4: 6.4) compared to trends based on all items (I1: 7.6, I2: 9.8, D2: 11.1).

Table 6 Trend estimates in country means and their linking errors for the PISA Trend from 2006 to 2009 in mathematics

cnt	Estimate						Linking error					
	I1	I2	I3	I4	D2	D4	I1	I2	I3	I4	D2	D4
AUS	−6.8	−9.5	−7.2	−9.8	−10.3	−9.8	2.7	7.3	1.6	5.9	7.6	5.9
AUT	−12.8	−8.2	**−16.4**	−13.1	−12.4	−13.1	9.0	13.4	9.5	12.0	13.8	12.0
BEL	−11.3	−10.6	−10.0	−10.7	−10.2	−10.7	2.6	7.2	2.1	6.2	7.5	6.2
CAN	−3.6	−1.0	−3.1	−2.4	−2.1	−2.4	1.5	3.7	3.0	4.7	4.7	4.7
CHE	2.7	0.3	3.1	1.4	1.2	1.4	3.7	5.9	2.8	5.8	3.7	5.8
CHL	2.3	7.4	**−0.3**	3.9	2.9	3.6	4.0	5.7	2.3	3.5	5.3	4.9
CZE	−20.4*	**−15.1***	**−23.1***	−17.8*	−17.9*	−17.8*	7.4	7.4	5.6	6.9	5.7	6.9
DEU	4.5	6.5	**1.2**	0.2	4.6	0.2	2.4	4.9	2.5	5.3	4.7	5.3
DNK	−15.5*	−15.1	**−11.3***	−22.1	−13.7	−22.1*	2.3	15.5	2.2	13.2	22.4	14.9
ESP	−0.5	−1.4	**−2.0**	−3.0	−3.6	−3.0	3.8	6.1	2.4	2.8	4.0	2.8
EST	−4.4	4.3	−4.1	4.3	3.0	4.3	3.6	18.9	6.2	13.6	17.3	12.2
FIN	−11.8*	−7.9	−11.5*	−8.7	−12.4	−12.3	5.6	9.2	7.6	10.1	4.3	5.2
FRA	−2.5	−2.6	−2.1	0.8	0.2	0.8	4.6	7.6	5.4	10.7	11.8	10.7
GBR	−4.0	−6.2	−3.3	−4.7	−4.9	−4.7	6.1	2.4	6.1	2.8	3.0	2.8
GRC	2.7	7.4	**−1.1**	**−2.6**	**−2.6**	**−2.6**	5.3	8.6	7.5	5.7	3.8	5.7
HUN	−3.4	−3.5	−3.1	−5.3	−5.5	−5.3	2.7	4.6	5.8	2.9	2.6	2.9
IRL	−16.4*	−17.6*	**−13.1***	−14.7*	−15.9*	−14.7*	3.7	5.1	2.4	2.7	3.2	2.7
ISL	−3.9	**2.1**	−2.4	**1.5**	**1.8**	**1.5**	6.6	6.4	8.9	6.3	7.3	6.3
ISR	−4.4	**−0.6**	**−0.7**	**1.9**	**1.7**	**1.9**	10.0	10.7	5.6	5.6	6.6	5.6
ITA	16.4*	16.4*	**11.3***	**11.0***	**10.1**	**11.0***	5.7	8.0	3.8	7.2	5.0	7.2
JPN	4.2	5.4	**8.4**	**10.8**	**11.4**	**10.8**	4.2	5.8	6.2	8.4	9.4	8.4
KOR	−1.0	−2.2	−1.7	−3.5	2.9	**4.7**	6.8	8.8	5.6	10.1	4.0	4.8
LUX	−4.8	−7.5	−4.8	−6.6	−8.4	−6.6	5.8	6.2	6.5	5.9	5.0	5.9
MEX	3.2	3.5	2.2	3.2	2.6	3.2	8.0	14.3	3.9	5.2	6.7	5.2
NLD	−7.0	−7.2	−5.0	−7.2	−3.7	−7.2	2.4	3.9	5.2	6.7	3.5	6.7
NOR	3.0	3.7	3.6	1.3	1.9	1.3	4.6	10.6	2.4	9.6	10.5	9.6
POL	−3.6	−0.9	**0.8**	**2.4**	**1.7**	**2.4**	3.4	8.0	5.4	3.7	4.7	3.7
PRT	16.3	18.3	**13.6**	11.1	**10.0**	**11.0**	2.6	8.7	3.3	5.5	2.9	4.0
SWE	−13.3	−12.0*	−13.9*	−13.9	−12.8*	−13.9	5.8	3.7	8.6	6.7	5.9	6.7
TUR	17.0*	23.4	15.5*	15.1	15.4	15.1	5.6	16.6	5.0	8.6	7.2	8.6

cnt country

* Statistically significant trend estimates ($p < .05$) with used standard error TE_{IPD} Estimates in I2, I3, I4, D2 and D4 that significantly differ from I1 are printed in bold

The linking error LE_{IPD} for the trend in standard deviations at the international metric was determined as 2.0 and was substantially smaller than the linking errors that also accounted for DIF and DIF × IPD effects. The median of the linking errors across countries was I1: 7.6, I2: 6.7, I3: 6.3, I4: 5.5, D2: 8.0, D4: 5.5. In line with the empirical findings of the variability in trend estimates in standard deviations, linking errors based on only link items turned out to be smaller. The proportion of significant trend estimates in standard deviations (based on TE_{IPD}) were I1: 47%, I2: 60%, I3: 37%, I4: 10%, D2: 53%, and D4: 13%. Strikingly, the methods that utilized only link items resulted in a lower proportion of significant trend estimates.

Table 7 Trend estimates in country means and their linking errors for the PISA Trend from 2006 to 2009 in science

	Estimate						Linking error					
cnt	I1	I2	I3	I4	D2	D4	I1	I2	I3	I4	D2	D4
AUS	−6.0	−5.2	**−3.1**	−4.0	−3.9	−4.0	7.9	7.4	5.3	7.7	7.6	7.7
AUT	−20.3*	−19.7*	−20.2*	−16.7*	−17.4*	−16.7*	5.3	5.7	5.5	6.6	3.7	6.6
BEL	−10.6	−8.7	**−14.1***	−10.5	−9.9	−10.5	8.0	7.1	4.4	4.3	5.9	4.3
CAN	−12.5*	−11.9*	**−9.3**	−9.1	−9.7	−9.1	4.7	6.6	5.2	6.1	6.1	6.1
CHE	−1.2	1.9	**0.3**	**3.0**	**3.4**	**3.0**	5.1	6.0	4.5	3.3	3.5	3.3
CHL	−10.9	−8.3	**−13.0**	−11.9	−11.2	−11.9	7.7	12.7	5.8	7.5	6.4	6.4
CZE	−18.1	−14.6	**−12.5**	**−9.8**	−12.4	−11.3	7.9	7.8	4.9	7.3	7.8	6.6
DEU	−0.4	1.7	**1.5**	**5.8**	**6.2**	**5.8**	6.3	9.1	4.2	5.4	7.5	5.4
DNK	−7.5	−3.8	**−0.9**	−4.1	−4.6	−4.1	4.9	4.0	5.0	6.0	5.3	5.6
ES	−11.3	−8.7	−10.9	−10.2	−10.4	−9.9	6.2	7.2	6.7	5.8	8.9	6.5
EST	−11.5*	−7.8	−12.7*	−9.2	−10.9	−11.0	9.0	6.5	5.6	6.5	4.8	4.9
FIN	−12.1*	−15.3*	−10.3	−15.6*	−16.9*	−15.6*	5.9	6.6	5.5	5.7	8.3	5.7
FRA	−7.5	−11.0	−6.3	−8.8	−2.9	−5.7	4.4	8.1	6.0	6.2	10.8	5.7
GBR	−7.1	−9.6	**−3.7**	−5.4	−5.3	−5.4	9.7	8.8	5.7	5.9	6.4	5.9
GRC	−15.4	−17.8	−16.3	−17.2	−11.7	−12.3	5.8	10.3	4.6	9.6	9.5	7.4
HUN	−8.6	−4.1	**−4.0**	−2.9	−5.5	−7.0	8.2	2.1	7.1	8.9	5.6	5.7
IRL	−7.7	−7.2	**−9.3**	−5.9	−7.1	−5.9	10.6	6.4	4.1	7.5	6.4	7.5
ISL	−2.4	0.2	−3.2	−1.4	−2.9	−1.5	5.7	11.5	4.9	10.6	9.5	10.5
ISR	−12.2	−13.9	**−8.7**	−8.2	**−7.6**	**−7.2**	4.9	5.7	7.3	8.8	4.0	5.2
ITA	3.2	3.0	**4.6**	4.3	4.0	4.5	10.8	8.6	5.6	4.2	4.9	4.4
JN	5.7	0.7	**−0.2**	−4.5	**−1.4**	**−3.5**	9.7	12.0	9.3	15.4	10.1	11.5
KOR	10.9	25.5*	**7.6**	13.8	19.3	5.1	7.5	17.9	7.0	22.0	42.4	24.2
LUX	−11.4*	−9.9	−12.5*	−11.7*	−10.7	−11.7*	5.1	4.8	4.3	6.7	3.5	6.7
MEX	−15.0*	−16.2*	−15.1*	−14.7*	−12.7	−13.2*	5.3	7.6	6.9	8.7	8.7	8.9
NLD	−7.4	−4.7	−6.6	−5.6	−5.0	−5.6	8.4	6.6	7.4	5.3	5.9	5.3
NOR	4.8	8.7	5.8	7.0	6.9	7.1	6.7	5.0	7.6	3.8	3.3	3.9
OL	2.1	2.5	**−1.2**	−2.2	−0.6	−2.2	4.7	5.9	5.7	7.8	7.9	7.8
RT	10.3	9.8	9.8	10.8	10.0	7.9	6.6	7.7	4.6	6.4	5.5	6.4
SWE	−15.3*	−17.2*	**−10.9**	−17.1*	−13.2	−13.4*	7.6	9.8	6.5	4.2	5.7	3.9
TUR	18.9*	19.3*	19.1*	19.5*	20.6*	19.6*	6.2	8.1	7.9	4.7	4.3	4.1

cnt country

* statistically significant trend estimates ($p < .05$) with used standard error TE_{IPD} Estimates in I2, I3, I4, D2 and D4 that significantly differ from I1 are printed in bold

Trend in country means for mathematics

Table 6 contains trend estimates in the country means for mathematics. As mathematics was a minor domain in PISA 2006 and PISA 2009, we expect smaller differences between different scaling methods compared to the trend for the reading domain. The median range in different trend estimates across models was 5.3 ($Min = 1.3$, $Max = 10.8$) and turned out to be smaller than for the reading domain. The smallest range was obtained for Belgium (BEL), whose trend estimates in the country means ranged between −11.3 and −10.0. The median of the country mean trend estimates were I1: −3.6, I2: −1.2, I3: −2.3, I4: −2.5, D2: −1.0, D4: −1.1. Hence, on average, the trends were very similar

for the different methods. The median of absolute values in trend estimates also did not differ substantially: I1: 4.4, I2: 6.9, I3: 3.9, I4: 5.0, D2: 4.8, D4: 5.0.

The linking error LE_{IPD} at the international metric due to IPD was estimated as 2.0. Again, our proposed linking errors based on jackknife has larger medians: I1: 4.4, I2: 7.4, I3: 5.3, I4: 6.1, D2: 5.2, D4: 5.9. Compared to the reading domain, the proportion of significant trend estimates in country means was smaller in mathematics: I1: 20%, I2: 13%, I3: 23%, I4: 10%, D2: 10%, D4: 13%.

Trend in country standard deviations for mathematics.

We now briefly summarize the findings for trends in country standard deviations for mathematics (not presented in a table). The median of the range across models was 3.6 ($Min = 1.0$, $Max = 7.5$), which can be considered relatively small. We also observed very similar median estimates of absolute values in trend estimates in standard deviations (ranging between 2.2 for D2 and 2.6 for I2).

The estimated linking error LE_{IPD} at the international metric was 1.1. The median of estimated linking errors based on jackknife did not differ substantially (ranging between 3.2 for I1 and 4.1 for I4). Only a few countries showed statistically significant trend estimates: I1: 10%, I2: 7%, I3: 0%, I4: 3%, D2: 10%, D4: 3%.

Trend in country means for science

Table 7 presents trend estimates in the country means for the science domain. Science was a major domain in PISA 2006 and a minor domain in PISA 2009. Hence, more variability in trend estimates compared to mathematics would be expected. The median of the range across the six different methods within a country was 4.7 ($Min = 1.6$, $Max = 20.4$). The largest range was obtained for South Korea (KOR), whose trend estimates varied between 5.1 and 25.5. The median of the trend estimates across countries was relatively similar across models: I1: −7.6, I2: −7.5, I3: −6.5, I4: −5.8, D2: −5.4, D4: −5.8. The empirical variability of absolute values using the median in the country mean trend estimates were also similar across models (ranging between 7.2 for D4 and 10.5 for I1). For science, we did not observe that the trend estimates based on link items were less variable than the corresponding estimates based on all items.

The estimated linking error LE_{IPD} at the international metric due to IPD was 3.9 and was smaller than the median values of linking errors that also account for DIF: I1: 6.5, I2: 7.3, I3: 5.6, I4: 6.5, D2: 6.3, D4: 6.0. For the science domain, about 20% of the countries had significant trend estimates in means (I1: 27%, I2: 23%, I3: 20%, I4: 20%, D2: 10%, D4: 20%).

Trend in country standard deviations for science

We now briefly report the results for trend estimates in standard deviations for science (not presented in a table). The median range across models was 3.7 ($Min = 1.1$, $Max = 10.7$). The median estimates of trend estimates in standard deviations did not differ much (ranging between 3.0 for D4 and 5.8 for I2). Moreover, the variability in trend estimates was also similar but slightly smaller for the models based on only link items (I1: 5.0, I2: 6.0, I3: 3.4, I4: 4.2, D2: 4.5, D4: 4.2).

The linking error LE_{IPD} at the international metric was estimated as 2.2. The median of linking errors based on jackknife was similar across models (ranging between 3.3 for I1 and 4.1 for I4). The proportion of significant trend estimates in standard deviations was I1: 23%, I2: 23%, I3: 17%, I4: 17%, D2: 27%, D4: 20%.

Discussion

In this article, we compared the original trend estimate (model I1) with alternative trend estimates for two time points. The proposed trend estimates differ regarding three aspects. First, one can distinguish between the trend estimates that involve all items (link items and unique items) and those that utilize link items only. Second, one can use country-specific scaling models that allow all item parameters to be noninvariant across countries, or one can rely on a full invariance assumption that employs item parameters obtained from a pooled calibration sample comprising all countries. Third, one can distinguish whether an indirect or a direct linking approach of a country should be carried out. In indirect linking approaches, two scaling models for the two points based on the calibration samples at the international metric were linked to countries. Hence, in this case, countries have to be linked twice to the international metric: at the first time and the second time point. When utilizing direct linking, the trend estimate of a country only relies on linking item parameters from country-specific scaling models for the two time points. Hence, there is only one link of the country to the international metric at the first time point.

In the simulation study involving the 1PL and the 2PL model, we found that alternative trend estimates to original trend estimates can increase the efficiency of trend estimates in the presence of DIF, IPD, and DIF × IPD effects. This means that alternatives to the currently reported trend estimates would result in less variable trend estimates. This observation was also found in the empirical application using the PISA trend from PISA 2006 to PISA 2009. Importantly, a linking based on only link items can be more efficient if there are substantial cross-sectional DIF effects and changes from minor to major domains (or the other way around) in PISA designs. By relying on only link items, potential DIF effects of unique items simply do not appear in differences (i.e., between country means and standard deviations across time) and thus do not contribute to trend estimates. There seems to be a preference for including all items (i.e., unique and link items) because more stable trend estimates should be obtained using all items (von Davier et al., 2019). However, the analyses in this article refute this statement (see also Heine & Robitzsch, 2022).

Hence, the general recommendation for always opting for concurrent scaling models with invariant item parameters that involve all items at the time points can be questioned for statistical reasons. We think that there is one obvious but convincing justification for relying on models that involve all items: Trend estimates should be computed as differences from reported cross-sectional estimates. Such a simple computation eases the interpretation of results for studies like PISA that target policymakers. Notably, a convincing and valid parameter estimate will typically not be the one that would be preferred from a statistical perspective (Robitzsch & Lüdtke, 2022b). However, there could also be good reasons to argue that different scaling models would be required to answer different questions. In this sense, marginal trend estimates (or indirect linking

methods) could be preferred over original trend estimates (or direct linking methods; see Carstensen, 2013, for such an argument).

Our simulation and analytical derivations demonstrated that indirect and direct linking approaches resulted in unbiased trend estimates. In previous literature, it is argued that using marginal trend estimates (as a direct linking method) reduces the comparability of trend estimates across countries compared to original trend estimates (as an indirect linking method). We do not think such a statement can be defended from a statistical point of view. In fact, Carstensen (2013) argued that original trend estimates reduce the stability of trend estimates because the original trend relies on cross-sectional estimates that are aimed to maximize the cross-sectional comparability of countries in the PISA test.

Furthermore, we analytically derived the variance of cross-sectional mean and standard deviation estimates for the 1PL model in the presence of DIF. The variance includes sampling error due to the sampling of persons and linking error due to the presence of DIF effects. Similar simple expressions for the variance of the mean and the standard deviation estimates could also be obtained for the 2PL model in the presence of uniform DIF. However, we think that there is no simple closed formula for the variance of the estimates in the presence of nonuniform DIF in the 2PL model.

In test designs involving two time points with the same items (i.e., no unique items), original trend estimates cannot be substantially improved in terms of efficiency. However, more efficient trend estimates for standard deviations were obtained if the scaling models utilized country-specific item parameters (i.e., they relied on noninvariance). Our findings did not substantially differ for the 1PL and the 2PL model. However, larger sample sizes were required for the 2PL model than the 1PL model to realize efficiency gains for models using only link items and noninvariant item parameters. Interestingly, the consequences of the choice of trend estimation methods were more substantial for country standard deviations (or their trends) than for country means. In empirical applications, this finding could have consequences for assessing performance gaps between groups of strongly differing abilities or quantiles. Similar findings were found regarding the choice of the scaling model (Robitzsch, 2022a) and the treatment of missing data (Robitzsch, 2021a).

Our analysis demonstrated that larger sample sizes are required in the 2PL model to realize efficiency gains. However, entirely relying on invariant item discriminations and item difficulties anchored at the international metric might be the other extreme compared to using country-specific separate 2PL scaling models. However, one could use ridge-type regularization penalties (Battauz, 2020) or hierarchical models (Fox & Verhagen, 2010; König et al., 2020) for stabilizing model estimation, which can subsequently result in more efficient country mean and standard deviation estimates in the linking approach. Alternatively, country-specific scaling models could use item discrimination parameters fixed to those at the international metric while freely estimating item intercepts. This approach would align with empirical findings that the extent of uniform DIF is more pronounced than of nonuniform DIF.

As with any simulation study, we only study a limited number of conditions. We expect efficiency gains (and losses) to be smaller for a larger number of items. Moreover, models that assume noninvariant item parameters will likely be more unstable

in smaller sample sizes than those studied in the simulation (e.g., $N = 250$). Finally, further simulation studies could involve more than 20 countries, although we do not think general findings would change.

Throughout the simulation study, we assumed a known functional form of the item response function. That is, it was assumed that the data-generating model (i.e., 1PL or 2PL model) coincided with the analysis model. In practice, the data-generating model will likely be more complex than the analysis model. The choice of an item response model and a particular estimation method might be seen as defining the parameter of interest. Hence, the analysis utilizes intentionally misspecified statistical models (Robitzsch & Lüdtke, 2022b; Robitzsch, 2022a). In future research, the consequences of using misspecified scaling models for linking and trend estimates could be investigated (but see Fischer et al., 2021).

In the simulation study, we only addressed the efficiency of trend estimates in country means and standard deviations. In the empirical example of the PISA trend, we also computed linking errors based on jackknife to take the uncertainty for trend estimates due to DIF, IPD, and DIF × IPD effects into account. We think utilizing adequate linking errors is vital in ILSA studies to not interpret results in trend estimates as statistically (and practically) significant because the effect of test designs and item samples remained unquantified in the reported uncertainty.

When utilizing the direct linking method, trend estimates are essentially carried out at the country level. Items that function differently across countries would induce uncertainty in trend estimates. Some researchers argued that longitudinal invariance (i.e., IPD) would be a prerequisite for meaningful and valid trend estimates (Carstensen, 2013; Fischer et al., 2019; Rohm et al., 2021; Wetzel & Carstensen, 2013). We tend to disagree with this opinion. One cannot expect that items homogeneously function across a long time range, and there is no reason one would require the absence of IPD. The presence of IPD is just another source of variance that must be accounted for in the quantification of uncertainty (i.e., represented in the linking error). Moreover, the restriction of trend analysis to invariant items is a potential threat to validity because the construct could be changed if it is only represented by a subset of items.

An anonymous reviewer wondered how our findings would generalize to the computation of trend estimates for three points (i.e., time points 1, 2, and 3). For example, the three time points could be studies of PISA 2009 (time point 1), PISA 2012 (time point 2), and PISA 2015 (time point 3). Suppose that a researcher would be interested in computing a trend estimate between time points 1 and 3; that is, the trend between PISA 2009 and PISA 2015. In operational practice, trend estimates are reported based on chain linking. In this approach, the first linking step is a linking between PISA 2009 and PISA 2012, while the studies PISA 2012 and PISA 2015 are linked in the second linking step. The trend estimate between PISA 2009 and PISA 2015 can be obtained as a consequence that the scores in the three studies are linked onto a common metric by utilizing chain linking. An alternative approach might be to conduct non-chained linking[8] between PISA 2009 (time point 1) and PISA 2015 (time point 3). The non-chained

[8] We use the term "non-chained linking" instead of the more obvious term "direct linking". Note that the term "direct linking" is already used in a different manner in this article.

linking might be preferred if the number of link items at time point 1 and time point 3 is (approximately) as large as the number of link items in each of the two steps in chain linking. If this is the case, there is no need to resort to chain linking from the statistical perspective of estimation efficiency. For chain linking or non-chained linking, indirect or direct linking approaches could be applied. Our general findings regarding the different performance of indirect and direct linking methods can also be applied to choosing appropriate methods in the linking steps involved in linking three time points.

If researchers cannot agree on model choice or there are a set of plausible statistical models, the variability due to model choice (i.e., model error; Longford, 2012) can be included as another source of uncertainty in the analysis. The findings for trend estimates in this article (Heine & Robitzsch, 2022) demonstrate that this variability cannot be considered negligible and can exceed standard errors due to the sampling of students (Robitzsch, 2022a).

Finally, the partial invariance approach (Joo et al., 2021; von Davier et al., 2019) is becoming popular in ILSA studies like PISA. This approach relies on the assumption that DIF effects are sparsely distributed; that is, only a few item parameters differ across countries, while the majority of items receive a common invariant parameter. We are not convinced by this approach for two reasons. First, our experience from empirical applications contradicts the assumption of sparsely distributed DIF effects. It seems that DIF effects are rather symmetrically and normally distributed. The partial invariance assumption of DIF effects seems to be as rare as unicorns in empirical applications (see Robitzsch & Lüdtke, 2022b). Second, we argued elsewhere that scaling models relying on partial invariance do not provide meaningful and valid cross-sectional and trend comparisons across countries because the set of items is allowed to differ for each pair of countries under comparison (Robitzsch & Lüdtke, 2022a, 2022b). In this sense, one compares apples with oranges in the partial invariance model. In contrast, the models utilized in this article rely on the same items. No item contribution would be downweighted or removed (see Robitzsch, 2021b), a property we consider a prerequisite for meaningful and valid comparisons. Andersson (2018) also provided computational arguments against using concurrent scaling with a multiple-group IRT model (such as the one that relies on partial invariance) involving data from all countries. First, it is more difficult to achieve convergence in the estimation in the concurrent model than in each of the country-specific scaling models. Second, it is also easier to diagnose potential estimation issues when estimating the item parameters separately. Third, the computation time of running a large concurrent scaling model frequently exceeds the required computation time for running many separate scaling models. Hence, one can recognize that the method of concurrent calibration involving thousands (or even millions) of students might be computationally feasible, we nevertheless doubt the practical utility of such concurrent scaling approaches.

Conclusion

In this article, we compared different trend estimation approaches for ILSA datasets. Previous literature mainly compared the two alternative original and marginal trend estimates. However, we suggest generally distinguishing trend estimation approaches regarding three factors: the type of linking to an international metric (indirect vs. direct), the set of items used (all items vs. only link items), and whether item parameters are

assumed to be invariant or noninvariant across countries. We showed that the direct linking approach (e.g., the marginal trend as a particular case) could result in more efficient trend estimates in the presence of DIF than indirect linking approaches (e.g., the original trend as a particular case that involves all items) if country DIF exists. As an alternative, indirect linking approaches could be used that only rely on link items. Moreover, the different performance of the original and marginal trend estimates can be explained by the fact that the marginal trend essentially only uses link items for linking, while the original trend relies on all items. Finally, we observed that the particular choice of a trend estimation method could be more consequential for trend estimates in country standard deviations than for country means if uniform DIF is present.

Appendix A: generation of synthetic data in the empirical example

To enable the replicability of the results of the empirical example by independent researchers, we created synthetic Datasets for the two PISA assessments that strongly resemble the original data. The principle of data generation relied on the approach of Jiang et al. (2022), which was also investigated by Grund et al. (2022). Synthetic datasets were produced at the country level for each of the three cognitive domains and the two PISA assessments, PISA 2006 and PISA 2009. To also represent the balanced incomplete block design in the synthetic datasets, we applied the synthesis model at the level of each administered booklet in the test.

The basic idea of generating data is to simulate a dataset whose distribution of multivariate variables is very close to the original datasets (Nowok et al., 2016). The original dataset can be decomposed into a vector of fixed variables **Z** that remains unaltered and a vector **X** of variables that should be synthesized. In our application, **Z** contained student weights and the proportion of correct item responses and their squared terms. The vector **X** consisted of all item responses of items that were administered in a respective test booklet. Synthetic data $\mathbf{X}_{\text{syn}}$ can be formally derived as a simulation draw from the posterior distribution (Grund et al., 2022)

$$\mathbf{X}_{\text{syn}} \sim \mathrm{P}(\mathbf{X}|\mathbf{Z}) = \frac{\mathrm{P}(\mathbf{X},\mathbf{Z})}{\mathrm{P}(\mathbf{Z})} \tag{22}$$

Jiang et al. (2022) proposed the brilliant idea of including noisy realizations of **X** and also using them as fixed variables. That is, they define the data-augmented variables $\mathbf{X}^*$ by adding random noise **e** to the original data

$$\mathbf{X^*} = \mathbf{X} + \mathbf{e}\,. \tag{23}$$

The components of **e** are typically chosen to be independent of each other. The variance of **e** can be chosen such that variable-specific reliabilities $\mathrm{Var}(X_i)/\mathrm{Var}(X_i^*)$ are relatively large, such as 0.90. This means that the noise variance $\mathrm{Var}(e_i)$ is determined such that $\mathrm{Var}(e_i)/\mathrm{Var}(X_i^*)$ equals 0.10. Instead of simulating a synthetic dataset from the posterior distribution $\mathrm{P}(\mathbf{X}|\mathbf{Z})$, Jiang et al. (2022) use the posterior distribution $\mathrm{P}(\mathbf{X}|\mathbf{Z},\mathbf{X}^*)$ of the augmented dataset. The posterior distribution has a closed form in the case of multivariate normality. Although this was not fulfilled in our application that involves

dichotomous item responses **X**, we nevertheless utilized linear regression models as working models for synthetic dataset generation. The reliability was chosen as 0.90 for all items.

The data synthesis used in this article followed a sequential approach. Let $\mathbf{X} = (X_1, \ldots, X_I)$ denote the vector of item responses and $\mathbf{X}_{(<i)} = (X_1, \ldots, X_{i-1})$ $(i > 1)$ contains all items with a variable index smaller than i. For synthesizing variable X_i, a linear regression of X_i on the predictors $\mathbf{X}_{(<i)}$, $\mathbf{Z}$, and $\mathbf{X}^*$ is computed. As sample sizes of the booklet level in our empirical example were frequently not very large compared to the number of predictors in the regression model, we first reduced the dimensionality of the predictors using partial least squares (PLS; Mevik & Wehrens, 2007) regression (see, e.g., Grund et al., 2021; Robitzsch, 2021a). We chose 20 PLS factors for dimension reduction in the linear regression model for all items.

Based on this regression model, linear predictions $\widehat{x}_{ni}$ are computed for all cases n. A normally distributed noise variable u_{ni} was simulated with a standard deviation that equals the residual standard deviation from the regression model. The vector containing all u_{ni} values can be residualized with respect to all predictors in the regression model to ensure that the original relationships between variables are (almost) equal in the synthetic datasets. Next, we computed $\widetilde{x}_{ni} = \widehat{x}_{ni} + u_{ni}$. In the case of continuous variables, these values could be used as synthesized values. However, our goal was to simulate dichotomous item responses. Moreover, we wanted to preserve the marginal distribution of the variable X_i. Therefore, we ordered the values of $\widetilde{x}_{ni}$ and assigned a corresponding value in the synthetic data by ordered values of the original data x_{ni} (the so-called "normrank" approach; see Nowok et al., 2016). Note that in the case of dichotomous item responses, these reordered values correspond to a vector of zeroes and ones, where the entries of ones follow the zero entries. By carrying out this principle, the marginal distribution is preserved. Because the proportion of correctly solved items was used as a fixed predictor in the synthesis model, it could be expected that the intra-class correlation referring to the clustered data structure of students nested within schools was approximately represented.

The synthetic datasets in our example were generated using the `miceadds::syn_da()` function from the R package miceadds (Robitzsch & Grund, 2022). See https://osf.io/n5zm6/?view_only=086ea651bbea49bb8b2aae44e3971db8 for a replication syntax.

Acknowledgements
Not applicable.

Author contributions
Both authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding
There is no external funding.

Availability of data and materials
Replication material can be found at https://osf.io/n5zm6/?view_only=086ea651bbea49bb8b2aae44e3971db8. The PISA 2006 and PISA 2009 datasets used in the empirical example can be downloaded from https://www.oecd.org/pisa/pisaproducts/database-pisa2006.htm and https://www.oecd.org/pisa/pisaproducts/pisa2009database-downloadabledata.htm, respectively.

Declarations

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

Received: 12 November 2022 Accepted: 7 July 2023
Published online: 19 July 2023

References

Andersson, B. (2018). Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Applied Psychological Measurement, 42*(3), 192–205. https://doi.org/10.1177/0146621617721249

Battauz, M. (2020). Regularized estimation of the four-parameter logistic model. *Psych, 2*(4), 269–278. https://doi.org/10.3390/psych2040020

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.

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

Cai, L., & Moustaki, I. (2018). Estimation methods in latent variable models for categorical outcome variables. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: a multidisciplinary reference on survey, scale and test* (pp. 253–277). New York: Wiley. https://doi.org/10.1002/9781118489772.ch9

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.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (pp. 221–256). Praeger Publisher.

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). Amsterdam: Springer. https://doi.org/10.1007/978-94-007-4458-5_12

Carstensen, C. H., Prenzel, M., & Baumert, J. (2008). Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? [Trend analyses in PISA: How did competencies in Germany develop between PISA 2000 and PISA 2006?]. In M. Prenzel & J. Baumert (Eds.), *Vertiefende Analysen zu PISA 2006* (pp. 11–34). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91815-0_2

Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. Springer. https://doi.org/10.1007/978-0-387-49771-6

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2019). Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: a comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7. *Psychological Test and Assessment Modeling, 61*(1), 37–64.

Fischer, L., Rohm, T., Carstensen, C. H., & Gnambs, T. (2021). Linking of Rasch-scaled tests: consequences of limited item pools and model misfit. *Frontiers in Psychology, 12*, 633896. https://doi.org/10.3389/fpsyg.2021.633896

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.

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*, 305–322.

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

Grund, S., Lüdtke, O., & Robitzsch, A. (2022). Using synthetic data to improve the reproducibility of statistical results in psychological research. *Psychological Methods*. https://doi.org/10.1037/met0000526

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (ETS Research Report ETS RR-09-40). *Princeton, ETS*. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*(3), 144–149. https://doi.org/10.4992/psycholres1954.22.144

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3–24. https://doi.org/10.1177/0146621602026001001

Hastedt, D., & Desa, D. (2015). Linking errors between two populations and tests: a case study in international surveys in education. *Practical Assessment, Research, and Evaluation, 20*, 14. https://doi.org/10.7275/yk4s-0a49

Heine, J.-H., & Robitzsch, A. (2022). Evaluating the effects of analytical decisions in large-scale assessments: analyzing PISA mathematics 2003–2012. *Large-Scale Assessments in Education, 10*, 10. https://doi.org/10.1186/s40536-022-00129-5

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

Jiang, B., Raftery, A. E., Steele, R. J., & Wang, N. (2022). Balancing inferential integrity and disclosure risk via model targeted masking and multiple imputation. *Journal of the American Statistical Association, 117*(537), 52–66. https://doi.org/10.1080/01621459.2021.1909597

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

Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. *Asia Pacific Education Review, 13*(2), 311–321. https://doi.org/10.1007/s12564-011-9197-2

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking*. Springer. https://doi.org/10.1007/978-1-4939-0317-7

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

König, C., Khorramdel, L., Yamamoto, K., & Frey, A. (2021). The benefits of fixed item parameter calibration for parameter accuracy in small sample situations in large-scale assessments. *Educational Measurement: Issues and Practice, 40*(1), 17–27. https://doi.org/10.1111/emip.12381

König, C., Spoden, C., & Frey, A. (2020). An optimized Bayesian hierarchical two-parameter logistic model for small-sample item calibration. *Applied Psychological Measurement, 44*(4), 311–326. https://doi.org/10.1177/0146621619893786

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210–231. https://doi.org/10.1007/s11336-013-9347-z

Longford, N. T. (2012). 'Which model?' is the wrong question. *Statistica Neerlandica, 66*(3), 237–252. https://doi.org/10.1111/j.1467-9574.2011.00517.x

Magis, D., & De Boeck, P. (2011). Identification of differential item functioning in multiple-group settings: a multivariate outlier detection approach. *Multivariate Behavioral Research, 46*(5), 733–755. https://doi.org/10.1080/00273171.2011.606757

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017).*Methods and Procedures in PIRLS 2016*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/publications/pirls/2016-methods.html

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 229–258). CRC Press.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing, 5*(3), 279–300. https://doi.org/10.1207/s15327574ijt0503_6

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

Mevik, B. H., & Wehrens, R. (2007). The pls package: principal component and partial least squares regression in R. *Journal of Statistical Software, 18*(2), 1–23. https://doi.org/10.18637/jss.v018.i02

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8*, 323–335.

Monseur, C., Sibberns, H., & Hastedt, D. (2008). Linking errors in trend estimation for international surveys in education. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 1*, 113–122.

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software, 74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11

OECD. (2009). *PISA 2006 technical report*. OECD Publishing.

OECD. (2012). *PISA 2009 technical report*. OECD Publishing.

OECD. (2014). *PISA 2012 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.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*, 1–21. https://doi.org/10.1080/15305058.2013.825265

Oliveri, M. E., & von Davier, M. (2017). Analyzing the invariance of item parameters used to estimate trends in international large-scale assessments. In H. Jiao & R. W. Lissitz (Eds.), *Test fairness in the new generation of large-scale assessment* (pp. 121–146). Information Age Publishing.

Pohl, S., Haberkorn, K., & Carstensen, C. H. (2015). Measuring competencies across the lifespan: challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds.), *Dependent data in social sciences research* (pp. 281–308). Cham: Springer. https://doi.org/10.1007/978-3-319-20585-4_12

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria. https://www.R-project.org/

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). 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. (2021b). 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). 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. (2022). *sirt: Supplementary item response theory models*. R package version 3.12–66. http://CRAN.R-project.org/package=sirt

Robitzsch, A. (2023). Linking error in the 2PL model. *J, 6*(1), 58–84. https://doi.org/10.3390/j6010005

Robitzsch, A., & Grund, S. (2022). *miceadds: Some additional imputation functions, Especially for mice*. R package version 3.16–4. https://github.com/alexanderrobitzsch/miceadds

Robitzsch, A., Kiefer, T., & Wu, M. (2022). *TAM: Test analysis modules*. R package version 4.1–4. http://CRAN.R-project.org/package=TAM

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. (2020). 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.

Robitzsch, A., & Lüdtke, O. (2022a). 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. (2022b). 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*, 9. https://doi.org/10.1186/s42409-022-00039-w

Rohm, T., Carstensen, C. H., Fischer, L., & Gnambs, T. (2021). The achievement gap in reading competence: the effect of measurement non-invariance across school types. *Large-Scale Assessments in Education, 6*(1), 23. https://doi.org/10.1186/s40536-021-00116-2

Rutkowski, D., & Rutkowski, L. (2022). The promise and methodological limits of international large-scale assessments. In L. I. Misiaszek, R. F. Arnove, & C. A. Torres (Eds.), *Emergent trends in comparative education: the dialectic of the global and the local* (pp. 253–268). Lankam: Rowman Littlefied.

Rutkowski, L., von Davier, M., & Rutkowski, D. (2014). *Handbook of international large-scale assessment*. Boca Raton: CRC Press. https://doi.org/10.1201/b16061

Sachse, K. A., & Haag, N. (2017). Standard errors for national trends in international large-scale assessments in the case of cross-national differential item functioning. *Applied Measurement in Education, 30*(2), 102–116. https://doi.org/10.1080/08957347.2017.1283315

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

von Davier, M., & Bezirhan, U. (2023). A robust method for detecting item misfit in large scale assessments. *Educational Psychological Measurement, 83*(4), 740–765. https://doi.org/10.1177/00131644221105819

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: Background, technical issues, and methods of data analysis* (pp. 155–174). London: CRC Press. https://doi.org/10.1201/b16061-12

von Davier, M., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & 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

Wang, W., Liu, Y., & Liu, H. (2022). Testing differential item functioning without predefined anchor items using robust regression. *Journal of Educational and Behavioral Statistics, 47*(6), 666–692. https://doi.org/10.3102/10769986221109208

Weeks, J., von Davier, M., & Yamamoto, K. (2014). Design considerations for the program for international student assessment. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 259–275). CRC Press.

Wetzel, E., & Carstensen, C. H. (2013). Linking PISA 2000 and PISA 2009: implications of instrument design on measurement invariance. *Psychological Test and Assessment Modeling, 55*(2), 181–206.

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

Xia, D.-F., Xu, S.-L., & Qi, F. (1999). A proof of the arithmetic mean-geometric mean-harmonic mean inequalities. *RGMIA Research Report Collection, 2*, 1. http://ajmaa.org/RGMIA/papers/v2n1/v2n1-10.pdf

Yuan, K. H., Cheng, Y., & Patton, J. (2014). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika, 79*(2), 232–254. https://doi.org/10.1007/s11336-013-9334-4

Publisher's Note